

基于文本的检索和编辑的多模态分子结构-文本模型

Shengchao Liu^{1,2}, Weili Nie³, Chengpeng Wang⁴, Jiarui Lu^{1,2}, Zhuoran Qiao⁵, Ling Liu⁶, Jian Tang^{*1,7}, Chaowei Xiao^{*3,8}, and Animashree Anandkumar^{*3,5}

¹米拉-魁北克人工智能研究所，加拿大魁北克省蒙特勒，H2S 3H1

²Universit  de Montr al, Montr al, QC H3T 1J4, Canada

³美国加利福尼亚州圣克拉拉市 95051 英伟达研究院

⁴美国伊利诺伊大学香槟分校，伊利诺伊州香槟市，61801

⁵美国加利福尼亚州帕萨迪纳市加利福尼亚理工学院 邮编：91125

⁶美国新泽西州普林斯顿市普林斯顿大学 邮编：08544

⁷蒙特里尔高等商学院，蒙特里尔，QC H3T 2A7, 加拿大

⁸美国亚利桑那州立大学，亚利桑那州坦佩 85281

摘要

人工智能在药物发现领域的应用越来越广泛。然而，现有研究利用机器学习主要是利用分子的化学结构，却忽视了化学中大量的文本知识。结合文本知识能让我们实现新的药物设计目标，适应基于文本的指令并预测复杂的生物活性。在此，我们通过对比学习策略，联合学习分子的化学结构和文本描述，提出了一种多模态分子结构-文本模型--MoleculeSTM。为了训练 MoleculeSTM，我们构建了一个大型多模态数据集，即 PubChemSTM，其中包含 280,000 多个化学结构-文本对。为了证明 MoleculeSTM 的有效性和实用性，我们设计了两个基于文本指示的具有挑战性的零点任务，包括结构-文本检索和分子编辑。MoleculeSTM 有两个主要特性：开放词汇和自然语言合成。在实验中，MoleculeSTM 在各种基准中对新的生化概念的泛化能力都达到了最先进的水平。

人工智能（AI）的最新进展有望为药物发现带来变革[1]。人工智能方法已被用于增强和加速当前的计算管道[2, 3, 4]，包括但不限于虚拟筛选[5, 6]、代谢特性预测[7, 8, 9]以及有针对性的化学结构生成和编辑[10, 11, 12, 13]。

现有的机器学习（ML）方法主要侧重于通过一维描述[14]、二维分子图[7, 15, 8]或三维几何结构[16, 17, 18]对分子的化学结构进行建模。它们还使用监督信号，如毒性标签、量子力学特性和结合亲和力测量。然而，这种有监督的设置需要对预先确定的标签类别进行昂贵的注释，从而阻碍了对未知类别和任务的应用[19]。为了克服这一问题，有人提出了在大规模数据库[20]上进行无监督预训练的方法，其主要优点是能够通过重建被遮蔽的拓扑[21]或几何[22]子结构，在没有监督注释的情况下学习化学结构。与有监督设置相比，虽然此类预训练模型[21, 22]已被证明能通过在少数标注示例上进行微调，更有效地泛化到各种下游任务中，但在没有此类标注示例或微调的情况下泛化未见类别和任务（即 ML 中所谓的 *zero-shot* 设置[23]），仍然是一项公开挑战。此外，现有的分子预训练方法大多只包含化学结构，对多模态表征的探索较少。

我们拥有大量人类可理解且易于获取的文本数据。目前，大规模的图像和视频多模态模型正在利用这些数据[24, 25, 26, 27]。自然语言界面是实现开放词汇和任务描述的直观方法。经过预训练的多模态模型可以很好地泛化到新的类别和任务中，即使是在零拍摄的情况下也是如此[24, 25, 26, 27]。它们还能让代理以交互方式学习解决新任务和探索新环境[28, 29]。我们相信，通过结合文献中大量的文本知识，分子模型也能获得类似的能力。

之前的研究 [30] 尝试利用文本知识来学习分子表征。不过，它只支持使用一维描述（简化分子输入行输入系统或 SMILES）建模，并在小规模数据集（10K 结构-文本对）上学习化学结构和文本描述。此外，它将两种模式统一到一个语言建模框架中，并需要对齐数据（即每个样本的化学结构和文本）进行训练。

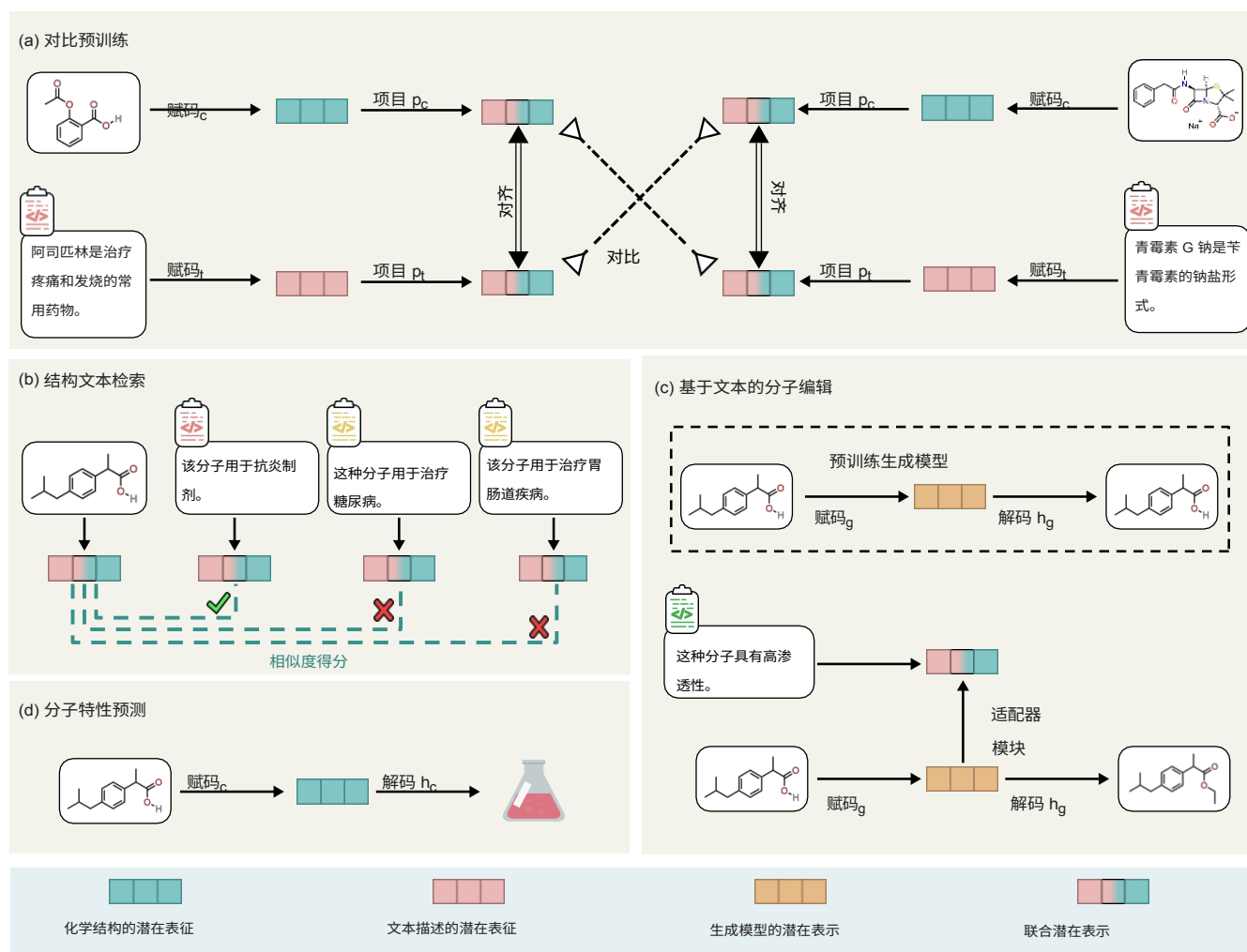


图 1.预训练和下游任务的流程。(a) MoleculeSTM 预训练有两个分支：化学结构（绿色）和文本描述（粉红色）。(b) 结构-文本检索下游任务。(c) 基于文本的分子编辑下游任务。(d) 分子性质预测下游任务。

因此，它无法采用现有的功能强大的预训练模型，而且对齐数据的可用性极为有限。

我们的方法我们为分子理解设计了一个多模式基础模型，该模型结合了分子结构信息和文本知识。我们利用基于文本的指令，对新药设计目标和新的复杂生物活性预测进行了零误差泛化演示，而无需标记示例或微调。我们提出的 MoleculeSTM 包含两个分支：化学结构分支和文本描述分支、分别处理分子的内部结构和外部领域知识。这种分离式设计可以 MoleculeSTM 可与分别针对每种模态训练的强大现有模型（即分子结构模型 [11, 31] 和科学语言模型 [32]）集成。鉴于这些预先训练好的模型，MoleculeSTM 通过对比学习范式[31, 33]将这两个分支连接起来。

为了将这两个分支与 MoleculeSTM 相结合，我们从 Pub- Chem [34]中构建了一个名为 PubChemSTM 的结构-文本数据集，这是迄今为止社区中最大的多模态数据集（比现有数据集 [30] 大 28 倍）。在 PubChemSTM 中，每个化学结构都与文字描述配对，相应地说明了化学和物理特性或高级生物活性。由于 MoleculeSTM 是在大规模的结构-文本配对数据集上进行训练的，而此类文本数据包含开放式的化学信息，因此它可以以 "0-shot "的方式推广到各种下游任务中。

为了证明引入语言模态的优势，我们设计了两个极具挑战性的下游任务：结构-文本检索任务和基于文本的分子编辑任务，并将预训练的 MoleculeSTM 以零点扫描的方式应用于这些任务。通过对这些任务的研究，我们总结出了 MoleculeSTM 的两个主要特性：开放词汇和组合性。(1) 词汇量开放是指我们提出的 MoleculeSTM 并不局限于一组固定的、预先定义的分子相关文本描述，它可以通过非绑定词汇支持探索广泛的生化概念。

的属性。在药物发现流水线中，这种属性可用于先导优化任务中基于文本的分子编辑，以及药物再利用任务中的新型疾病-药物关系提取。(2) 组合性意味着我们可以将一个复杂的概念分解成几个简单的概念来表达。这可以应用于基于文本的多目标先导优化任务[35]，该任务的目标是同时生成满足多种特性的分子。从经验上看，与最先进的方法相比，MoleculeSTM 在 6 项零次检索任务（准确率最高提高 50%）和 20 项零次文本编辑任务（命中率最高提高 40%）中表现最佳。此外，在分子编辑任务中，目视检查显示 MoleculeSTM 可以成功检测到文本描述中隐含的关键结构。此外，我们还探讨了 MoleculeSTM 能否通过微调提高标准分子性质预测基准 [9] 的性能。我们的结果表明，MoleculeSTM 可以在标准分子性质预测基准[9]中获得最佳整体性能。在八项财产预测任务中，九条基线的性能表现。

成果

概述和前言

在本节中，我们首先将概述 MoleculeSTM。然后，我们将介绍如何对 MoleculeSTM 进行预训练，并将预训练后的 MoleculeSTM 应用于三种下游任务（图 1）。

概述。MoleculeSTM 包括两个分支：化学结构分支和文本描述分支（ x_c 和 x_t ）。化学结构分支说明分子中原子的排列。我们考虑了两种编码器 f_c ：SMILES 字符串上的 Transformer [36] 和二维分子图上的 GNNs [7, 8, 15]。文字描述分支提供了分子功能的高级描述，我们使用最近一项研究成果[37]中的语言模型作为编码器 f_t 。**预训练。**在这一设计中，MoleculeSTM 的目标是通过对比学习（contrastive learning）[31, 33]，使用两个投影器（ p_c 和 p_t ）将从两个分支提取的表征映射到一个联合空间。对比学习的基本思想是缩小同一分子的化学结构和文字描述对之间的表征距离，增大不同分子的化学结构和文字描述对之间的表征距离。具体来说，我们使用预训练的单模态检查点（single-modal checkpoints）初始化这两个分支编码器[11, 31, 32]，然后在收集的数据集 PubChemSTM 上进行端到端的对比预训练。具体来说，PubChemSTM 是根据 PubChem [34] 构建的。我们提取了带有文本描述字段的分子，从而得到了 281K 个化学结构和文本对。更多详情可参见补充 A.1。

下游任务设计的两个原则

我们想强调的是，对于这些下游任务，预训练 MoleculeSTM 中的语言模型揭示了分子建模和药物发现的某些吸引人的属性。我们总结了以下两个关键点。

开放词汇。语言的本质是开放词汇和自由形式[38]。大型语言模型已在各种艺术相关应用中证明了其通用能力 [24, 25, 26]，我们发现它也为药物发现任务提供有前景、有洞察力的观察结果。因此，我们的方法并不局限于一套固定的、预先定义的分子相关注释，而是可以支持探索具有非约束词汇的新生化概念。其中一个例子是药物再利用。假设我们有一种新疾病或蛋白质靶点功能的文字描述。在这种情况下，我们可以使用 MoleculeSTM 获取其与所有现有药物的相似性，并检索出排名最高的药物，这些药物可用于临床试验等后期阶段。另一个例子是基于文本的先导优化。我们使用自然语言来描绘一种全新的特性，这种特性可以在优化后生成的分子中得到体现。

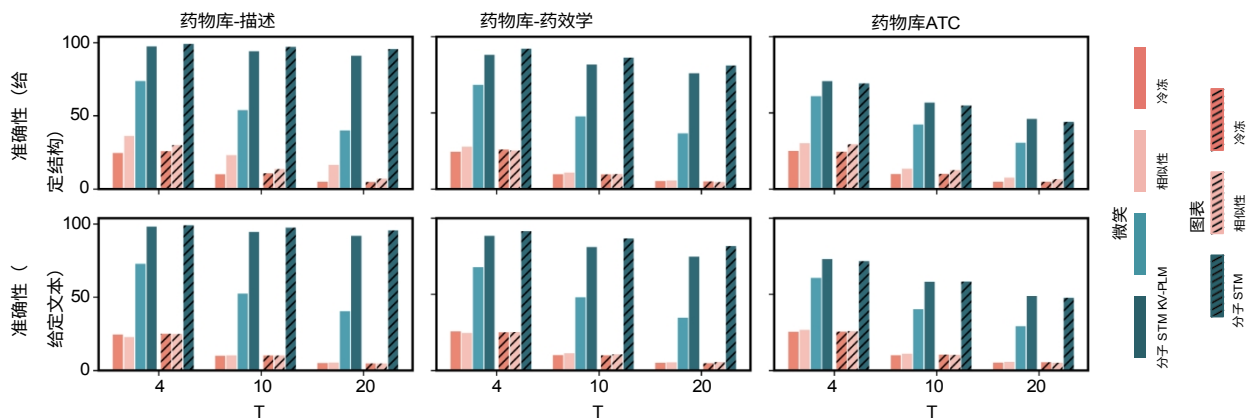
组合性。另一个属性是组合性。在自然语言中，复杂的概念可以通过将其分解成简单的概念来表达。这对于某些特定领域的任务至关重要，例如多目标先导优化[35]，我们需要同时生成具有多种所需属性的分子。现有的解决方案有两种：（1）针对每种所需的特性学习一个分类器，然后在一个大的候选库中进行筛选[10]；或者（2）优化

检索数据库，修改分子以实现多目标目标[12]。其主要局限性在于，成功率在很大程度上取决于训练分类器或检索数据库的标注数据的可用性。对于 MoleculeSTM 中的语言模型，我们提供了另一种解决方案。我们首先制作一个自然文本，称为文本提示，作为任务描述。文本提示可以是多目标的，包括每个属性的描述（例如 "分子可溶于水且具有高渗透性"）。有了化学结构和文本描述之间的预训练联合空间，MoleculeSTM 可以将分子属性组成问题转化为语言组成问题，而使用语言模型则更容易解决这个问题。

下游：零镜头结构文本检索

实验为了进行零点检索，我们从 DrugBank [39] 中构建了三个数据集。DrugBank 是迄今为止最全面的类药物分子数据库。在这里，我们提取了 DrugBank 中的三个字段：描述字段、药效学字段和解剖治疗化学（ATC）字段。这些字段说明了目标生物体的化学特性和药物作用。然后，检索任务可视为一个 T 选一的多选题，其中

(a) 结构文本检索结果



(b) 来自 DrugBank-ATC 的药物再利用案例研究

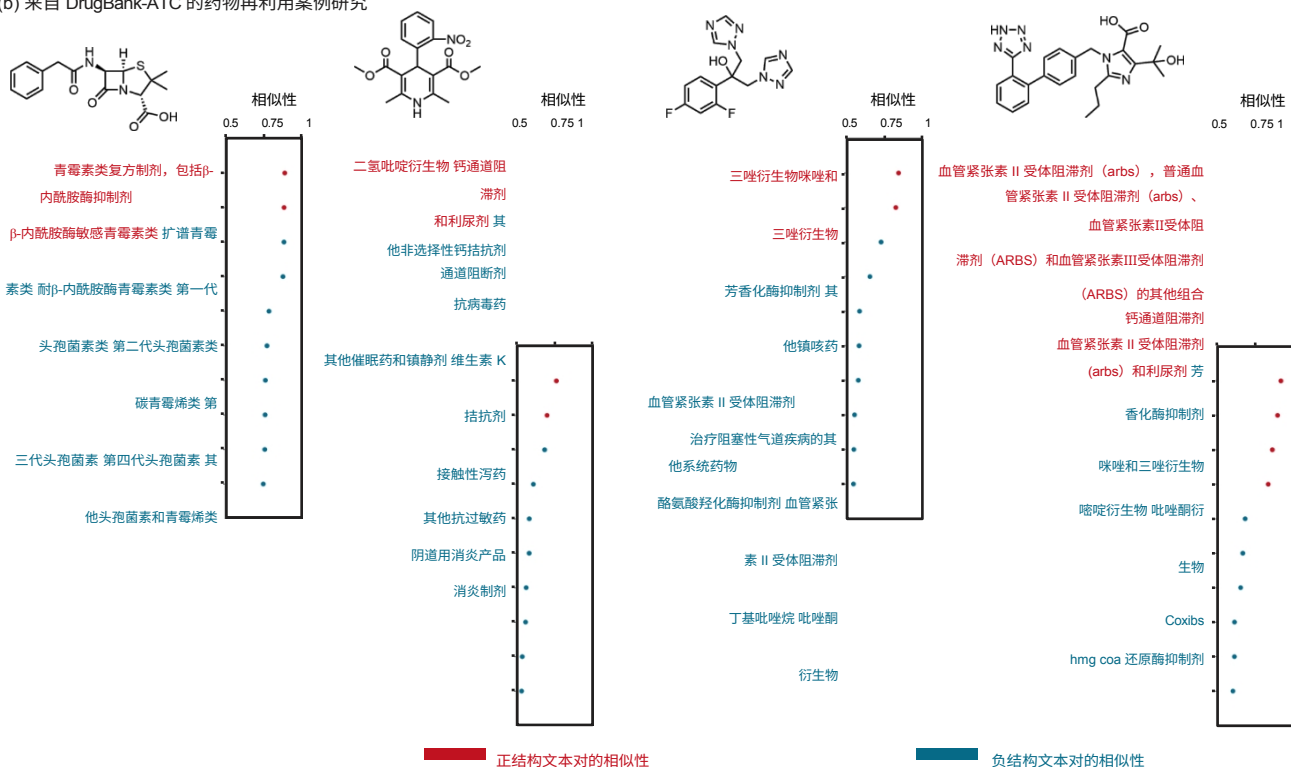


图 2. 零镜头结构-文本检索结果。(a) 在三个 DrugBank 数据集上进行零镜头结构-文本检索的准确率。(b) DrugBank-ATC 检索的四个案例研究。HMG-CoA 是 β -羟基 β -甲基戊二酰-CoA。

