预测分子特性的多模态学习：基于图像和图形结构的框架

Zhuoyuan Wang[1] , Jiacong Mi[1] , Shan Lu[2] , Jieyue He[1,*]
[1]东南大学计算机科学与工程学院，教育部计算机网络与信息集成重点实验室，江苏南京，210018
[2]南京风火天地通信技术有限公司，中国江苏省南京市，211161
wangzhuoy@seu.edu.cn, mijiacong@seu.edu.cn, bfcat.cn@gmail.com, jieyuehe@seu.edu.cn

***摘要*--在人工智能药物发现（AIDD）领域，准确预测药物分子特性是一项基本挑战。药物分子的有效表征是实现这一目标的关键要素。当代的前沿研究主要采用自监督学习（SSL）技术，从大规模、无标记的分子数据中提取有意义的结构表征，然后针对一系列下游任务对这些表征进行微调。然而，这些研究的一个固有缺陷在于它们只依赖一种分子信息模式，如分子图像或 SMILES 表征，从而忽视了各种分子模式的潜在互补性。针对这一局限性，我们提出了 MolIG 模型，这是一种基于图像和图形结构预测分子特性的新型 MultiModaL 分子预训练框架。MolIG 模型创新性地利用了分子图和分子图像之间的一致性和相关性来执行自我监督任务，有效地融合了两种分子表示形式的优势。这种整体方法可以捕捉关键的分子结构特征和高级语义信息。预训练完成后，图形神经网络（GNN）编码器将用于下游任务的预测。与先进的基线模型相比，MolIG 在 MoleculeNet Benchmark Group 和 ADMET Benchmark Group 等基准组中与分子性质预测相关的下游任务中表现出更高的性能。**

***索引词条*--*分子*性质预测、对比学习、分子图、分子图像**

I. 导言

在计算化学和药物发现领域，分子表征学习是一项重要任务，旨在开发有效的方法来表征分子结构并预测其性质[1]-[3]。准确预测化学分子的性质可以极大地促进药物开发过程，降低研究成本，提高新药的成功率，为药物设计提供更多的信息和指导，具有重要的实用价值[4]。早期的分子表征学习方法主要包括基于拓扑学的方法和基于物理化学的方法。基于拓扑学的方法通过分析分子结构中原子间的化学键连接来描述分子，而基于物理化学的方法则通过分析分子结构中原子间的化学键连接来描述分子。扩展连接指纹（ECFP）[5] 是最经典的方法之一。ECFP 将原子周围的邻接信息转换成固定长度的二进制字符串作为特征表示。另一方面，基于物理化学的方法则通过计算分子的物理化学性质（如电荷状态、极性、电子亲和力、电离能等）来描述分子结构。这些方法需要事先计算理化性质，然后将其用作特征表示。例如分子量子力学（MQM）、分子力学（MM）、密度泛函理论（DFT）[6] 等。然而，早期的方法缺乏对分子结构的深刻理解和捕捉复杂分子特征的能力，因此在分子性质预测任务中存在一定的局限性。

近年来，图神经网络（GNN）的出现为一系列与图相关的任务带来了显著的进步[7]，从而激发了将其应用于分子结构学习的灵感。基于分子结构的 GNN 模型概念的核心是将分子内原子和化学键的拓扑结构视为图，其中原子和化学键分别对应于节点和边。初始特征集是根据原子类型、化学键类型等固有物理化学特性制定的，并通过相邻节点之间的迭代信息交换执行聚合操作[8]。与传统的基于描述符的方法相比，GNN 可以囊括更广泛的分子特征，包括但不限于局部相互作用和环状结构，从而提高预测的精确度。迄今为止，已经提出了许多基于 GNN 的分子表征学习方法，如消息传递神经网络（MPNN）[9] 和注意力 FP [4] 等。

与此同时，随着 GNN 的发展，图对比学习 [10] 被应用到分子表征学习领域，弥补了标注分子数据的 不足，极大地推动了这一领域的发展。现有的基于图对比学习的分子表征学习方法通常是

采用分子图增强策略。然 而 ，由于分子结构具有特定的化学规则和限制，分子领域的数据增强策略并不简单，这可能需要额外的领域知识和经验来设计合适的分子增强策略。除了将分子视为节点和边的拓扑视图外，分子还可以以图像的形式呈现[11]-[13]。图形模式的输入为模型提供了分子结构的 详细信息，如化学键类型和原子类型。这种明确的信息传递使模型能够直观地理解分子的微观方面。然而，必须注意的是，在图神经网络中，每一层都试图通过聚合来自邻近节点的信息来更新节点的特征表示。这种聚合往往会导致节点之间的高度相似性，造成过度平滑，从而限制了处理复杂结构时的表现力。从这个意义上说，图形模态可以被视为一种局部模态，强调微观细节的表达。相比之下，图像模态的输入并不能为原子化学键等细节提供直接指导。虽然模型并不了解分子结构的微观细节，但卷积运算的全局性使每个像素都能感知图像的整体信息。这种全局视角可以帮助模型更好地理解整体环境，而无需关注微观细节，从而避免过度平滑的问题。因此，图像模式可被视为一种全局模式，侧重于捕捉整体结构。

虽然以前基于图形或 SMILES 的模型取得了不错的成绩，但单一模态提供的信息是有限的。受计算机视觉（CV）和自然语言处理（NLP）领域利用大量无标记数据取得巨大成功的启发，我们将分子图像模态作为一种增强模态引入对比学习，并提出了一种名为 MolIG 的新型分子预训练框架。 考虑到分