T 是选择的数量。具体来说，我们有两种设置：(1) 给定化学结构检索文字描述；(2) 给定文字描述检索化学结构。检索准确率被用作评估指标。**基线。**我们首先考虑使用预训练的单模态编码器 [11, 31, 32] 的两种基线。(1) **冻结**是指我们使用两个分支的预训练编码器和两个随机初始化的投影器。(2) **相似性**是指我们只从单个分支中提取相似性。例如，在第一种情况下，当给定化学结构时，我们从 PubChemSTM 中检索最相似的化学结构，然后将 PubChemSTM 中相应的成对文本表示作为代理表示。在此基础上，我们可以计算代理表示法与 T 个请求文本表示法之间的相似度得分。(3) 我们进一步考虑第三条基线，即针对知识渊博、多才多艺的人的预训练语言模型。

对 SMILES 文本对进行机器阅读 (KV-PLM) [30]。

结果 零镜头检索结果如图 2 (a) 所示。首先，我们发现所有算法的精确度在两种设置下都非常相似。然后，正如预期的那样，我们发现基线 Frozen 的表现并不比随机猜测好，因为投影仪是随机初始化的。相似性基线的表现略好于偶然性，这验证了预训练的单一模态确实能学习语义信息，但却不能在不同模态之间很好地泛化。另一方面，KV-PLM 可以从 SMILES 文本对中学习到有语义意义的信息，因此在三个数据集上取得了更高的准确率。就 MoleculeSTM 而言，在描述和药效学方面，来自 GNN 的图表示法比来自转换器模型的 SMILES 表示法具有更高的准确率；然而，在三个数据集和两种设置上，这两种方法都远远优于所有其他方法。例如，与 $T = 20$ 的最佳基线相比，准确率分别提高了约 50%、40% 和 15%。如此大的改进差距验证了

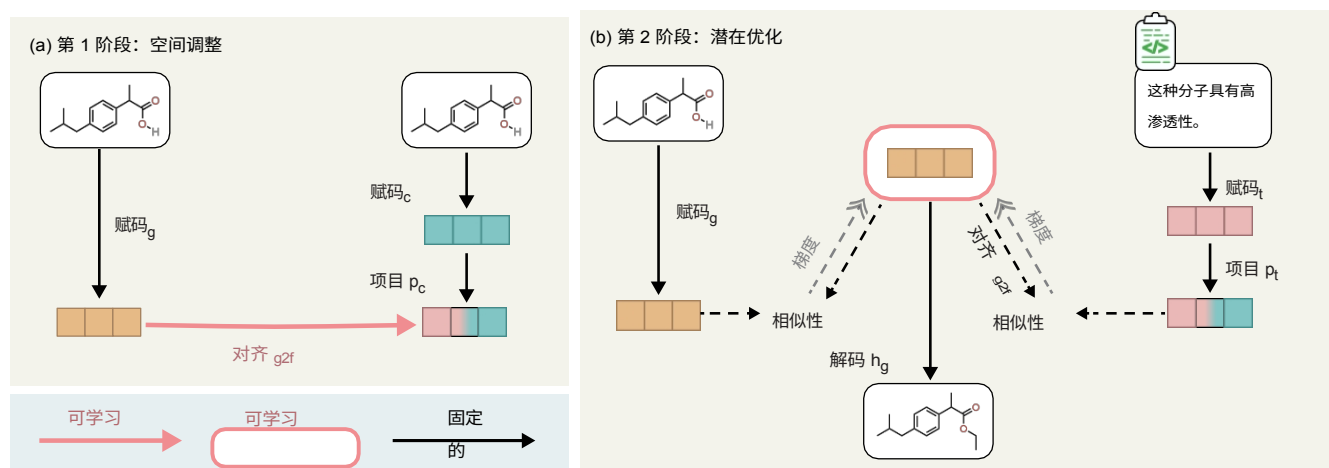


图 3.基于零镜头文本的分子编辑流程。(a) 空间对齐步骤将预训练的分子生成模型的表示空间与 MoleculeSTM 的表示空间对齐。(b) 潜在优化步骤学习与输入分子和文本描述相似的潜在表征。

认为 MoleculeSTM 可以在理解和连接两种分子模式方面发挥更好的作用。

药物再利用分析案例研究。在图 2 (b) 中，我们进一步展示了关于 ATC 检索质量的四个案例研究。具体来说，在给定分子化学结构的情况下，我们从 600 个分子中提取 10 个最相似的 ATC 标签。据观察，MoleculeSTM 可以检索到排名靠前的真实 ATC 标签。

下游：基于文本的零镜头分子编辑

实验。在分子编辑方面，我们从 ZINC [20] 中随机抽取 200 个分子和一个文本提示作为输入。文本提示分为四类：(1) **单目标编辑**是指使用与药物相关的单一属性进行编辑的文本提示，如“溶解度高的分子”和“更像药物的分子”。(2) **多目标（组合性）编辑**是同时应用多个属性进行编辑的文本提示，如“分子具有高溶解性和高渗透性”。(3) **基于结合亲和力的编辑**是化验描述的文本提示，其中每个化验对应一个结合亲和力任务。一个具体的例子是 ChEMBL 1613777 [40]，其提示为“该分子在酶蛋白的抑制剂和底物的检测中呈阳性。它利用分子氧将一个氧原子插入底物，并将第二个氧原子还原成水分子。输出分子应具有更高的结合亲和力分数”。(4) **药物相关性编辑**是通过文本提示使分子在结构上与某些常见药物相似，例如“这个分子看起来像青霉素”。我们希望输出的分子比输入的药物与目标药物更相似。有关文本提示的更多详细说明，请查阅补充 D。评价标准是满意命中率，如果输出与输入之间的度量差异超过阈值 Δ ，则为命中。 Δ 值取决于具体任务，我们考虑了两种典型情况： $\Delta = 0$ 表示条件宽松，而 $\Delta > 0$ 则表示条件严格，具有较大的积极影响。我们在图 3 中提供了算法流水线，更多详情请参见方法部分。

基线。我们考虑了四种基线。前三种基线 [13] 修改了输入分子的表征，然后对分子空间进行解码。**随机**是指我们将随机噪声作为输入分子表征的扰动。**PCA**是指我们将特征向量作为潜在方向，特征向量是在使用原理成分分析 (PCA) 对输入分子的潜在表示进行分解后得到的。**高方差**是指我们将方差最大的潜在表示维度作为编辑的语义方向，并对其进行单次编码。此外，我们还考虑了直接修改分子空间的基线--**遗传搜索 (GS)**。它是图遗传算法 [41] 的一种变体，不同之处在于，GS 是随机搜索，而不是由奖励函数引导的搜索，因为在零次搜索设置中没有检索数据库。

结果首先，我们在图 4 中提供了四种编辑任务类型中 20 个编辑任务的定量结果。实证结果表明，在所有 20 个任务中，MoleculeSTM 的满意命中率是最好的。这证明，对于 SMILES 和分子图编码器，MoleculeSTM 都能更好地理解自然语言的语义，从而探索出具有所需属性的输出分子。接下来，我们在图 5 中对输出分子的质量进行了详细分析，具体如下。

单目标分子编辑的可视化分析。我们利用单目标特性对输入和输出分子之间的差异进行可视化分析。典型的修改是添加、移除和替换分子的官能团或核心。例如，图 5(a)和(b)显示了对同一分子的两种不同编辑，分别导致

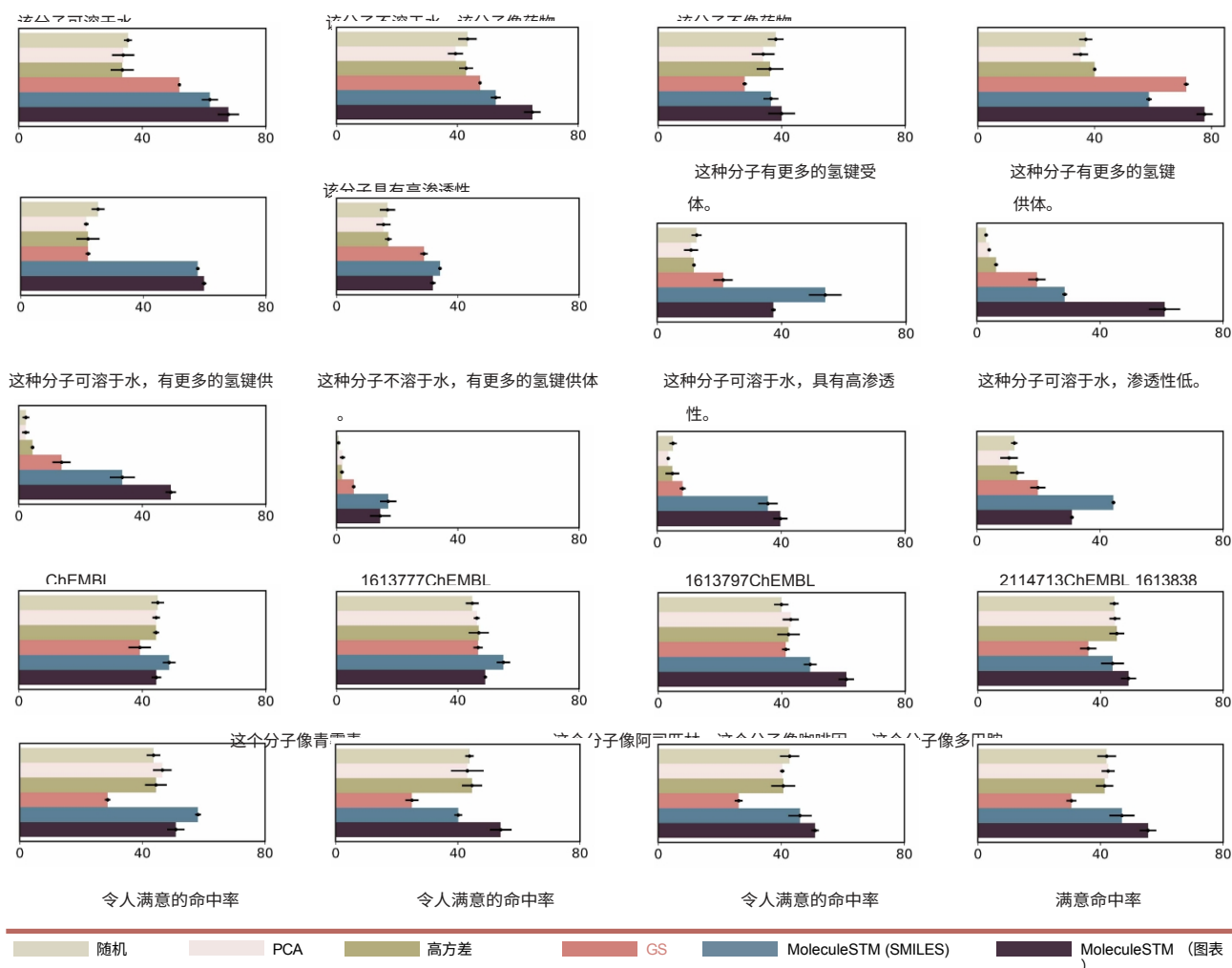


图 4.基于零镜头文本的分子编辑的可视化结果。四种基于文本编辑任务的满意命中率 (%)：八种单目标、四种多目标、四种基于 ChEMBL 结合亲和力的编辑任务

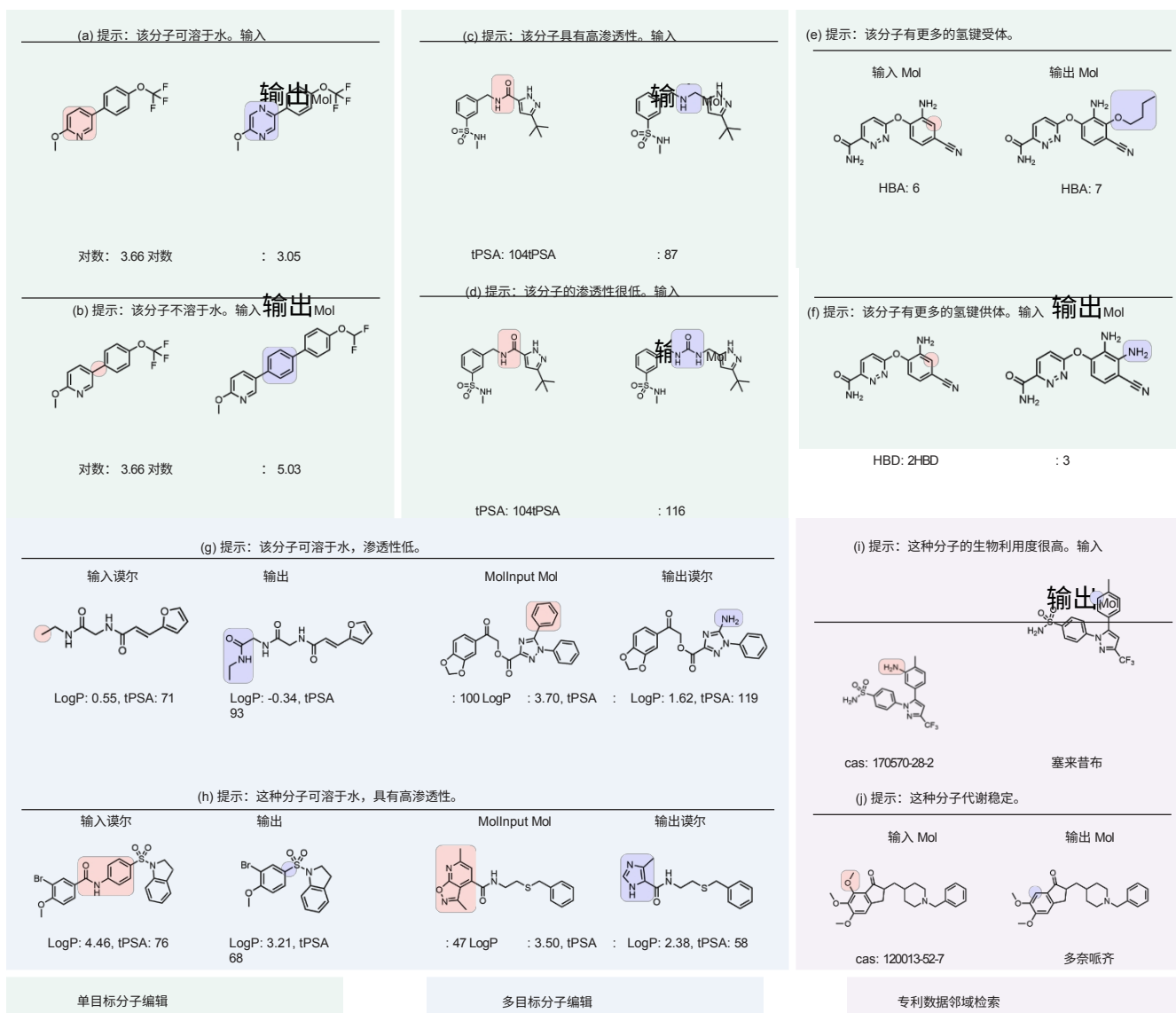
(在此基础上，研究人员还进行了四项药物相关性编辑任务（预训练的随机森林作为评价器，详细的文本提示见补充 D）和四项药物相关性编辑任务。所有可视化结果的满意阈值 (Δ) 均为 0。每个任务运行三个随机种子，每个误差条的长度代表标准偏差。

根据文本提示，溶解度的变化方向相反。将吡啶替换为吡嗪核心可提高溶解度，而插入苯连接则会产生不溶解的分子。在图 5(c)和(d)中，将酰胺连接改为烷基胺和脲分别会使编辑后的分子具有更高和更低的渗透性。最后，图 5(e)和(f)在分子的准确位置添加了丁基醚和伯胺，分别带来了更多的氢键受体和供体。

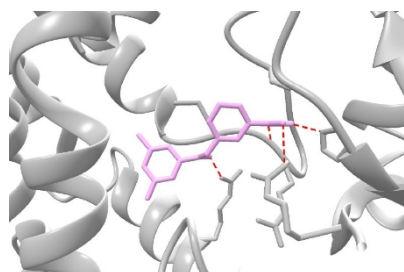
多目标分子编辑的可视化分析。我们进一步分析了多目标（成分）特性编辑。当在分子中引入极性基团并去除亲油性碳氢化合物时，水溶性的提高和渗透性的降低是一致的，如图 5 (g) 中用酰胺或伯胺取代甲基或苯基。不过，如果去除极性官能团或减少极性官能团和疏水成分的数量，则可获得更高的溶解度和渗透性。例如，在图 5 (h) 中，左边的酰胺和苯连接都被去除，右边的[1,2]恶唑并[5,4-b]吡啶取代基被极性表面较小的水溶性咪唑取代。**专利药物分子邻域搜索案例研究。**在药物发现过程中，改善先导分子的类药物特性对于找到候选药物至关重要[35]。在此

，我们展示了两个根据文本提示解决专利类似物的性质缺陷，从而从其专利类似物中生成已获批准药物的实例。图 5 (i) 从其氨基取代衍生物生成塞来昔布[42]，去掉氨基后，分子的肠道渗透性增强，生物利用度提高[43]。在图 5 (j) 中，多奈哌齐（Donepezil）中的三甲氧基苯分子是一种富电子的芳香族化合物，已知会进行氧化 I 期代谢[44]。为代谢稳定的分子。

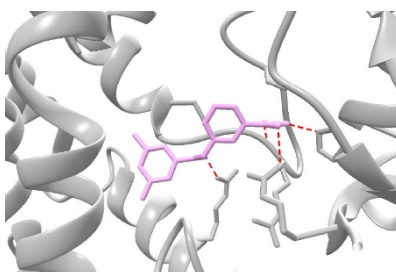
总之，我们在四种类型和 20 个基于文本的分子编辑任务上进行了丰富的实验，其中令人满意的有



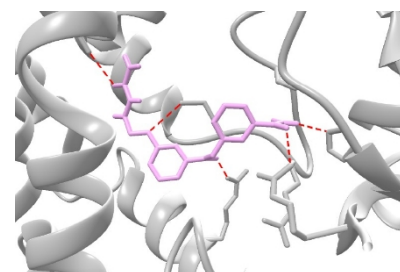
(k) ChEMBL1613777 (UniProt P33261) 的对接可视化



输入分子 (对接得分: -9.055)



带 GS 的输出分子 (对接得分: -8.843)



用 MoleculeSTM 输出分子 (对接得分: -10.35)

图 5. 基于文本的分子编辑可视化分析。溶解度编辑 (a,b)、渗透性编辑 (c,d)、受体和供体编辑 (e,f)、溶解度和渗透性编辑 (g,h) 以及专利数据邻域搜索 (i,j) 的案例研究。粉色和蓝色区域标出了编辑前后的官能团，并列出了化学文摘服务 (CAS) 登记号。(k) 显示基于结合亲和力的编辑，红色虚线标记潜在的结合。

MoleculeSTM 的命中率优于基准方法。此外，我们的编辑结果也符合基于化学领域知识的预期结果。定量和定性结果都表明，MoleculeSTM 可以学习对领域应用有用的语义信息，这激励我们在未来利用 MoleculeSTM 探索更具挑战性的任务。

下游：分子特性预测

实验。 MoleculeSTM 的一个优势是预训练的化学结构表征与外部领域知识共享信息，这种隐含的偏差有利于性质预测任务。与之前的分子预训练工作[21, 31]类似，我们采用了 MoleculeNet 基准[9]。它包含八个单模态二元分类数据集，用于评估预训练分子表示方法的表达能力。评估

方法		BBBP \uparrow	Tox21 \uparrow	ToxCast \uparrow	Sider \uparrow	ClinTox \uparrow	MUV \uparrow	艾滋病病毒 \uparrow	Bace \uparrow	平均值 \uparrow
SMILES	- 随机初始化	66.54 \pm 0.95	71.18 \pm 0.67	61.16 \pm 1.15	58.31 \pm 0.78	88.11 \pm 0.70	62.74 \pm 1.57	70.32 \pm 1.51	80.02 \pm 1.66	69.80
	MegaMolBART	68.89 \pm 0.17	73.89 \pm 0.67	63.32 \pm 0.79	59.52 \pm 1.79	78.12 \pm 4.62	61.51 \pm 2.75	71.04 \pm 1.70	82.46 \pm 0.84	69.84
	KV-PLM	70.50 \pm 0.54	72.12 \pm 1.02	55.03 \pm 1.65	59.83 \pm 0.56	89.17 \pm 2.73	54.63 \pm 4.81	65.40 \pm 1.69	78.50 \pm 2.73	68.15
	分子 STM	70.75 \pm 1.90	75.71 \pm 0.89	65.17 \pm 0.37	63.70 \pm 0.81	86.60 \pm 2.28	65.69 \pm 1.46	77.02 \pm 0.44	81.99 \pm 0.41	73.33
图	- 随机初始化	63.90 \pm 2.25	75.06 \pm 0.24	64.64 \pm 0.76	56.63 \pm 2.26	79.86 \pm 7.23	70.43 \pm 1.83	76.23 \pm 0.80	73.14 \pm 5.28	69.99
	AttrMask	67.79 \pm 2.60	75.00 \pm 0.20	63.57 \pm 0.81	58.05 \pm 1.17	75.44 \pm 8.75	73.76 \pm 1.22	75.44 \pm 0.45	80.28 \pm 0.04	71.17
	上下文预设	63.13 \pm 3.48	74.29 \pm 0.23	61.58 \pm 0.50	60.26 \pm 0.77	80.34 \pm 3.79	71.36 \pm 1.44	70.67 \pm 3.56	78.75 \pm 0.35	70.05
	信息图表	64.84 \pm 0.55	76.24 \pm 0.37	62.68 \pm 0.65	59.15 \pm 0.63	76.51 \pm 7.83	72.97 \pm 3.61	70.20 \pm 2.41	77.64 \pm 2.04	70.03
	MolCLR	67.79 \pm 0.52	75.55 \pm 0.43	64.58 \pm 0.07	58.66 \pm 0.12	84.22 \pm 1.47	72.76 \pm 0.73	75.88 \pm 0.24	71.14 \pm 1.21	71.32
	GraphMVP	68.11 \pm 1.36	77.06 \pm 0.35	65.11 \pm 0.27	60.64 \pm 0.13	84.46 \pm 3.10	74.38 \pm 2.00	77.74 \pm 2.51	80.48 \pm 2.68	73.50
	分子 STM	69.98 \pm 0.52	76.91 \pm 0.51	65.05 \pm 0.39	60.96 \pm 1.05	92.53 \pm 1.07	73.40 \pm 2.90	76.93 \pm 1.84	80.77 \pm 1.34	74.57

表 1.八项 MoleculeNet 二进制分类任务的结果。报告了在三个随机种子上测试 ROC-AUC 的平均值和标准偏差。

指标是接收者操作特征曲线下面积 (ROC-AUC) [45]。

基线。我们考虑了两种类型的化学结构：SMILES 字符串和分子图。对于 SMILES 字符串，我们采用了三种基线：随机初始化模型和两种预训练语言模型 (MegaMolBART [11] 和 KV-PLM [30])。对于分子图，除了随机初始化外，我们还将五种基于预训练的方法作为基线：AttrMasking [21]、ContextPred [21]、InfoGraph [46]、MolCLR [47] 和 GraphMVP [8]。

结果如表 1 所示，我们首先发现，与随机初始化方法相比，基于预训练的方法提高了整体分类准确率。与三种基线方法相比，在 SMILES 字符串上的 MoleculeSTM 在八项任务中的六项上都取得了一致的改进。分子图上的 MoleculeSTM 在八项任务中的四项任务中表现最佳，而在其他四项任务中的表现与最佳基线相当。在这两种情况下，MoleculeSTM 的总体性能（即所有八项任务的平均值）都是所有方法中最好的。

讨论

在这项工作中，我们提出了一个多模态模型 MoleculeSTM，以说明结合文本描述进行分子表征学习的有效性。与现有方法相比，我们证实了 MoleculeSTM 在两个新提出的零点任务和一个标准性质预测基准上的持续改进性能。此外，我们还观察到，MoleculeSTM 可以检索新的药物-靶标关系，并成功修改分子子结构以获得所需的特性。这些功能可加速各种下游药物发现实践，如再利用和多目标先导优化。此外，这些下游任务的结果与化学专家的反馈一致，反映了 MoleculeSTM 的领域知识探索能力。

这项工作的局限性之一是数据不足。尽管 PubChemSTM 比现有工作中使用的数据集大 28 倍，但它还可以进一步改进，未来可能需要整个社区的支持。这项工作的第二个瓶颈是化学结构模型的表现力，包括 SMILES 编码器、GNN 编码器和基于 SMILES 的分子生成模型。开发更具表现力的体系结构与这项工作息息相关，而且可以适用于我们的多模态预训练框架。

对于未来的发展方向，我们希望将 MoleculeSTM 从化学信息学扩展到具有更丰富文本信息的生物信息学任务。这使我们能够考虑基于结构的药物设计问题，如蛋白质配体结合和片段设计。此外，三维几何信息对于小分子和聚合物来说变得更加重要，因此也可以并入我们的基础模型中。最后但并非最不重要的一点是，文本描述的标记化可能需要额外的努力。某些任务拥有丰富的术语（例如 DrugBank-ATC 中的 ATC 代码），因此整体性能会受到影响。此类基本问题应谨慎处理。

方法

本节简要介绍预训练和下游任务中的某些模块。数据集构建、模型架构和超参数等详细说明见补充 A。

MoleculeSTM 预培训

数据集构建。在结构-文本预训练中，我们将 PubChem 数据库 [34] 作为数据源。PubChem 包含 1.12 亿个分子，是最大的分子公共数据库之一。PubChem 数据库有很多字段，之前的工作[30]使用同义词字段与学术论文语料库[48]进行匹配，得到了一个包含 10K 个结构-文本对的数据集。与此同时，PubChem 数据库中还有一个名为 "string "的字段，它具有更全面、更灵活的功能。

分子注释。我们利用这一领域构建了一个名为 PubChemSTM 的大型数据集，其中包括 25 万个分子和 28.1 万个结构-文本对。

此外，尽管 PubChemSTM 是最大的文本描述数据集，但与其他领域的同类数据集相比（如视觉语言领域的 4 亿数据集 [24]），其数据集规模相对较小。为了缓解这种数据不足的问题，我们采用了现有检查点的预训练模型，然后进行端到端的预训练，这将在下文中讨论。

化学结构分支 f_c 。这项工作考虑了两种类型的化学结构：SMILES 字符串将分子视为序列，二维分子图分别将原子和化学键作为节点和边。然后，基于化学结构，我们应用深度学习编码器 f_c 获得作为分子表示的潜向量。具体来说，对于 SMILES 字符串，我们采用 MegaMolBART [11] 的编码器，该编码器已在 ZINC 数据库 [49] 的 5 亿个分子上进行了预训练。对于分子图，我们使用 Graph-MVP 预训练 [31] 的预训练图同构网络（GIN） [15]。GraphMVP 正在对来自 GEOM 数据集 [50] 的 25 万个构象进行二维拓扑和三维几何之间的多视角预训练。因此，尽管我们没有明确利用三维几何图形，但最先进的预训练 GIN 模型可以隐含地编码此类信息。

文字描述分支 f_t 。文字描述分支提供了分子功能的高层次描述。我们可以将该分支视为加强分子表征的领域知识。这些领域知识采用自然语言形式，我们使用 BERT 模型 [37] 作为文本编码器 f_t 。我们进一步调整了预训练的 SciBERT [32]，该模型在化学和生物领域的文本数据上进行了预训练。

对比预训练。对于 MoleculeSTM 预训练，我们采用了对比学习策略，例如 EBM-NCE [31] 和 InfoNCE [33]。EBM-NCE 和 InfoNCE 对同一分子的结构-文本对进行对齐，同时对不同分子的结构-文本对进行对比。我们认为选择对比预训练方法是一个重要的超参数。EBM-NCE 和 InfoNCE 的目标是

$$\begin{aligned} \text{LEBM-NCE} &= - \mathbb{E}_{x, x_t} \log \sigma(E(x_c, x_t)) + \mathbb{E}_{x, x_t'} \log(1 - \sigma(E(x_c, x_t'))) + \mathbb{E}_{x, x_t} \log \sigma(E(x_c, x_t)) + \mathbb{E}_{x, x_t'} \log(1 - \sigma(E(x_t', x_t))) \\ \text{信息} &= - \mathbb{E}_{x, x_t} \log \frac{\exp(E(x_c, x_t))}{\exp(E(x_c, x_t)) + \sum_{x_t'} \exp(E(x_c, x_t'))} + \log \frac{\exp(E(x_c, x_t))}{\exp(E(x_c, x_t)) + \sum_{x_t'} \exp(E(x_c, x_t'))} \end{aligned} \quad (1)$$

其中， σ 是 sigmoid 激活函数， x_c 和 x_t 构成每个分子的结构-文本对， x_c' 和 x_t' 是随机从噪声分布中采样的负样本，我们使用经验数据分布。 $E(-)$ 是具有灵活表述的能量函数，我们使用联合学习空间上的点积，即 $E(x_c, x_t) = \langle p_c \circ f_c(x_c), p_t \circ f_t(x_t) \rangle$ ，其中 \circ 是函数组成。

下游：零镜头结构文本检索

给定一个化学结构和 T 个文本描述，检索任务就是根据联合表示空间计算的得分，选择与化学结构相似度最高的文本描述（反之亦然）。这对特定的药物发现任务很有吸引力，例如药物再利用或适应症扩展 [30, 51]。我们要强调的是，预训练模型是在零点场景下用于检索的，也就是说，这一检索任务不需要对模型进行优化。现有研究 [52] 发现了一个潜在的问题，即仅仅利用化学结构是不够的，而 MoleculeSTM 通过采用文本描述和利用分子的高级功能实现了一种新的视角。

在这种 "零镜头" 任务设置中，所有编码器 (f_c, f_t) 和投影器 (p_c, p_t) 都是通过 MoleculeSTM 预先训练的，并在下游任务中保持冻结。设置 (1) 的检索任务示例如下

$$\text{Retrieval}(x_c) = \arg\max_{x_t} p_c \circ f_c(x_c), p_t \circ f_t(x_t) \quad x_t \in T \text{ 文本描述} \quad (2)$$

下游：基于文本的零镜头分子编辑

分子编辑任务的目的是修改分子的化学结构，如官能团变化 [53] 和支架跳跃 [54, 55]。传统的分子编辑方法高度依赖领域专家，可能存在主观性或偏差 [56, 57]。ML 方法为解决这一问题提供了另一种策略。给定一个固定的预训练分子生成模型（编码器 f_g 和解码器 h_g ），ML 编辑方法在潜表征（或潜代码）空间上学习一个有语义意义的方向。然后，解码器 h_g 沿着该方向移动，生成具有所需属性的输出分子。在 MoleculeSTM 中，有了预训练的联合表征空间，我们就可以通过零点注入文本描述的方式完成这项任务。如图 3（a、b）所示，我们需要两个阶段。第一阶段是空间对齐，我们要训练一个适配器模块，将生成模型的表示空间与 MoleculeSTM 的联合表示空间对齐。第二阶段是潜码优化，在这一阶段，我们使用两个相似性模块直接学习潜码。

分数作为目标函数。最后，对优化后的潜码进行解码，就可以得到输出的分子。请注意，在这一编辑过程中，MoleculeSTM (f_c, p_c, f_t, p_t) 和预训练的分子生成模型 (f_g, h_g) 都被冻结。

第 1 阶段：空间对齐。在这一阶段，目标是学习一个适配器模块，将生成模型的表示空间与 MoleculeSTM 的联合表示空间对齐。根据高斯分布，目标函数为

$$L = m \|g_{2f} \circ f_g(x_c) - p_c \circ f_c(x)\|_d^2, \quad (3)$$

其中， \circ 是函数组成函数， $m_{g_{2f}}$ 是为对齐两个潜空间而优化的适配器模块。

第二阶段：潜在优化。在这一阶段，给定输入分子 $x_{c,in}$ 和文本提示 x_t ，目标是直接优化潜码 w 。最优的 w 应同时接近 $x_{c,in}$ 和 x_t 的表示，如图所示：

$$w = \underset{w \in W}{\operatorname{argmin}} -L_{\text{cosine-sim}}(m_{g_{2f}}(w), p_t \circ f_t(x)_t) + \lambda - L_{l_2}(w, f_g(x) \circ c_{in}) \quad (4)$$

其中， W 是潜码空间， $L_{\text{cosine-sim}}$ 是余弦相似度， L_{l_2} 是 l_2 距离， λ 是平衡这两个相似度的系数。最后，在优化潜码 w 之后，我们将使用预训练生成模型的解码器进行解码，从而得到输出分子： $x_{c,out} = h_g(w)$ 。

评估。评价指标是满意命中率。假设我们有一个输入分子 $x_{c,in}$ 和一个文本提示 x_t ，编辑算法将生成一个输出分子 $x_{c,out}$ 。然后，我们用命中率来衡量输出分子是否满足文本提示中的条件。

$$\text{hit}(x_{c,in}, x_t) = \begin{cases} 1, & \exists \lambda, \text{ s.t. } x_{c,out} = h_g(w; \lambda) \wedge \text{满足}(x_{c,in}, x_{c,out}, x_t) \\ 0, & \text{否则} \end{cases}, \quad \text{hit}(t) = \frac{\sum_{i=1}^N \text{命中}_{c,t}^i}{N} \quad (5)$$

其中， N 是编辑输出的总数， $\text{satisfy}(-)$ 是满足条件。它与具体任务相关，下面我们列出了五个关键点。(1) 对于基于属性的单目标编辑，我们使用分配系数对数 (LogP)、药物相似性定量估计 (QED) 和拓扑极性表面积 (tPSA) 分别作为衡量分子溶解度[58]、药物相似性[59]和渗透性[60]的代用指标。氢键受体 (HBA) 和氢键供体 (HBD) 的数量是明确计算出来的。一旦输入分子和输出分子之间的测量差值超过一定的阈值 Δ ，则成功命中。(2) 对于基于多目标属性的编辑，我们输入描述多个属性组成的文本提示。 Δ 由每个属性的阈值组成，成功的命中需要同时满足所有属性。(3) 对于基于结合亲和性的编辑，我们利用 ChEMBL 的地面实况数据来训练二元分类器，并测试输出分子是否比输入分子具有更高的置信度， Δ 被固定为 0。(4) 对于药物相关性编辑，我们使用谷本相似性来量化结构相似性[61]。如果输出分子与目标药物之间的相似度得分高于输入分子与目标药物之间的相似度阈值 Δ ，则为命中。(5) 此外，满意度阈值 Δ 的选择也是针对具体任务的，其值越高，满意度条件就越严格。关于阈值的详细信息，请参见补充 D。

下游：分子特性预测

在建模方面，我们采用预训练编码器 f_c 并添加预测头 h_c 来预测分类值或标量值分子特性，如结合亲和力或毒性。 f_c 和 h_c 都经过优化，以适应目标特性， ∇ 微调方式 [21, 31]。

数据可用性

所有数据集都在[这个 "拥抱的脸" 链接](#)中提供。具体到 PubChemSTM 的发布，我们在文本数据许可证方面遇到了很大的挑战。经与 PubChem 集团确认，对这些数据进行研究并不违反他们的许可证；然而，PubChem 并不拥有文本

数据的许可证，这就需要对 PubChemSTM 中 280 个结构-文本对中的每一个进行广泛的许可证评估。这阻碍了 PubChemSTM 的发布。尽管如此，我们还是(1) 在补充 A.1 中描述了详细的预处理步骤，(2) 在 PubChemSTM 中提供了[带有 CID 文件的分子](#)，(3) 还提供了详细的[预处理脚本](#)。利用这些脚本，用户可以轻松地重建 PubChemSTM 数据集。

代码可用性

源代码可在 [GitHub 存储库](#) 和 Zenodo [62] 中找到。[这里](#) 提供了预训练和三个下游任务的脚本。预训练模型的检查点在这个[拥抱脸链接](#) 中提供。除了上述方法外，为了帮助用户试用我们的 MoleculeSTM 模型，本版本还包括[笔记本中的演示](#)。此外，用户还可以通过查看[数据集文件夹](#) 来定制自己的数据集。

致谢

这项工作是刘胜超在英伟达研究院实习期间完成的。作者感谢 Michelle Lynn Gill、Abe Stern 以及 AIAIgo 和英伟达 Clara 团队的其他成员提出的宝贵意见。作者还要感谢来自 PubChem 的 Teresa Dierks、Evan Bolton、Paul Thiessen 等人在确认 PubChem 许可证方面提供的帮助。

作者投稿声明

S.L.、W.N.、C.W.、Z.Q.、C.X.和 A.A.构思并设计了实验。S.L. 进行了实验。S.L. 和 C.W. 分析了数据。S.L.、C.W.和J.L.提供了分析工具。S.L.、W.N.、C.W.、J.L.、Z.Q.、L.L.、J.T.、C.X.和A.A.撰写论文。J.T.、C.X.和A.A.对本项目的指导做出了同等贡献。

竞争利益声明

作者声明不存在利益冲突。

参考资料

- [1] 托马斯-沙利文"艰难的道路：开发一种新药的成本为 26 亿美元；进入临床开发阶段的药物批准率不到 12%"。见《政策与医学》（2019 年）：《政策与医学》（2019 年）。
- [2] Atanas Patronov, Kostas Papadopoulos 和 Ola Engkvist。"人工智能影响了药物发现吗？ In：《药物设计中的人工智能》。Springer, 2022, pp.
- [3] Madura KP Jayatunga, Wen Xie, Ludwig Ruder, Ulrik Schulze 和 Christoph Meier。"小分子药物发现中的人工智能：即将到来的浪潮"。 In： *Nat.Rev. Drug Discov* 21 (2022), pp.
- [4] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Židek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis."利用 AlphaFold 高精度预测蛋白质结构"。 In： *自然* 596.7873 (2021), 第 583-589 页。
- [5] Sebastian G. Rohrer 和 Knut Baumann。"基于 PubChem 生物活性数据的虚拟筛选最大无偏验证（MUV）数据集"。 In： *Journal of Chemical Information and Modeling* 49.2 (2009).PMID: 19161251, pp.DOI: [10.1021/ ci8002649](https://doi.org/10.1021/ci8002649). eprint: <https://doi.org/10.1021/ci8002649>.URL: <https://doi.org/10.1021/ci8002649>.
- [6] Shengchao Liu, Moayad Alnammi, Spencer S Ericksen, Andrew F Voter, Gene E Ananiev, James L Keck, F Michael Hoffmann, Scott A Wildman, and Anthony Gitter."前瞻性虚拟筛选的实用模型选择"。 In： *化学信息与建模杂志* 59.1 (2018), 第 282-293 页。
- [7] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, Ryan P Adams。"用于学习分子指纹的图卷积网络"。 In： *神经信息处理系统进展* 28 (2015)。
- [8] Shengchao Liu, Mehmet F Demirel, and Yingyu Liang."N-Gram Graph：图的简单无监督表示法及其在分子中的应用"。 In： *神经信息处理系统进展* 32 (2019)。
- [9] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay

- Pande."分子网：分子机器学习的基准"。In: *化学科学* 9.2 (2018), 第 513-530 页。
- [10] Wengong Jin, Regina Barzilay, and Tommi Jaakkola."利用结构主题分层生成分子图》。In: *国际机器学习会议*。PMLR.2020, 第 4839-4848 页。
- [11] Ross Irwin、Spyridon Dimitriadis、Jiazhen He 和 Esben Jannik Bjerrum。"Chemformer: a pre-trained transformer for computational chemistry".In: *机器学习：科学与技术* 3.1 (2022), 第 015022 页。
- [12] 王志超、聂伟力、乔卓然、肖超伟、理查德-巴兰纽克和阿尼玛-阿南德库马尔。"基于检索的可控分子生成》。见: *arXiv preprint arXiv:2208.11126* (2022)。
- [13] 刘胜超、王成鹏、聂伟力、王汉臣、陆家瑞、周伯磊和唐健。"GraphCG：无监督发现图中的可引导因子"。In: *NeurIPS 2022 研讨会：图学习的新前沿》*。2022.URL: https://openreview.net/forum?id=BhR44NzeK_1。
- [14] Mario Krenn、Florian Häse、AkshatKumar Nigam、Pascal Friederich 和 Alan Aspuru-Guzik。"自参照嵌入字符串（SELFIES）：100%稳健的分子字符串表示法"。In: *机器学习：科学与技术* 1.4 (2020), 第 045024 页。

- [15] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. "图神经网络有多强大?" In: *arXiv preprint arXiv:1810.00826* (2018).
- [16] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko 和 K-R Müller. "SchNet--分子和材料的深度学习架构"。In: *化学物理学报* 148.24 (2018), 第 241722 页。
- [17] Victor Garcia Satorras, Emiel Hoogeboom 和 Max Welling. "E (n) 等变图神经网络"。In: *arXiv preprint arXiv:2102.09844* (2021).
- [18] Kenneth Atz, Francesca Grisoni 和 Gisbert Schneider. "分子表征的几何深度学习"。In: *自然-机器学习* 3.12 (2021), 第 1023-1032 页。
- [19] 纪元 峰、张璐、吴嘉翔、吴秉哲、黄隆凯、徐廷扬、荣宇、李兰青、任杰、薛丁等: 《DrugOOD: Out-of-Distribution (OOD) Dataset Curator and Benchmark for AI-aided Drug Discovery-A Focus on Affinity Prediction Problems with Noise Annotations》。In: *arXiv preprint arXiv:2201.09637* (2022).
- [20] John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad 和 Ryan G Coleman. "ZINC: 发现生物学化学的免费工具"。In: *化学信息与建模杂志* 第 52.7 期 (2012 年), 第 1757-1768 页。
- [21] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. "预训练图神经网络的策略"。In: *学习表征国际会议, ICLR.2020*.
- [22] Shengchao Liu, Hongyu Guo, and Jian Tang. "利用se (3)-不变量去噪距离匹配进行分子几何预训练"。In: *arXiv preprint arXiv:2206.13602* (2022).
- [23] Hugo Larochelle, Dumitru Erhan 和 Yoshua Bengio. "新任务的零数据学习"。In: *AAAI.Vol.2*. 2008, p. 3.
- [24] 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》。In: *国际机器学习会议*. PMLR.2021, pp.
- [25] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever 和 Mark Chen. "Glide: 利用文本引导的扩散模型实现逼真图像生成和编辑"。In: *arXiv preprint arXiv:2112.10741* (2021).
- [26] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu 和 Mark Chen. "使用剪辑潜变量的分层文本条件图像生成"。In: *arXiv preprint arXiv:2204.06125* (2022).
- [27] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or 和 Dani Lischinski. "Styleclip: 文本驱动的样式表图像处理"。In: *IEEE/CVF 计算机视觉国际会议论文集*。2021年, 第2085-2094页。
- [28] 李爽、泽维尔-普伊格、杜一伦、克林顿-王、埃金-阿库雷克、安东尼奥-托拉尔巴、雅各布-安德烈亚斯和伊戈尔-莫尔达奇。"用于互动决策的预训练语言模型"。In: *arXiv preprint arXiv:2202.01771* (2022).
- [29] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. "Menedojo: 用互联网规模的知识构建开放式的具身代理"。In: *arXiv preprint arXiv:2206.08853* (2022).
- [30] 曾 哲妮、姚远、刘志远和孙茂松。"连接分子结构和生物医学文本的深度学习系统, 其理解能力可媲美人类专业人士"。In: *自然通讯* 13.1 (2022), 第 1-11 页。
- [31] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. "用三维几何预训练分子图表示法"。In: *学习表征国际会议*. 2022.URL: <https://openreview.net/forum?id=xQUelpOKPam>.
- [32] Iz Beltagy, Kyle Lo 和 Arman Cohan. "SciBERT: 科学文本的预训练语言模型"。In: *EMNLP.2019*. 电子版: [arXiv:1903.10676](https://arxiv.org/abs/1903.10676).

- [33] Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。"对比预测编码的表征学习"。In: *arXiv preprint arXiv:1807.03748* (2018).
- [34] Sunghwan Kim、Jie Chen、Tiejun Cheng、Asta Gindulyte、Jia He、Siqian He、Qingliang Li、Benjamin A Shoemaker、Paul A Thiessen、Bo Yu 等：《2021 年的 PubChem：新数据内容和改进的网络界面》。In: *Nucleic acids research* 49.D1 (2021), pp.
- [35] James P Hughes、Stephen Rees、S Barrett Kalindjian 和 Karen L Philpott。"早期药物发现的原则"。In: *British journal of pharmacology* 162.6 (2011), pp.
- [36] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。"关注就是一切"。In: *神经信息处理系统进展* 30 (2017)。
- [37] 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"。In: *arXiv preprint arXiv:1810.04805* (2018).
- [38] Xiuye Gu、Tsung-Yi Lin、Weicheng Kuo 和 Yin Cui。"通过视觉和语言知识提炼实现开放词汇对象检测"。In: *arXiv preprint arXiv:2104.13921* (2021).

- [39] David S Wishart、Yannick D Feunang、An C Guo、Elvis J Lo、Ana Marcu、Jason R Grant、Tanvir Sajed、Daniel Johnson、Carin Li、Zinat Sayeeda 等:《DrugBank 5.0: DrugBank 数据库 2018 年重大更新》。核酸研究 *核酸研究* 46.D1 (2018)、pp.D1074-D1082.
- [40] David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, María Gordillo-Marañón、菲奥娜-亨特、劳拉-琼科、格蕾丝-穆古姆巴特、米拉格罗斯-罗德里格斯-洛佩斯、弗朗西斯-阿特金森、尼古拉斯-博斯克、克里斯-J-拉杜、阿尔多-塞古拉-卡布雷拉、安妮-赫西和安德鲁-R-利奇。"ChEMBL: towards direct deposition of bioassay data".In: *Nucleic Acids Research* 47.D1 (Nov. 2018), pp.ISSN: 0305-1048.DOI: [10.1093/nar/gky1075](https://doi.org/10.1093/nar/gky1075). eprint: <https://academic.oup.com/nar/article-pdf/47/D1/D930/27437436/gky1075.pdf>。URL:<https://doi.org/10.1093/nar/gky1075>.
- [41] Jan H Jensen."探索化学空间的基于图的遗传算法和生成模型/蒙特卡洛树搜索"。In: *化学科学* 10.12 (2019), 第 3567-3572 页。
- [42] John J Talley、Thomas D Penning、Paul W Collins、Donald J Rogier Jr、James W Malecha、Julie Miyashiro、Stephen R Bertenshaw、Ish K Khanna、Matthew J Graneto、Roland S Rogers 等*用于治疗炎症的取代吡唑苯磺酰胺类药物*。美国专利 5,760,068.美国专利 5,760,068.
- [43] David Dahlgren 和 Hans Lennernäs。"肠道渗透性和药物吸收: 预测性实验、计算和体内方法"。In. *Pharmaceutics* 11.8 (2019) : *Pharmaceutics* 11.8 (2019).ISSN: 1999-4923.DOI: [10.3390/pharmaceutics11080411](https://doi.org/10.3390/pharmaceutics11080411).URL: <https://www.mdpi.com/1999-4923/11/8/411>.
- [44] Gordon Guroff、Jean Renson、Sidney Udenfriend、John W Daly、Donald M Jerina 和 Bernhard Witkop。"羟化诱导迁移: NIH 转变": 最近的实验揭示了芳香族化合物酶促羟基化的一个意想不到的普遍结果"。In: *科学* 157.3796 (1967), 第 1524-1530 页。
- [45] 安德鲁-P-布拉德利"在机器学习算法评估中使用 ROC 曲线下面积"。In: *模式识别* 30.7 (1997), 第 1145-1159 页。
- [46] 孙凡云、乔丹-霍夫曼、维卡斯-维尔马和唐健。"信息图: 通过互信息最大化实现无监督和半监督图表示学习"。In: *学习表征国际会议, ICLR.2020*.
- [47] 王宇阳、王建仁、曹中林和 Amir Barati Farimani。"Molclr: 通过图神经网络进行表征的分子对比学习"。In: *arXiv preprint arXiv:2102.10056* (2021).
- [48] Kyle Lo、Lucy Lu Wang、Mark Neumann、Rodney Kinney 和 Dan S Weld。"S2ORC: The semantic scholar open research corpus"。In: *arXiv preprint arXiv:1911.02782* (2019).
- [49] Teague Sterling 和 John J Irwin."ZINC 15--人人都能发现的配体"。In: *化学信息与建模杂志* 55.11 (2015), 第 2324-2337 页。
- [50] Simon Axelrod 和 Rafael Gomez-Bombarelli。"GEOM, 用于性质预测和分子生成的能量注释分子构象"。In: *科学数据* 9.1 (2022), 第 1-14 页。
- [51] Saurabh Aggarwal."癌症靶向疗法"。In: *自然评论. 药物发现* 9.6 (2010), 第 427 页。
- [53] Peter Ertl、Eva Altmann 和 Jeffrey M McKenna。"生物活性分子中最常见的官能团及其受欢迎程度的演变"。In: *药物化学杂志* 63.15 (2020), 第 8408-8418 页。
- [54] Hans-Joachim Böhm、Alexander Flohr 和 Martin Stahl。"脚手架跳跃"。In: *今日药物发现: 技术* 1.3 (2004)、pp.217-224.

- [55] Ye Hu、Dagmar Stumpfe 和 Jurgen Bajorath。"支架跳跃的最新进展：迷你透视"。In: *Journal of medicinal chemistry* 60.4 (2017), pp.
- [56] Jürgen Drews."药物发现：历史视角"。In. *Science* 287.5460 (2000), pp: *Science* 287.5460 (2000), pp.
- [57] 劳伦特-戈麦斯"药物化学决策：直觉的力量"。In. *ACS Medicinal Chemistry Letters* 9.10 (2018), pp: *ACS Medicinal Chemistry Letters* 9.10 (2018), pp.
- [58] Albert Leo、Corwin Hansch 和 David Elkins."分馏系数及其用途"。见《化学评论》71.6 (1971 年)，第 525-616 页：《化学评论》71.6 (1971 年)，第 525-616 页。DOI: [10.1021/cr60274a001](https://doi.org/10.1021/cr60274a001)。电子版：<https://doi.org/10.1021/cr60274a001>。URL: <https://doi.org/10.1021/cr60274a001>。
- [59] G Richard Bickerton、Gaia V Paolini、Jérémy Besnard、Sorel Muresan 和 Andrew L Hopkins。"量化药物的化学美"。In: *自然化学* 4.2 (2012)，第 90-98 页。
- [60] Peter Ertl、Bernhard Rohde 和 Paul Selzer。"基于片段贡献之和的分子极表面积快速计算及其在药物运输特性预测中的应用"。In: *药物化学杂志* 43.20 (2000).PMID: 11020286, pp.DOI: [10.1021/jm000942e](https://doi.org/10.1021/jm000942e). eprint: <https://doi.org/10.1021/jm000942e>.URL: <https://doi.org/10.1021/jm000942e>。

- [61] 达科-布蒂娜"基于日光指纹和谷本相似性的无监督数据库聚类：对小型和大型数据集进行聚类的快速自动方法"。In: *化学信息与计算机科学杂志* 39.4 (1999 年) , 第 747-750 页。
DOI: [10.1021/ci9803381](https://doi.org/10.1021/ci9803381). eprint: <https://doi.org/10.1021/ci9803381>. URL: <https://doi.org/10.1021/ci9803381>.
- [62] 刘胜超、聂伟力、王成鹏、陆家瑞、乔卓然、刘玲、唐健、肖超伟和阿尼玛-阿南德库马尔。"基于文本编辑和检索的多模态分子结构-文本模型"。In: (Aug. 2023). DOI: [10.5281/zenodo.8303265](https://doi.org/10.5281/zenodo.8303265).

补充信息

预培训

A.1 PubChemSTM 建设

我们构建了一个名为 PubChemSTM 的化学结构-文本对数据集，该数据集从 PubChem 数据库 [1] 中提取。下面我们将解释数据集构建的关键步骤。

1. 我们使用 [PUG View](#)（一种 REST 类型的网络服务）下载分子的文本描述。它总共有 290 页，每一页都以 XML 格式下载。[这里](#)有一个示例页面（第一页）供参考。XML 数据中有一个 "字符串" 字段，我们将其视为分子的文字描述。构建后，我们有 250K 个分子（具有唯一的 PubChem ID）和 281K 个化学结构-文本对。请注意，每个分子可以有来自不同资源的多个注释。
 - 大多数分子注释都以通用名称或国际纯粹与应用化学联合会（IUPAC）名称开头。我们既可以使用原始描述（带有通用名称或 IUPAC 名称），也可以用文本模板（如“该分子是……”）来替换。
 - 因此，我们构建了两个版本的 PubChemSTM 数据集：PubChemSTM-raw 和 PubChemSTM-extracted，分别对应于使用原始注释或用文本提示替换分子名称。除分子名称外，这两个版本的 PubChemSTM 共享分子。
2. 我们从 PubChem [FTP 服务](#) 下载 326 个 SDF 文件。每个 SDF 文件都包含一批分子的结构信息（如 SMILES 字符串和分子图）。
3. 我们使用 PubChem ID 对前两个步骤中的每个分子的注释和化学结构进行匹配，第一步中的大部分分子都包含 SDF 文件中的相应化学结构。具体来说，只有 12 个分子未能从 SDF 文件中找到有效的 SMILES，我们忽略了这些分子。
4. 通过以上三个步骤，最终将得到一个包含 281K 个结构-文本配对和 250K 个独特分子的数据集。请注意，PubChem 数据库[1]经常在线更新，上述数据收集于 2022 年 3 月。

预处理细节 PubChem 数据库中有一个名为 "名称" 的字段，其中包括每个分子的通用名称或 IUPAC 名称。请注意，对 IUPAC 进行标记化并非易事。因此，我们使用了两个版本来测试其效果，即 PubChemSTM 原始版本和 PubChemSTM 提取版本。我们发现，PubChemSTM-raw 中存在几种文本描述模式，而 PubChem-extract 则进一步利用这些模式来提取更简洁的分子描述版本。详细说明如下：

- 最常见的模式是分子注释以 "XXX（名称）是/是/是/出现/出现/代表/属于/退出……"。我们通过人工提取获得了大部分分子名称，并将其替换为 "该分子……" 或 "这些分子……"。
- **多余的 "纯" 字。**有些分子注释以 "纯 xxx ……" 开头，我们去掉了 "纯" 字。
- **错别字。**例如，"Mercurycombines …" 应为 "Mercury combines …"。

数据集示例 我们在表 2 中提供了 PubChemSTM 原始数据集和 PubChemSTM 提取数据集的四个示例。

可重复性 由于 PubChem 数据库[1]经常在线更新，因此我们提供了本工作中使用的所有预处理数据集，以实现可重复性。此外，我们还提供了上述步骤的源代码，供今后使用。

比较 如前所述，我们采用了预先训练好的 SciBERT 模型[2]，并继续在 PubChemSTM 上进行训练。SciBERT 是专为科学发现训练的 BERT 模型。它从 Semantic Scholar [3] 中随机抽取了 114 万篇论文，其中约 18% 的论文来自计算机科学领域，82% 的论文来自广泛的生物医学领域。其语料库有 3.17B 标记，词汇量为 31K。此外，SciBERT 是针对论文全文而非摘要进行训练的。一个潜在的问题是从语义学者到 PubChemSTM 的词汇转移。虽然我们在这项工作中采用了 SciBERT 的预训练检查点（连同其词汇），但我们仍然希望仔细检查文本数据的词汇。

在表 3 中，我们列出了 PubChemSTM-raw 和 PubChemSTM-extract 在三种标记化方法下的词汇量：使用留白、spaCy [4] 和 SciBERT 标记化器。我们可以看到，与使用空白和 spaCy 的方法相比，使用 SciBERT 标记化器的 PubChemSTM-raw 和 PubChemSTM-extract 之间的差异非常小。因此，我们认为词汇也是一个重要因素，而 SciBERT 标记符号化器已经显示出相当稳定的标记效果。未来，我们需要更全面的标记化和词汇来推进这一研究方向，~~即~~实现用于药物发现的大语言模型。但这超出了本文的范围，需要整个社区的努力。

A.2 建筑细节

我们有两个分支，即化学结构分支 f_c 和文本描述分支 f_t 。

表 2. PubChemSTM 上的示例。这里只列出了化学结构的 SMILES 字符串，因为二维拓扑图可以通过 RDKit 软件包获得。

PubChemSTM-raw	PubChemSTM 提取的		
	SMILES: c1ccccc1		
苯是一种具有甜味的无色液体。它很快	蒸发到空气中，	在水中略有溶解。	很快蒸发到空气中，
中，稍溶于水			
	SMILES: Oc1ccccc1		
苯酚既是一种人造化学品，也是一种天然物质。	这种分子是，既是一种人造化学品，也是一种天然物质。它在纯净		
时是一种无色至白色的固体	物质纯净时为无色至白色固体。		
	SMILES: CC(=O)Oc1ccccc1C(=O)O		
乙酰水杨酸呈无味的白色结晶或结晶--	该分子在，	呈无味的白色结晶或结晶性水杨碱粉末，	略带苦味。粉末
，略带苦味			
	SMILES: CC1(C)SC2C(NC(=O)Cc3ccccc3)C(=O)N2C1C(=O)O		
苄青霉素是一种青霉素，其中位于戊环第 6 位的取代基为	该分子是一种青霉素，其戊环第 6 位的取代基为苯乙酰胺		
苯乙酰胺基。它具有抗菌药、表位和药物过敏原的作用。	基。它具有抗菌药、表位和药物过敏原的作用。		

表 3. 词汇比较。

数据来源	标记化方法	词汇量	与科学计算机科学研究中心的重叠
语义学者 (SciBERT 中使用)	SciBERT 标记符号生成器	31,090	-
PubChemSTM-raw	空白	315,704	7,635
	水疗	114,976	719
	SciBERT 标记符号生成器	18,320	18,320
PubChemSTM-extract	空白	100,877	7,562
	水疗	27,519	691
	SciBERT 标记符号生成器	17,442	17,442

化学结构分支 f_c 这项工作考虑了两种类型的化学结构：SMILES 字符串将分子视为序列，二维分子图分别将原子和化学键作为节点和边。然后，基于化学结构，我们应用深度学习编码器 f_c 获得作为分子表示的潜向量。具体来说，对于 SMILES 字符串，我们采用 MegaMolBART [5] 的编码器，该编码器已在 ZINC 数据库 [6] 的 5 亿个分子上进行了预训练。对于分子图，我们使用 GraphMVP 预训练 [8] 的预训练图同构网络 (GIN) [7]。GraphMVP 正在对来自 GEOM 数据集 [9] 的 25 万个构象进行二维拓扑和三维几何之间的多视角预训练。因此，虽然我们没有明确利用三维几何图形，但最先进的预训练 GIN 模型可以隐含地编码此类信息。

文本描述分支 f_t 文本描述分支提供了分子功能的高级描述。我们可以把这个分支看作是加强分子表征的领域知识。这些领域知识采用自然语言形式，我们使用 BERT 模型 [10] 作为文本编码器 f_t 。我们进一步调整了预训练的

SciBERT [2], 该模型是在化学和生物领域的文本数据上进行预训练的。

表 4.模型规格。每个模型中的 # 个参数。

分支		型号# 参数
化学 结构	GIN	1,885,206
	MegaMolBART	10,010,635
文字说明 SciBERT		109,918,464

A.3 培训前详情

预训练目标 对于 MoleculeSTM 预训练, 我们采用对比学习法。更具体地说, 我们从 EBM-NCE [8] 和 InfoNCE [11] 中选择一种。二者本质上做的是同一件事, 但 EBM-NCE 被认为比 InfoNCE 更有效。

对图形数据有效 [8, 12]。EBM-NCE 的目标是：

$$L = -\frac{1}{2} \mathbb{E}_{x_c, x_t} \log \sigma(E(x_c, x_t)) + \mathbb{E}_{x_c, x_t'} \log(1 - \sigma(E(x_c, x_t'))) - \frac{1}{2} \mathbb{E}_{x_c, x_t} \log \sigma(E(x_c, x_t)) + \mathbb{E}_{x_c, x_t'} \log(1 - \sigma(E(x_c, x_t'))) \quad (6)$$

其中， x_c 和 x_t 构成每个分子的结构-文本对， x_c' 和 x_t' 是随机从噪声分布中抽取的负样本，我们使用经验数据分布。 $E(-)$ 是具有灵活表述的能量函数，我们使用联合学习空间上的点积，即 $E(x_c, x_t) = \langle p_c \circ f_c(x_c), p_t \circ f_t(x_t) \rangle$ 。同样，我们将 InfoNCE 的目标设为

$$L = -\frac{1}{2} \mathbb{E}_{x_c, x_t} \log \frac{\exp(E(x_c, x_t))}{\exp(E(x_c, x_t)) + \sum_{x_t'} \exp(E(x_c, x_t'))} + \log \frac{\exp(E(x_c, x_t))}{\exp(E(x_c, x_t)) + \sum_{x_c'} \exp(E(x_c', x_t))} \quad (7)$$

超参数 我们分别以 SMILES 字符串和二维分子图为输入，列出了用于 MoleculeSTM 预训练的关键超参数。

表 5.MoleculeSTM 预训练的超参数规格。

输入	超参数	值 历
SMILES 字符串	时	{32}
	文本分支的学习率{1e-	4}
	支的学习率{1e-5, 3e-5}	化学结构分
二维分子图	目标函数	{ EBM-NCE, InfoNCE}
	纪元	{32}
	文本分支的学习率{1e-	4}
	支的学习率{1e-5, 3e-5}	化学结构分
	目标函数	{ EBM-NCE, InfoNCE}

运行时间 我们分别以 SMILES 字符串和二维分子图为输入，列出了 MoleculeSTM 的运行时间。

表 6.MoleculeSTM 预训练的运行时间。

输入运行时间	
SMILES 字符串	44 分钟/历时
2D 分子图	42 分钟/历时

B 下游任务的设计原则

本节将讨论设计下游任务的关键原则。

适用评估 视觉语言领域的基础模型与我们的 MoleculeSTM 最大的区别之一体现在评估方面。大多数视觉和语言任务都可以看作是艺术问题，*即*不存在适用于评估的标准和精确的解决方案。例如，我们可以检测出图像是 "一匹骑着宇航员的马" 还是 "一只正在制作拿铁艺术品的熊猫"[13]，但只能从视觉上而非计算上进行检测，因此无法进行大规模评估。药物发现则不然，因为这是一项科学任务，其结果（如编辑任务中输出分子的属性）可以在体外或硅学中精确评估。因此，物理实验通常成本高昂且持续时间长，所以在这项工作中，我们希望将重点放在那些在计算上可行的评估任务上。

模糊匹配 具体到分子编辑任务，文本提示应遵循 "模糊匹配" 标准，因为可能存在多个输出分子。这与 "精确匹配" 是矛盾的，因为 "精确匹配" 的输出分子是确定的。例如，对于官能团变化，我们可以输入 "将环中的第三个氮改为氧" 这样的提示。这个提示非常明确，有精确的解决方案，而且有基于规则的化学工具可以完美地处理这个问题。因此，基于文本的编辑无法在这一轨道上显示其优势。相反，在模糊匹配设置中，基于文本的编辑可以在潜空间中的语义方向上游荡，从而带来更多好处。这也反映了我们一直关注的语言模型的 *开放词汇* 属性。

C 下游：零次结构文本检索

C.1 数据集构建

DrugBank 数据库[14]有许多字段，这些字段对探索药物发现任务很有意义。在此，我们提取了每种小分子药物的三个字段，用于零点检索任务：描述字段、药效学字段和解剖治疗化学（ATC）字段，详情如下：

- **DrugBank-Description.**描述字段提供了药物化学特性、历史和监管状态的高级审查。
- **DrugBank-药效学。**这说明了药物如何改变或影响所使用的生物体。这一领域可能包括药物在体内产生的预期效果和不预期效果（也称为副作用）。
- **DrugBank-ATC.**解剖治疗化学物（ATC）是一种分类系统，它根据分子作用的器官或系统及其治疗、药理和化学特性将分子分为不同的组别。

我们将数据集构建的关键步骤列举如下：

1. 我们从[网站](#)上下载完整的 DrugBank 数据库（XML 格式）和小型化学结构文件（SDF 格式）。
2. 我们解析 XML 文件，提取三个字段的数据：描述、药效学和 ATC。
3. 我们将提取的文件与 SDF 文件中的化学结构进行映射。对于 DrugBank-Description 和 DrugBank-Pharmacodynamics 数据集，我们剔除了在 PubChemSTM 中出现过的分子，并用规范 SMILES 进行了过滤。同时，对于 DrugBank-ATC，我们会同时排除符合以下两个标准的分子：

- **化学结构过滤** 如果具有相同规范 SMILES 的分子已在 PubChemSTM 中出现；
- **文本数据过滤** 我们首先需要定义两个文本数据之间的相似度，如公式（8）所示，其中 textDrugBank 和 textPubChemSTM 分别是 DrugBank 和 PubChemSTM 中同一分子的文本数据， $\text{len}()$ 是文本数据的长度， $\text{Levenshtein}()$ 是两个文本数据之间的列文森距离。因此，第二个条件是：如果 DrugBank 文本和 PubChemSTM 文本之间的相似度高于某个阈值（例如 0.6）。

另一个细节是，在 DrugBank-ATC 中，每个小分子都有多个 ATC 字段（ textDrugBank ）。在 PubChemSTM 中，每个分子也有多个文本描述（ textPubChemSTM ）。因此，在文本数据过滤步骤中，对于 DrugBank 和 PubChemSTM 之间的每个共享分子，我们会计算所有 $\text{textDrugBank-textPubChemSTM}$ 文本对的相似度，如果存在相似度超过阈值 0.6 的文本对，则排除该分子。

4. 表 7 列出了一些基本的数据集统计数据。请注意，ATC 有多个级别，我们在本工作中使用第 5 级进行检索。

$$\text{sim}_{\text{textDrugBank, text PubChemSTM}} = 1 - \frac{\text{Levenshtein}(\text{textDrugBank}, \text{textPubChemSTM})}{\text{len}(\text{textDrugBank})} \quad (8)$$

表 7. 药物库中三个字段的统计数据。过滤步骤如上所示。

现场	# 结构-文本对 子不在 PubChemSTM 中	在 PubChemSTM		总计
		中共享的 # 结构-文本对分子 但文本相似度低于 0.6		
药物库-描述	1,154	-	-	1,154
药物库-药效学	1,005	-	-	1,005
药库-ATC	1,507		1,500	3,007

C.2 实验

在实验方面，我们在正文中介绍了三种基线。作为概念验证，我们还进行了另一项名为随机的基线实验。对于 Random，两个编码器 (f_c 和 f_i) 都是随机初始化的。表 8 至表 10 显示了三个数据集的零次检索结果。

表 8. 药物库-描述 T 选一检索的准确率 (%)。

		给定化学结构			给定文本		
	T	4	10	20	4	10	20
微笑	随机	24.59 ± 1.14	10.12 ± 1.38	4.97 ± 0.42	24.54 ± 0.97	9.97 ± 0.81	5.09 ± 0.37
	冷冻	25.07 ± 1.24	10.22 ± 1.19	5.12 ± 0.65	24.69 ± 1.87	10.20 ± 1.38	5.37 ± 1.15
	相似性	36.35 ± 0.59	23.22 ± 0.58	16.40 ± 0.59	22.74 ± 0.24	10.31 ± 0.24	5.34 ± 0.24
	KV-PLM	73.80 ± 0.00	53.96 ± 0.29	40.07 ± 0.38	72.86 ± 0.00	52.55 ± 0.29	40.33 ± 0.00
	分子 STM	97.50 ± 0.46	94.18 ± 0.46	91.12 ± 0.46	98.21 ± 0.00	94.54 ± 0.37	91.97 ± 0.46
图表	随机	25.78 ± 1.43	10.71 ± 0.97	4.83 ± 1.00	24.98 ± 0.32	10.20 ± 0.40	4.80 ± 0.21
	冷冻	24.01 ± 1.34	9.39 ± 0.92	4.85 ± 0.52	24.00 ± 1.66	9.91 ± 0.71	5.07 ± 0.75
	相似性	30.03 ± 0.38	13.63 ± 0.27	7.07 ± 0.10	24.81 ± 0.27	10.22 ± 0.24	4.74 ± 0.24
	分子 STM	99.15 ± 0.00	97.19 ± 0.00	95.66 ± 0.00	99.05 ± 0.37	97.50 ± 0.46	95.71 ± 0.46

表 9. 药物数据库-药效学 T 选一检索的准确率 (%)。

		给定化学结构			给定文本		
	T	4	10	20	4	10	20
微笑	随机	24.49 ± 0.68	9.73 ± 0.34	5.14 ± 0.57	25.61 ± 0.62	10.10 ± 0.91	5.07 ± 0.69
	冷冻	25.47 ± 1.12	10.55 ± 0.75	5.48 ± 0.70	25.34 ± 0.41	9.86 ± 0.44	4.84 ± 0.26
	相似性	27.85 ± 0.03	10.75 ± 0.02	5.67 ± 0.01	24.58 ± 0.03	11.25 ± 0.03	5.29 ± 0.02
	KV-PLM	68.38 ± 0.03	47.59 ± 0.03	36.54 ± 0.03	67.68 ± 0.03	48.00 ± 0.02	34.66 ± 0.02
	分子 STM	88.07 ± 0.01	81.70 ± 0.02	75.94 ± 0.02	88.46 ± 0.01	81.01 ± 0.02	74.64 ± 0.03
图表	随机	26.00 ± 0.37	9.65 ± 0.88	4.95 ± 0.36	25.11 ± 0.63	9.99 ± 0.62	4.82 ± 0.54
	冷冻	25.49 ± 1.82	10.19 ± 1.47	4.74 ± 0.56	25.55 ± 0.45	10.15 ± 0.77	4.88 ± 0.55
	相似性	25.33 ± 0.27	9.89 ± 0.52	4.61 ± 0.08	25.28 ± 0.03	10.64 ± 0.02	5.47 ± 0.02
	分子 STM	92.14 ± 0.02	86.27 ± 0.02	81.08 ± 0.05	91.44 ± 0.02	86.76 ± 0.03	81.68 ± 0.03

表 10. 分子-ATC T 选一检索的准确率 (%)。

		给定化学结构			给定文本		
	T	4	10	20	4	10	20
微笑	随机	25.03 ± 0.33	9.83 ± 0.19	4.80 ± 0.22	25.44 ± 1.21	10.03 ± 0.94	5.11 ± 0.79
	冷冻	25.05 ± 0.94	10.17 ± 0.63	4.99 ± 0.54	25.35 ± 0.78	10.32 ± 0.44	5.22 ± 0.34
	相似性	30.03 ± 0.00	13.35 ± 0.02	7.53 ± 0.02	26.74 ± 0.03	11.01 ± 0.00	5.62 ± 0.00
	KV-PLM	60.94 ± 0.00	42.35 ± 0.00	30.32 ± 0.00	60.67 ± 0.00	40.19 ± 0.00	29.02 ± 0.00
	分子 STM	70.84 ± 0.07	56.75 ± 0.05	46.12 ± 0.07	73.07 ± 0.03	58.19 ± 0.03	48.97 ± 0.06
图表	随机	24.48 ± 0.66	9.97 ± 0.25	4.81 ± 0.34	25.48 ± 0.59	10.40 ± 0.37	5.38 ± 0.30
	冷冻	24.19 ± 0.77	10.24 ± 0.71	4.87 ± 0.47	24.95 ± 1.52	10.07 ± 0.80	5.06 ± 0.36
	相似性	29.46 ± 0.00	12.34 ± 0.00	6.52 ± 0.00	25.78 ± 1.53	10.23 ± 0.70	5.06 ± 0.67
	分子 STM	69.33 ± 0.03	54.83 ± 0.04	44.13 ± 0.05	71.81 ± 0.05	58.34 ± 0.07	47.58 ± 0.05

C.3 消融研究：固定预训练编码器

在主体部分，我们通过采用预训练的单模态检查点进行预训练，即针对 f_c 采用 GraphMVP 和 MegaMolBART，针对 f_t 采用 SciBERT。然后，针对 MoleculeSTM 预训练，我们使用对比学习并更新所有模型参数。在此，我们进行了一项消融研究，只对两个分支的联合空间 (p_c, p_t) 的投影层进行优化，同时保持两个编码器 (f_c, f_t) 固定不变。三个数据集的结果如表 11 至表 13 所示。

表 11. DrugBank-Description T -choose-one 检索的准确率 (%)。

		给定化学结构			给定文本		
	T	4	10	20	4	10	20
微笑	随机	24.59 ± 1.14	10.12 ± 1.38	4.97 ± 0.42	24.54 ± 0.97	9.97 ± 0.81	5.09 ± 0.37
	冷冻	25.07 ± 1.24	10.22 ± 1.19	5.12 ± 0.65	24.69 ± 1.87	10.20 ± 1.38	5.37 ± 1.15
	相似性	36.35 ± 0.59	23.22 ± 0.58	16.40 ± 0.59	22.74 ± 0.24	10.31 ± 0.24	5.34 ± 0.24
	分子 STM	47.64 ± 0.40	29.21 ± 0.47	19.69 ± 0.47	52.60 ± 0.46	32.24 ± 0.37	21.45 ± 0.37
图表	随机	25.78 ± 1.43	10.71 ± 0.97	4.83 ± 1.00	24.98 ± 0.32	10.20 ± 0.40	4.80 ± 0.21
	冷冻	24.01 ± 1.34	9.39 ± 0.92	4.85 ± 0.52	24.00 ± 1.66	9.91 ± 0.71	5.07 ± 0.75
	相似性	30.03 ± 0.38	13.63 ± 0.27	7.07 ± 0.10	24.81 ± 0.27	10.22 ± 0.24	4.74 ± 0.24
	分子 STM	51.28 ± 0.00	31.99 ± 0.41	20.71 ± 0.47	55.27 ± 0.00	33.08 ± 0.00	21.77 ± 0.00

表 12. 药物数据库-药效学 T 选一检索的准确率 (%)。

		给定化学结构			给定文本		
	T	4	10	20	4	10	20
微笑	随机	24.49 ± 0.68	9.73 ± 0.34	5.14 ± 0.57	25.61 ± 0.62	10.10 ± 0.91	5.07 ± 0.69
	冷冻	25.47 ± 1.12	10.55 ± 0.75	5.48 ± 0.70	25.34 ± 0.41	9.86 ± 0.44	4.84 ± 0.26
	相似性	27.85 ± 0.03	10.75 ± 0.02	5.67 ± 0.01	24.58 ± 0.03	11.25 ± 0.03	5.29 ± 0.02
	分子 STM	46.43 ± 0.00	27.42 ± 0.47	18.24 ± 0.47	52.53 ± 0.41	30.53 ± 0.00	19.98 ± 0.00
图表	随机	26.00 ± 0.37	9.65 ± 0.88	4.95 ± 0.36	25.11 ± 0.63	9.99 ± 0.62	4.82 ± 0.54
	冷冻	25.49 ± 1.82	10.19 ± 1.47	4.74 ± 0.56	25.55 ± 0.45	10.15 ± 0.77	4.88 ± 0.55
	相似性	25.33 ± 0.27	9.89 ± 0.52	4.61 ± 0.08	25.28 ± 0.03	10.64 ± 0.02	5.47 ± 0.02
	分子 STM	46.29 ± 0.03	27.18 ± 0.02	17.73 ± 0.02	50.95 ± 0.04	31.65 ± 0.03	23.00 ± 0.03

表 13. DrugBank-ATC T -choose-one 检索的准确率 (%)。

		给定化学结构			给定文本		
	T	4	10	20	4	10	20
微笑	随机	25.03 ± 0.33	9.83 ± 0.19	4.80 ± 0.22	25.44 ± 1.21	10.03 ± 0.94	5.11 ± 0.79
	冷冻	25.05 ± 0.94	10.17 ± 0.63	4.99 ± 0.54	25.35 ± 0.78	10.32 ± 0.44	5.22 ± 0.34
	相似性	30.03 ± 0.00	13.35 ± 0.02	7.53 ± 0.02	26.74 ± 0.03	11.01 ± 0.00	5.62 ± 0.00
	分子 STM	43.41 ± 0.12	25.66 ± 0.06	15.69 ± 0.06	48.75 ± 0.11	29.44 ± 0.06	19.75 ± 0.03
图表	随机	24.48 ± 0.66	9.97 ± 0.25	4.81 ± 0.34	25.48 ± 0.59	10.40 ± 0.37	5.38 ± 0.30
	冷冻	24.19 ± 0.77	10.24 ± 0.71	4.87 ± 0.47	24.95 ± 1.52	10.07 ± 0.80	5.06 ± 0.36
	相似性	29.46 ± 0.00	12.34 ± 0.00	6.52 ± 0.00	25.78 ± 1.53	10.23 ± 0.70	5.06 ± 0.67
	分子 STM	42.53 ± 0.07	24.34 ± 0.00	14.78 ± 0.03	48.91 ± 0.03	28.77 ± 0.07	19.28 ± 0.07

D 下游：基于文本的零镜头分子编辑

分子编辑或可控分子生成是指根据给定和预训练的分子生成模型改变分子结构。在这项工作中，借助 MoleculeSTM 中的大型语言模型，我们能够实现基于零镜头文本的分子编辑。首先，我们想列举视觉领域和分子领域编辑任务的两个关键挑战，具体如下：

- **骨干生成模型。**对于视觉领域来说，基于 StyleGAN [15]（一种很好地纠缠的骨干模型）的图像可控生成是相当可行的。然而，这对于深度分子生成模型来说并非易事。最近的一项研究 GraphCG [16] 探讨了基于图的可控分子生成方法的不纠缠特性，结论是，尽管骨干生成模型并不是完美不纠缠的，但仍然存在对分子图或点云等高结构化数据进行可控生成的方法。同时，由于 MoleculeSTM 的编辑解决方案是与模型无关的，可以很容易地推广到未来的模型中，因此开发一种新颖的解缠分子生成模型不在本文的研究范围之内。
- **评价。**图像可控生成是一个艺术问题，~~即~~它是主观的，可以有多种（甚至无限多）答案。相反，可控分子生成是一个科学问题，~~即~~它是客观的，只有少数几个答案。附录 B 对此进行了讨论。

D.1 实验设置

两个关键的超参数是学习率 $\{1e-2, 1e-3\}$ 和 $\lambda \in \{1e1, 1e0, 1e-1, 1e-2, 1e-3\}$ 。为了公平比较，对于基线，我们采用 $w = w_{in} + \alpha \cdot D$ 的形式，其中 D 是通过随机、PCA 和方差得到的， $\lambda \in \{1.0, 1.5, 2.0, 2.5, 3.0\}$ 。对于 GS，我们对每个输入分子重复随机取样五次。

接下来，我们将对四种类型的编辑任务以及三个案例研究进行基于零镜头文本的分子编辑，具体讨论如下：

- 附录 D.2 中的单目标分子编辑（八项任务）。
- 附录 D.3 中的多目标分子编辑（六项任务）。
- 附录 D.4 中基于结合亲和力的分子编辑（六项任务）。
- 附录 D.5 中的药物相关性编辑（四项任务）。
- 附录 D.6（三个案例研究）中的专利药物分子邻近搜索。

由于篇幅限制，我们只在正文中展示了四项多目标编辑任务和四项基于绑定亲和力的编辑任务。在此，我们将展示更全面的成果。

值得一提的是，在进行单目标和多目标编辑时，我们从 ZINC 中随机选择了 200 个分子作为输入分子。这 200 个输入分子中没有一个是出现在 PubChemSTM 中。此外，随机选择过程确保了这 200 个分子的性质分布与整个数据集保持一致。下图（图 6 和图 7）是分子性质的三个示例：LogP（测量水溶性）、tPSA（测量渗透性）和分子量。

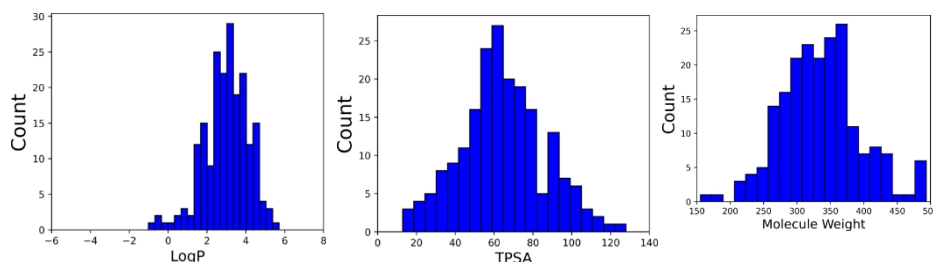


图 6.200 个随机抽样编辑分子的三种属性分布。

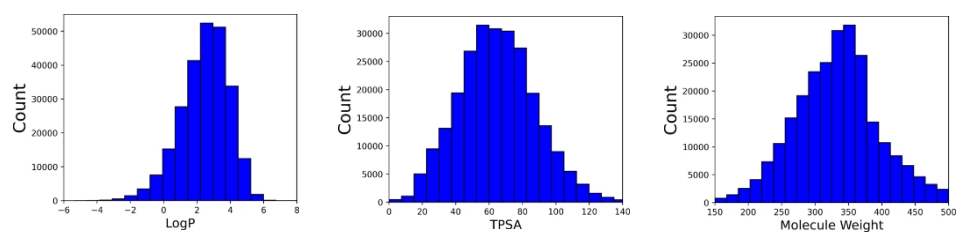


图 7.来自 ZINC250K 的 250K 分子的三种属性分布。

D.2 单目标分子编辑

我们首先考虑分子编辑的八个单目标特性。如 "方法 "部分所示，满意度函数和阈值 Δ 的定义是根据每项任务具体确定的，如

- 我们使用 LogP 来评估溶解度和不溶解度。我们将 0 和 0.5 作为不同的阈值。
- 我们使用 QED 来评估药物相似性。我们将 0 和 0.1 作为不同的阈值。
- 我们使用 tPSA 来评估高渗透性和低渗透性。我们将 0 和 10 作为不同的阈值。
- 对于氢键受体（HBA）和氢键供体（HBD），我们可以直接计算它们在分子中的数量，并使用 0 和 1 作为不同的阈值。

对于 Δ 而言，它是一个阈值，只有高于这个阈值的差异才能被视为命中。因此， Δ 越大，意味着编辑标准越严格。下面我们将展示八个单一目标属性分子编辑结果的定量和定性结果。

表 14.八个单一目标分子编辑的结果。输入是从 ZINC 中随机抽样的 200 个分子，评价指标是属性变化的命中率。潜在优

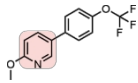
化是使用 MoleculeSTM 进行基于文本的分子编辑，分别使用 SMILES 字符串和分子图。

	Δ	基准线				潜优化	
		随机	PCA	高差异	GS-Mutate	微笑	图表
这种分子可溶于水。	0	35.33 ± 1.31	33.80 ± 3.63	33.52 ± 3.75	52.00 ± 0.41	61.87 ± 2.67	67.86 ± 3.46
	0.5	11.04 ± 2.40	10.66 ± 3.24	10.86 ± 2.56	14.67 ± 0.62	49.02 ± 1.84	54.44 ± 3.99
这种分子不溶于水。	0	43.36 ± 3.06	39.36 ± 2.55	42.89 ± 2.36	47.50 ± 0.41	52.71 ± 1.67	64.79 ± 2.76
	0.5	19.75 ± 1.56	15.12 ± 2.93	18.22 ± 0.33	12.50 ± 0.82	30.47 ± 3.26	47.09 ± 3.42
这种分子就像毒品。	0	38.06 ± 2.57	33.99 ± 3.72	36.20 ± 4.34	28.00 ± 0.71	36.52 ± 2.46	39.97 ± 4.32
	0.1	5.27 ± 0.24	3.97 ± 0.10	4.44 ± 0.58	6.33 ± 2.09	8.81 ± 0.82	14.06 ± 3.18
这种分子与药物不同。	0	36.96 ± 2.25	35.17 ± 2.61	39.99 ± 0.57	71.33 ± 0.85	58.59 ± 1.01	77.62 ± 2.80
	0.1	6.16 ± 1.87	5.26 ± 0.95	7.56 ± 0.29	27.67 ± 3.79	37.56 ± 1.76	54.22 ± 3.12
这种分子具有高渗透性。	0	25.23 ± 2.13	21.36 ± 0.79	21.98 ± 3.77	22.00 ± 0.82	57.74 ± 0.60	59.84 ± 0.78
	10	17.41 ± 1.43	14.52 ± 0.80	14.66 ± 2.13	6.17 ± 0.62	47.51 ± 1.88	50.42 ± 2.73
这种分子的渗透性很低。	0	16.79 ± 2.54	15.48 ± 2.40	17.10 ± 1.14	28.83 ± 1.25	34.13 ± 0.59	31.76 ± 0.97
	10	11.02 ± 0.71	10.62 ± 1.86	12.01 ± 1.01	15.17 ± 1.03	26.48 ± 0.97	19.76 ± 1.31
这种分子有更多的氢键受体。	0	12.64 ± 1.64	10.85 ± 2.29	11.78 ± 0.15	21.17 ± 3.09	54.01 ± 5.26	37.35 ± 0.79
	1	0.69 ± 0.01	0.90 ± 0.84	0.67 ± 0.01	1.83 ± 0.47	27.33 ± 2.62	16.13 ± 2.87
这种分子有更多的氢键供体。	0	2.97 ± 0.61	3.97 ± 0.55	6.23 ± 0.66	19.50 ± 2.86	28.55 ± 0.76	60.97 ± 5.09
	1	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	1.33 ± 0.24	7.69 ± 0.56	32.35 ± 2.57

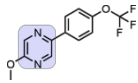
表 15.根据分子的辛醇-水分配系数（LogP）的对数测量溶解度的文本编辑可视化。一般来说，LogP 较小的分子更易溶于水。

为了生成可溶于水的分子，我们可以在输入分子中添加极性成分（如氧原子和硝基）、去除疏水分子（如苯和环己烷）或用极性官能团取代疏水基团。为了生成不溶于水的分子，我们可以对输入分子进行相反的修饰。粉色和蓝色区域分别表示输入和输出分子中的修饰结构。

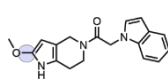
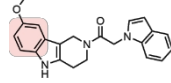
文字提示：这种分子可溶于水。



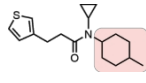
输入



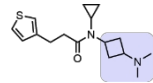
分子输出 分子输入 分子输出 分子



输入 分子



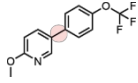
输出 分子



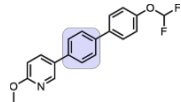
输入 分子式

LogP: 3.66LogP : 3.05LogP : 3.72LogP : 2.56LogP: 4.25LogP : 2.76

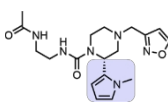
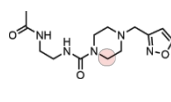
文字提示：这种分子不溶于水。



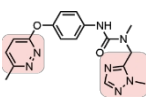
输入



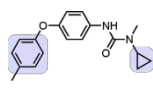
分子输出 分子输入 分子输出 分子



输入 分子



输出 分子



输入 分子式

LogP: 3.66LogP : 5.03LogP : -0.36LogP : 0.72LogP: 2.37LogP : 4.41

表 16.通过分子的拓扑极性表面积（tPSA）测量的基于文本编辑的渗透性可视化。一般来说，tPSA 越小的分子渗透性越强。为了生成高渗透性分子，我们可以从输入分子中去除高极性的官能团或杂环，如酰胺、磺酰胺、脒、硝基和含氮烷。为了生成低渗透性分子，我们可以对输入分子进行相反的修饰。粉色和蓝色区域分别表示输入和输出分子中的修改结构。

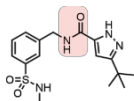
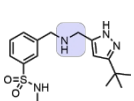
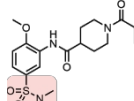
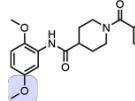
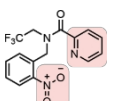
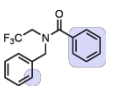
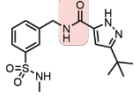
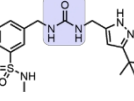
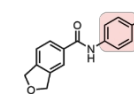
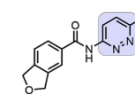
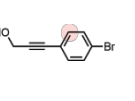
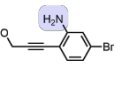
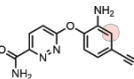
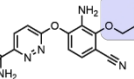
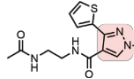
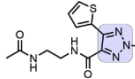
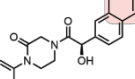
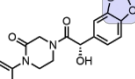
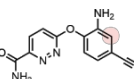
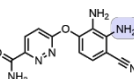
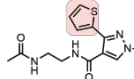
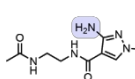
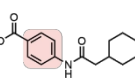
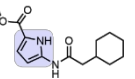
文字提示：这种分子具有高渗透性。					
					
输入	分子输出 分子输入 分子输出 分子	输入 分子	输出 分子	输入 分子式	
tPSA: 104	tPSA: 87	tPSA: 96	tPSA: 68	tPSA: 76	tPSA: 20
文字提示：这种分子的渗透性很低。					
					
输入	分子输出 分子输入 分子输出 分子	输入 分子	输出 分子	输入 分子式	
tPSA: 104	tPSA: 116	tPSA: 42	tPSA: 67	tPSA: 20	tPSA: 46

表 17.基于文本编辑的氢键受体（HBA）和氢键供体（HBD）的可视化。要生成具有更多氢键受体的分子，我们可以在输入分子中添加氧、氮和硫等杂原子，或用含杂原子的结构基团替换现有基团。为了生成更多的 HBD 分子，我们可以添加附有氢的杂原子，如胺等官能团和吡咯等杂环。粉色和蓝色区域分别表示输入和输出分子中的修改结构。

文字提示：该分子有更多的氢键受体。					
					
输入	分子输出 分子输入 分子输出 分子	输入 分子	输出 分子	输入 分子式	
HBA: 6	HBA: 7	HBA: 5	HBA: 6	HBA: 3	HBA: 5
文字提示：该分子有更多的氢键供体。					
					
输入	分子输出 分子输入 分子输出 分子	输入 分子	输出 分子	输入 分子式	
HBD: 2	HBD: 2	HBD: 3	HBD: 1	HBD: 2	

D.3 多目标分子编辑

然后，我们考虑了分子编辑的六个多目标特性。如 "方法 "部分所示，满意度函数和阈值 Δ 的定义具体基于每项任务。首先，对于每个单目标，我们遵循附录 D.2 中的评价指标，包括溶解度、渗透性以及 HBA 和 HBD 的数量。然后，对于多目标评价，我们考虑两种情况：

- 阈值宽松的**简单**情况，如溶解度和渗透性同时为 0 和 0 的阈值。
- 具有严格阈值的**挑战性**案例，如溶解度和 HBA/HBD 同时阈值为 0.5 和 1，溶解度和渗透性同时阈值为 0.5 和 10。

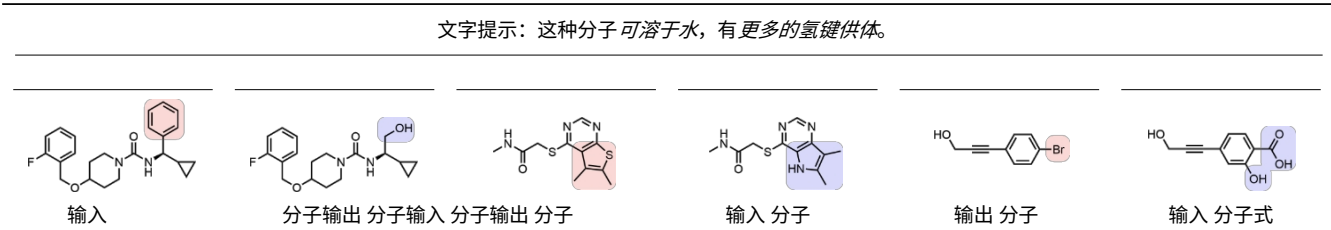
那么，一个成功的命中需要同时满足这两个条件。下面我们将展示六个多目标属性分子编辑结果的定量和定性结果。

表 18.六个多目标分子编辑的结果。输入是从 ZINC 中随机抽样的 200 个分子，评价指标是属性变化的命中率。潜在优化是使

用 MoleculeSTM 进行基于文本的分子编辑，分别使用 SMILES 字符串和分子图。

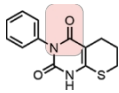
	Δ	基质优化					
		随机	PCA	高差异	GS-Mutate	微笑	图表
该分子可溶于水 并有更多的氢键受体。	0 - 0	9.88 \pm 1.03	8.64 \pm 2.06	9.09 \pm 1.25	14.00 \pm 2.48	27.87 \pm 3.86	27.43 \pm 3.41
	0.5 - 1	0.23 \pm 0.33	0.45 \pm 0.64	0.22 \pm 0.31	0.67 \pm 0.62	8.80 \pm 0.04	11.10 \pm 1.80
这种分子不溶于水 并有更多的氢键受体。	0 - 0	2.99 \pm 0.38	2.00 \pm 0.58	2.45 \pm 0.67	7.17 \pm 0.85	8.55 \pm 2.75	8.21 \pm 0.81
	0.5 - 1	0.45 \pm 0.32	0.00 \pm 0.00	0.22 \pm 0.31	0.17 \pm 0.24	2.93 \pm 0.30	0.00 \pm 0.00
该分子可溶于水 并有更多的氢键供体。	0 - 0	2.28 \pm 1.15	2.23 \pm 1.16	4.44 \pm 0.58	13.83 \pm 2.95	33.51 \pm 4.08	49.23 \pm 1.71
	0.5 - 1	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	9.98 \pm 1.03	23.94 \pm 1.09
这种分子不溶于水 并有更多的氢键供体。	0 - 0	0.69 \pm 0.58	1.96 \pm 0.87	1.79 \pm 0.66	5.67 \pm 0.62	17.03 \pm 2.75	14.42 \pm 3.43
	0.5 - 1	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	0.00 \pm 0.00	2.59 \pm 1.14	3.84 \pm 0.71
该分子可溶于水 并具有高渗透性。	0 - 0	5.06 \pm 1.21	3.53 \pm 0.38	4.88 \pm 2.21	8.17 \pm 1.03	35.69 \pm 3.19	39.74 \pm 2.26
	0.5 - 10	1.16 \pm 0.68	0.67 \pm 0.55	0.66 \pm 0.54	0.00 \pm 0.00	19.15 \pm 0.73	22.66 \pm 1.90
该分子可溶于水 渗透性低。	0 - 0	12.17 \pm 1.05	10.43 \pm 2.88	13.08 \pm 2.28	19.83 \pm 2.46	44.35 \pm 0.68	30.87 \pm 0.62
	0.5 - 10	6.20 \pm 0.64	6.23 \pm 2.31	6.67 \pm 0.53	4.83 \pm 0.85	28.67 \pm 2.22	20.06 \pm 1.26

表 19.基于文本编辑的多目标（组成）属性可视化：溶解度和氢键供体（HBD），以分子的 LogP 和 HBD 数量衡量。具有更多 HBD 的分子也可能溶于水，例如在输入分子中用极性基团或含有氢键杂原子的环（醇、氮杂吡啶、羧酸等）取代疏水基团（苯、噻吩、溴等）。不过，我们也可以在输入分子中添加 HBD，同时降低其溶解度，例如用亲水性较弱的 HBD（吡啶、硫醇等）取代输入分子中的高极性结构基团（酰胺、内酯等）。粉色和蓝色区域分别表示输入和输出分子中的修饰结构。



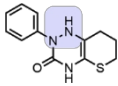
LogP: 4.67, HBD: 1LogP : 2.29, HBD: 2LogP: 2.15, HBD: 1LogP : 1.41, HBD : 2LogP: 1.79, HBD: 1LogP : 0.43, HBD: 3

文字提示：这种分子不溶于水，氢键供体较多。

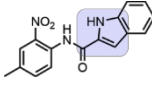
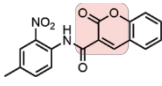


输入

LogP: 1.56, HBD: 1LogP : 2.42, HBD: 2LogP: 3.26, HBD: 1LogP

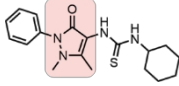


分子输出 分子输入 分子输出 分子



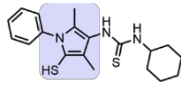
输入 分子

: 3.64, HBD: 2LogP



输出 分子

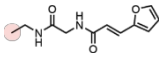
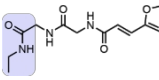
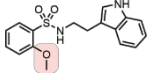
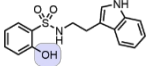
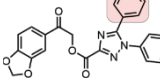
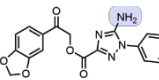
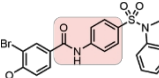
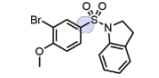
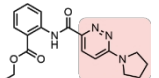
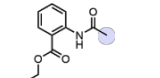
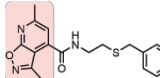
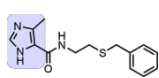
: 3.10, HBD: 2LogP



输入 分子式

: 5.00, HBD: 3

表 20.基于文本编辑的多目标（组成）属性可视化：溶解度和渗透性，以分子的 LogP 和 tPSA 度量。渗透性低的分子也可能溶于水，例如在输入分子中添加极性官能团（如酰胺、胺）和去除碳氢化合物（如甲基、苯基）。不过，我们也可以同时提高分子的溶解度和渗透性，如同时去除碳氢化合物和极性分子，或缩小输入分子中杂环的尺寸（如将[1,2]恶唑并[5,4-b]吡啶改为咪唑）。粉色和蓝色区域分别表示输入和输出分子中的修改结构。

文字提示：这种分子可溶于水，渗透性低。					
					
输入分子	输出分子	输入分子	输出分子	输入分子	输出分子
LogP: 0.55, tPSA: 71	LogP: -0.34, tPSA: 100	LogP: 2.70, tPSA: 71	LogP: 2.39, tPSA: 82	LogP: 3.70, tPSA: 93	LogP: 1.62, tPSA: 119
文字提示：这种分子可溶于水，具有高渗透性。					
					
输入分子	输出分子	输入分子	输出分子	输入分子	输出分子
LogP: 4.46, tPSA: 76	LogP: 3.21, tPSA: 47	LogP: 2.51, tPSA: 84	LogP: 1.82, tPSA: 55	LogP: 3.50, tPSA: 68	LogP: 2.38, tPSA: 58

D.4 基于结合亲和力的分子编辑

我们进一步对结合亲和力测定应用了基于文本的编辑。具体来说，我们从 ChEMBL [17] 中提取了六个结合亲和力任务。如表 21 所列，每项检测都有文字说明。

表 21.ChEMBL 检测说明。	
ChEMBL	IDAssay 说明
	1613777 这种分子在 检测酶蛋白的抑制剂和底物时呈阳性。它利用分子氧将一个氧原子插入底物，并将第二个氧原子还原成水分子。
1613797 该	分子在 炭疽致死 试验中检测呈阳性，炭疽致死试验是一种蛋白酶，可裂解大多数双特异性丝裂原活化蛋白激酶激酶的 N-末端。
2114713	该分子在 检测 ClpP 激活剂时呈阳性，ClpP 激活剂在需要 ATP 水解的过程中裂解各种蛋白质中的肽，在缺乏 ATP 结合亚基的情况下，其肽酶活性有限。
	1613838 在 检测参与内体和反式高尔基体网络之间蛋白质转运的活化剂时，该分子呈阳性。
	1614236 这种分子是一种蛋白质的抑制剂，它通过抑制泛素化来触发抗病毒转导信号，并抑制细胞前 mRNA 的转录后处理，从而阻止细胞抗病毒状态的建立。
1613903 这种	分子在 高通量筛选试验中 检测呈阳性，该试验旨在确定 SARS 冠状病毒 3C 样蛋白酶的抑制剂，该蛋白酶可在 11 个位点上裂解复制酶多聚蛋白的 C-末端。

我们按照 "方法 "部分进行评估。回想一下，每次结合亲和力检测都可以对应带有正标签和负标签的分子。因此，我们可以在这些数据点上训练分类器，这里的满足标准是输出分子的置信度是否高于输入分子，其中置信度是使用分类器对每个任务进行预测的。流程见图 8。



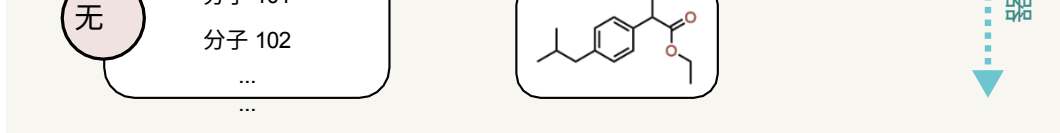


图 8.基于结合亲和力的分子编辑流程。输入的分⼦是从 ZINC 中随机抽样的，文本提示是检测说明。为了进行评估，每个检测项目的小分子都用于训练二元分类器，并考虑了两种类型的模型（随机森林和逻辑回归）。

命中率结果如表 22 所示。请注意，为了更好地证明结果的有效性，我们对每种检测方法训练了两个分类器：随机森林（RF）和逻辑回归（LR），并将指纹作为特征化特征。

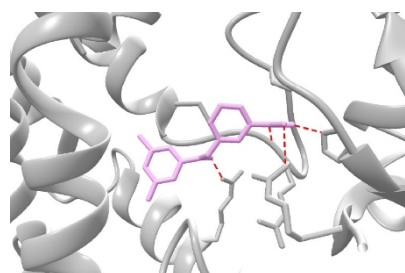
表 22.六项 ChEMBL 检测编辑的结果。每个 ChEMBL 检测都是二元任务，我们训练分类器以获得每个分子（输入和输出分子）的置信度得分。输入是从 ZINC 中随机抽样的 200 个分子，评估是置信度变化的命中率。潜在优化是使用

MoleculeSTM 进行基于文本的分子编辑，分别使用 SMILES 字符串和分子图。									
ChEMBL ID								基质优化	
		随机	PCA	高差异	GS-Mutate	微笑	图表		
1613777	RF	44.99 ± 2.08	44.49 ± 1.22	44.45 ± 1.01	39.17 ± 3.66	48.70 ± 2.06	44.53 ± 1.60		
	LR	47.34 ± 5.53	49.13 ± 0.86	49.69 ± 6.75	51.50 ± 2.86	54.09 ± 1.94	50.55 ± 3.14		
1613797	RF	44.76 ± 2.18	46.25 ± 0.97	46.92 ± 3.34	46.67 ± 1.55	55.03 ± 2.23	49.03 ± 0.03		
	LR	48.40 ± 3.71	49.92 ± 4.31	48.67 ± 1.64	49.17 ± 3.01	57.98 ± 3.34	54.95 ± 3.74		
2114713	RF	39.87 ± 2.32	42.91 ± 2.64	42.19 ± 3.68	41.33 ± 1.25	49.20 ± 2.11	60.93 ± 2.53		
	LR	51.39 ± 1.15	52.62 ± 1.64	52.24 ± 1.07	50.50 ± 1.47	56.93 ± 3.67	58.77 ± 2.41		
1613838	RF	44.49 ± 1.48	44.71 ± 1.80	45.30 ± 2.47	36.00 ± 2.68	43.94 ± 3.75	49.13 ± 2.52		
	LR	50.22 ± 4.23	49.73 ± 2.33	44.69 ± 2.41	41.33 ± 3.17	47.50 ± 2.28	56.13 ± 1.50		
1614236	RF	41.33 ± 3.59	42.28 ± 1.91	42.85 ± 2.88	45.33 ± 1.65	57.90 ± 2.39	35.71 ± 4.19		
	LR	46.57 ± 0.51	49.34 ± 1.80	50.62 ± 3.86	56.00 ± 1.08	65.78 ± 5.67	46.36 ± 2.53		
1613903	RF	44.28 ± 0.77	43.83 ± 2.65	42.00 ± 3.19	46.17 ± 0.85	56.82 ± 3.96	58.70 ± 1.43		
	LR	53.94 ± 3.30	48.63 ± 4.49	56.19 ± 2.51	56.33 ± 0.94	58.31 ± 2.98	64.64 ± 5.23		

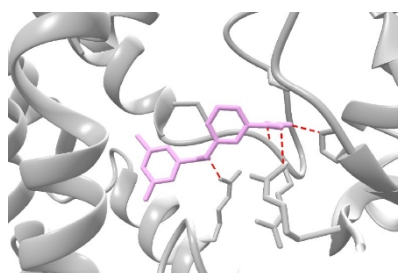
然后在图 9 中添加可视化对接。我们选择了 ChEMBL 1613777 和可用的 PDB 结构。具体来说，我们首先使用 MoleculeSTM 提取出置信度（RF 和 LR）高于基线生成的输出分子。然后运行分子对接软件得出结果。对接设置详情如下。

- 我们使用 RDKit [19] 中提供的默克分子力场（MMFF）[18] 为每个分子嵌入（生成）三维构象。MMFF 的介电常数设定为 80，优化的最大迭代次数为 1000，每个分子中最多有 5 个构象被用于进一步分析。
- 对于结合目标，我们考虑化验 P450（CYP）2C19 [20]（ChEMBL id: 1613777），并选择蛋白质数据库（PDB）中的相应晶体结构（PDB id: 4GQS）。然后，我们选择链 A 进行对接运行。之后进行结合时，将结合口袋与 PDB 复合物晶体结构中的原始配体对齐：中心设置为 (-81.48, 16.55, -41.6)，方框为 (20.0, 23.0, 25.0)。
- 然后，我们进行预处理，补充氢原子并添加部分电荷。我们使用 meeko v0.3.3 处理小分子，使用 AutoDock Flexible Receptor (ADFR) suite v1.2 处理蛋白质。
- 对接时，我们使用 AutoDock Vina v1.2.3 [21]。每个分子构象都以 *exhaustiveness* 为 32 进行对接，选出对接得分最高（最低）的姿态用于可视化。在可视化方面，我们使用了 UCSF Chimera。

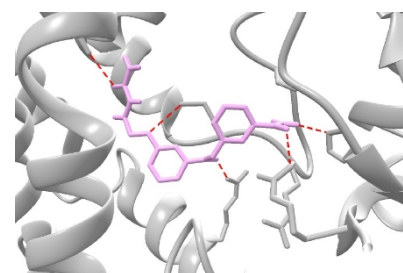
(a) 第 1 组，输入分子（SMILES）：Cc1cc(F)cc(C(=O)Oc2ccccc(C(N)=O)c2)c1



输入分子（对接得分：
-9.055）

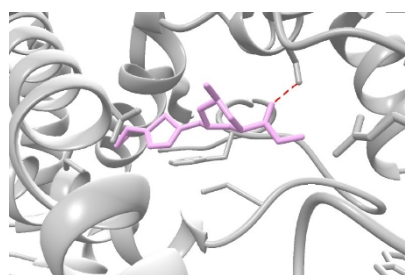


带 GS 的输出分子（文档
评分：-8.843）

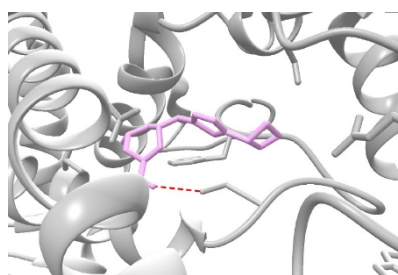


用 MoleculeSTM 输出分子（对接得
分：-10.35）

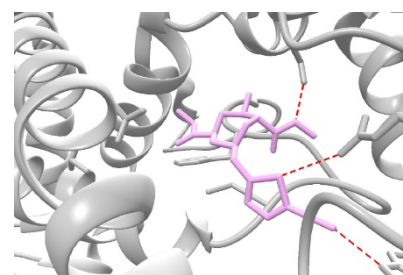
(b) 第 2 组，输入分子（SMILES）：COC(=O)[C@@H]1CN(Cc2cnc(C3CC3)s2)C[C@@H](C)O1



输入分子（对接得分：
-7.441）



带 GS 的输出分子（文档
评分：-7.747）



用 MoleculeSTM 输出分子（对接得
分：-11.363）

图 9.两组基于结合亲和力的分子编辑对接可视化图。文本提示来自 ChEMBL 1613777（“该分子在酶蛋白的抑制剂和底物的检测中呈阳性。它利用分子氧将一个氧原子插入底物，并将第二个氧原子还原成水分子。”）为便于可视化，显示了输入分子和输出分子与 GS 和 MoleculeSTM 的关系。可以看出，MoleculeSTM 可以生成对接得分最低的分子（氢键最多，用红色虚线标出）。在第 1 组（a）中，输出分子共享相同的分子支架。在第 2 组（b）中，使用 MoleculeSTM 生成的分子的主题也发生了变化。

D.5 药物相关性编辑

作为概念验证，我们进一步对常见药物编辑任务进行了四项编辑。这里使用的文本提示是让输入的分子看起来像现有的药物，例如 "这个分子看起来像青霉素"。在 "方法 "一节中，使用的满足函数是 Tanimoto 相似度，阈值 Δ 值为 0 和 0.05。

表 23.四种常见药物分子编辑结果。输入是从 ZINC 中随机抽取的 200 个分子，评价指标是与常见药物的 Tanimoto 相似度增加

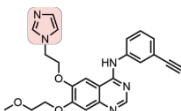
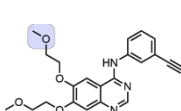
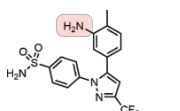
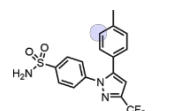
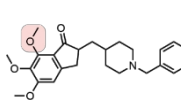
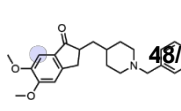
		基质优化					
		随机	PCA	高差异	GS-Mutate	微笑	图表
这种分子 看起来像青霉素。	0	43.61 \pm 2.23	46.51 \pm 3.02	44.42 \pm 3.56	28.67 \pm 0.94	58.13 \pm 0.97	50.91 \pm 2.80
	0.05	0.69 \pm 0.55	0.23 \pm 0.32	0.89 \pm 0.30	0.67 \pm 0.62	11.01 \pm 0.58	3.64 \pm 0.57
这种分子 看起来像阿司匹林。	0	43.82 \pm 1.41	43.12 \pm 5.35	44.63 \pm 3.33	25.00 \pm 2.16	40.13 \pm 1.33	54.05 \pm 3.58
	0.05	2.99 \pm 0.38	3.08 \pm 0.82	2.45 \pm 0.33	0.33 \pm 0.47	4.28 \pm 1.22	10.84 \pm 1.26
这种分子 看起来像咖啡因。	0	42.71 \pm 3.16	40.33 \pm 0.71	40.64 \pm 3.89	26.17 \pm 1.31	46.08 \pm 3.81	51.01 \pm 1.22
	0.05	0.69 \pm 0.01	0.23 \pm 0.32	0.44 \pm 0.31	0.33 \pm 0.24	1.61 \pm 0.67	0.61 \pm 0.01
这种分子 看起来像多巴胺。	0	42.00 \pm 3.08	42.50 \pm 2.12	41.33 \pm 2.86	30.50 \pm 1.63	47.00 \pm 4.11	55.50 \pm 2.73
	0.05	0.00 \pm 0.00	0.44 \pm 0.31	0.22 \pm 0.31	0.83 \pm 0.24	2.30 \pm 0.44	6.24 \pm 0.56

D.6 专利药物分子邻域检索案例研究

为了证明基于文本的分子编辑的实用性，我们展示了三个从类似物生成已批准药物的案例研究。先导分子优化是药物发现的一个关键阶段，在这一阶段，以先导分子为基础，制造密切相关的化合物，旨在改善其药效和 DMPK（药物代谢和药代动力学）特性，最终确定候选药物[22]。因此，要求获得更多类似药物特性的文本提示将有助于改进先导分子的不足之处，加速药物发现研究。

具体而言，输入分子是每个已批准药物分子的专利类似物，输入文本提示是单目标的，如附录 D.2 所示。这里的目标是检查是否能成功生成已获批准的药物作为输出分子，其结构变化与文本提示中反映的性质改进相一致。例如，在表 24 (a)中，厄洛替尼通过将咪唑取代基替换为甲氧基[23]，成功地从类似物生成。这一变化反映出 tPSA 从 83 降至 75，与文本提示一致，表明渗透性更高。表 24(b)是塞来昔布的氨基取代衍生物[24]，去掉氨基后，分子的肠道渗透性更强，生物利用度更高。生物利用度是药物分子进入全身循环的部分，是口服药物吸收的关键因素[25]。最后，表 24(c)说明了如何通过基于文本的编辑来解决分子中潜在的代谢问题。一个要求代谢稳定的分子的文本成功地将多奈哌齐（Donepezil）中的三甲氧基炔变成了二甲氧基炔[26]，前者代表一种富电子的芳香族化合物，已知会发生氧化I期代谢[27]。

表 24.根据文本提示生成已批准药物的三个药物类似物单目标分子编辑的可视化效果。粉色和蓝色区域分别突出显示了输入和输出分子中的修改结构。

(a) 提示：这种分子具有高渗透性。		(b) 提示：这种分子的生物利用率很高。		(c) 提示：这种分子代谢稳定。	
输入分子	输出分子	输入分子	输出分子	输入分子	输出分子
					

CAS:
(Donepezil)

183320-43-6Tarceva (Erlotinib) CAS:

170570-28-2Celebrex (Celecoxib) CAS:

120013-52-7Aricept

E 下游：分子特性预测

在本节中，我们将回顾用于分子性质预测下游任务的两大类数据集，它们分别来自 MoleculeNet 和分子基准工作 [28, 29]。

分子特性：血脑屏障穿透（BBBP） [30] 数据集测量分子是否会穿透中枢神经系统。Tox21 [31]、ToxCast [28] 和 ClinTox [32] 这三个与毒性相关的数据集都与分子化合物的毒性有关。副作用资源（SIDER）[33] 数据集存储了上市药物数据库中的药物不良反应。

分子特性：生物物理学最大无偏验证（MUV） [34] 是 PCBA 的另一个子数据库，是通过精炼的近邻分析获得的。HIV 来自药物治疗计划（DTP）艾滋病抗病毒筛选[35]，旨在预测对 HIV 复制的抑制作用。BACE 衡量的是 β -分泌酶 1（BACE-1）抑制剂的结合结果，收集于 MoleculeNet [28]。

表 25.分子化学数据集摘要。

数据集	任务	# 任务	# 分子
BBBP	分类	1	2,039
Tox21	分类	12	7,831
ToxCast	分类	617	8,576
Sider	分类	27	1,427
临床毒理学	分类	2	1,478
MUV	分类	17	93,087
艾滋病病毒	分类	1	41,127
贝丝	分类	1	1,513

在数据拆分方面，我们采用了脚手架拆分法[28]。脚手架衡量的是分子的骨架结构，脚手架拆分是指我们将具有较常见脚手架的分子放入训练，其余的放入验证和测试，以模拟分布外（OOD）环境。OOD 环境在实际场景中更为常见，因此更适合用来测试预训练的分子表征能力。

实现细节 对于 SMILES 字符串，我们使用 MegaMolBART [5] 作为骨干变换器模型。对于分子图，我们使用相同的骨干 GIN 模型，并使用丰富的特征（用于 GraphMVP [8] 中的回归任务）。下面我们将列出主要的超参数。

表 26.分子性质预测的超参数规格。

	超参数值	历时	{100}
培训前基线	学习率		{1e-3}
	权重衰减		{0}
下游	纪元		{100}
	学习率	{1e-3}	{5e-4}
	权重衰减		{0}

骨干模型的选择。我们想说明的是，MoleculeSTM 与每种模式的骨干编码器（如分子表示模型）无关。

- 在骨干模型方面，我们使用 GIN 模型作为固定的二维 GNN 骨干编码器。换句话说，MoleculeSTM 的性能受到二维主干模型的限制。
- 在分子预训练研究领域（如 AttrMask [36]、MolCLR [37]、GraphMVP [8]、MoleculeSDE [38]），所有这些工作都采用 GIN 作为二维骨干模型，作为对照来测试各种预训练算法的有效性。我们提出的 MoleculeSTM 也

是类似的情况。

- 未来，我们希望探索更先进的分子 GNN 模型。

补充参考资料

- [1] Sunghwan Kim、Jie Chen、Tiejun Cheng、Asta Gindulyte、Jia He、Siqian He、Qingliang Li、Benjamin A Shoemaker、Paul A Thiessen、Bo Yu 等：《2021 年的 PubChem：新数据内容和改进的网络界面》。In: *Nucleic acids research* 49.D1 (2021), pp.
- [2] Iz Beltagy、Kyle Lo 和 Arman Cohan。"SciBERT：科学文本的预训练语言模型"。In: *EMNLP.2019*. 电子版：[arXiv:1903.10676](https://arxiv.org/abs/1903.10676).
- [3] Waleed Ammar、Dirk Groeneveld、Chandra Bhagavatula、Iz Beltagy、Miles Crawford、Doug Downey、Jason Dunkelberger、Ahmed Elgohary、Sergey Feldman、Vu Ha 等：《语义学者中文献图的构建》。In: *arXiv preprint arXiv:1805.02262* (2018).
- [4] M Honnibal、I Montani 和 S Van Landeghem。"博伊德"。A. spaCy: industrial-strength natural language processing in Python: *A. spaCy: 用 Python 进行工业级自然语言处理* (2020).
- [5] Ross Irwin、Spyridon Dimitriadis、Jiazhen He 和 Esben Jannik Bjerrum。"Chemformer: a pre-trained transformer for computational chemistry".In: *机器学习：科学与技术* 3.1 (2022), 第 015022 页。
- [6] Teague Sterling 和 John J Irwin."ZINC 15--人人都能发现的配体"。In: *化学信息与建模杂志* 55.11 (2015), 第 2324-2337 页。
- [7] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka."神经网络有多强大？"In: *arXiv preprint arXiv:1810.00826* (2018).
- [8] Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang."用三维几何预训练分子图表示法"。In: *学习表征国际会议*. 2022.URL: <https://openreview.net/forum?id=xQUelpOKPam>.
- [9] Simon Axelrod 和 Rafael Gomez-Bombarelli。"GEOM，用于性质预测和分子生成的能量注释分子构象"。In: *科学数据* 9.1 (2022), 第 1-14 页。
- [10] 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"。In: *arXiv preprint arXiv:1810.04805* (2018).
- [11] Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。"对比预测编码的表征学习"。In: *arXiv preprint arXiv:1807.03748* (2018).
- [12] Kaveh Hassani 和 Amir Hosein Khasahmadi。"图上的对比多视图表示学习"。In: *International Conference on Machine Learning*.PMLR.2020, 第 4116-4126 页。
- [13] Chitwan Saharia、William Chan、Saurabh Saxena、Lala Li、Jay Whang、Emily Denton、Seyed Kamyar Seyed Ghasemipour、Burcu Karagol Ayan、Sara Mahdavi、Rapha Gontijo Lopes 等：《具有深度语言理解能力的逼真文本到图像扩散模型》。In: *arXiv preprint arXiv:2205.11487* (2022).
- [14] David S Wishart、Yannick D Feunang、An C Guo、Elvis J Lo、Ana Marcu、Jason R Grant、Tanvir Sajed、Daniel Johnson、Carin Li、Zinat Sayeeda 等：《DrugBank 5.0：DrugBank 数据库 2018 年重大更新》。核酸研究 *核酸研究* 46.D1 (2018)、pp.D1074-D1082.
- [15] Tero Karras、Samuli Laine 和 Timo Aila。"基于风格的生成式对抗网络生成器架构"。In: *IEEE/CVF 计算机视觉与模式识别会议论文集*。2019, pp.
- [16] 刘胜超、王成鹏、聂伟力、王汉臣、陆家瑞、周伯磊和唐健。"GraphCG：无监督发现图中的可引导因子"。In: *NeurIPS 2022 研讨会：图学习的新前沿*。2022.URL: https://openreview.net/forum?id=BhR44NzeK_1.

- [17] David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, María Gordillo-Marañón、菲奥娜-亨特、劳拉-琼科、格蕾丝-穆古姆巴特、米拉格罗斯-罗德里格斯-洛佩斯、弗朗西斯-阿特金森、尼古拉斯-博斯克、克里斯-J-拉杜、阿尔多-塞古拉-卡布雷拉、安妮-赫西和安德鲁-R-利奇。"ChEMBL: towards direct deposition of bioassay data".In: *Nucleic Acids Research* 47.D1 (Nov. 2018), pp.ISSN: 0305-1048.DOI: [10.1093/nar/gky1075](https://doi.org/10.1093/nar/gky1075). eprint: <https://academic.oup.com/nar/article-pdf/47/D1/D930/27437436/gky1075.pdf>。URL: <https://doi.org/10.1093/nar/gky1075>。
- [18] 托马斯-A-哈尔格伦"默克分子力场。I. MMFF94 的基础、形式、范围、参数化和性能"。In: *Journal of computational chemistry* 17.5-6 (1996), pp.
- [19] Greg Landrum et al.: *用于化学信息学、计算化学和预测建模的软件套件*。2013。
- [20] R Leila Reynald、Stefaan Sansen、C David Stout 和 Eric F Johnson。"人类细胞色素 P450 2C19 的结构特征: P450s 2C8、2C9 和 2C19 的活性位点差异"。In: *生物化学杂志* 287.53 (2012)、pp.44581-44591。
- [21] Oleg Trott 和 Arthur J Olson。"AutoDock Vina: 利用新的评分函数、高效优化和多线程提高对接速度和准确性"。In: *计算化学杂志* 31.2 (2010 年) , 第 455-461 页。

- [22] James P Hughes、Stephen Rees、S Barrett Kalindjian 和 Karen L Philpott. "早期药物发现的原则"。In: *British journal of pharmacology* 162.6 (2011), pp.
- [23] Rodney Caughren Schnur 和 Lee Daniel Arnold. 炔基和叠氮取代的 4-氨基喹唑啉。美国专利 5,747,498.1998 年 5 月。
- [24] John J Talley、Thomas D Penning、Paul W Collins、Donald J Rogier Jr、James W Malecha、Julie Miyashiro、Stephen R Bertenshaw、Ish K Khanna、Matthew J Graneto、Roland S Rogers 等用于治疗炎症的取代吡唑苯磺酰胺类药物。美国专利 5,760,068.美国专利 5,760,068.
- [25] David Dahlgren 和 Hans Lennernäs. "肠道渗透性和药物吸收：预测性实验、计算和体内方法"。In: *Pharmaceutics* 11.8 (2019) : *Pharmaceutics* 11.8 (2019).ISSN: 1999-4923.DOI: [10.3390/pharmaceutics11080411](https://doi.org/10.3390/pharmaceutics11080411).URL: <https://www.mdpi.com/1999-4923/11/8/411>.
- [26] Hachiro Sugimoto、Youichi Iimura、Yoshiharu Yamanishi 和 Kiyomi Yamatsu. "乙酰胆碱酯酶抑制剂的合成与结构活性关系：1-苄基-4-[(5,6-二甲氧基-1-氧代茚满-2-基)甲基]哌啶盐酸盐及相关化合物"。In: *药物化学杂志* 38.24 (1995 年)。PMID: 7490731, pp.DOI: [10.1021/jm00024a009](https://doi.org/10.1021/jm00024a009). eprint: <https://doi.org/10.1021/jm00024a009>.URL: <https://doi.org/10.1021/jm00024a009>.
- [27] Gordon Guroff、Jean Renson、Sidney Udenfriend、John W Daly、Donald M Jerina 和 Bernhard Witkop. "羟化诱导迁移：NIH 转变"：最近的实验揭示了芳香族化合物酶促羟基化的一个意想不到的普遍结果"。In: *科学* 157.3796 (1967)，第 1524-1530 页。
- [28] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande."分子网：分子机器学习的基准"。In: *化学科学* 9.2 (2018)，第 513-530 页。
- [29] 王瀚辰、Jean Kaddour、刘胜超、唐健、Matt Kusner、Joan Lasenby 和刘琦。分子图嵌入的自我监督学习评估。2022.URL: <https://openreview.net/forum?id=ctX2eXYIW3>.
- [30] Ines Filipa Martins、Ana L Teixeira、Luis Pinheiro 和 Andre O Falcao. "硅学血脑屏障渗透建模的贝叶斯方法"。In: *化学信息与建模杂志* 第 52.6 期 (2012 年)，第 1686-1697 页。
- [31] Tox21 数据挑战赛。"2014年Tox21数据挑战赛"。见: <https://tripod.nih.gov/tox21/challenge/> (2014)。
- [32] Kaitlyn M Gayvert、Neel S Madhukar 和 Olivier Elemento. "预测临床试验成败的数据驱动方法"。细胞化学生物学 23.10 (2016): *细胞化学生物学* 23.10 (2016)，第 1294-1301 页。
- [33] Michael Kuhn、Ivica Letunic、Lars Juhl Jensen 和 Peer Bork. "药物和副作用 SIDER 数据库"。In: *核酸研究* 44.D1 (2015)，第 D1075-D1079 页。
- [34] Sebastian G. Rohrer 和 Knut Baumann. "基于 PubChem 生物活性数据的虚拟筛选最大无偏验证 (MUV) 数据集"。In: *Journal of Chemical Information and Modeling* 49.2 (2009).PMID: 19161251, pp.DOI: [10.1021/ci8002649](https://doi.org/10.1021/ci8002649). eprint: <https://doi.org/10.1021/ci8002649>.URL: <https://doi.org/10.1021/ci8002649>.
- [35] 丹尼尔-扎哈雷维茨艾滋病抗病毒筛选数据。2015.
- [36] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec."预训练图神经网络的策略"。In: *学习表征国际会议, ICLR.2020*.
- [37] 王宇阳、王建仁、曹中林和 Amir Barati Farimani. "Molclr: 通过图神经网络进行表征的分子对比学习"。In: *arXiv preprint arXiv:2102.10056* (2021).
- [38] Shengchao Liu, Weitao Du, Zhi-Ming Ma, Hongyu Guo, and Jian Tang."分子多模态预训练的组对称随机微分方程模型"。In: *国际机器学习大会*. PMLR.2023, pp.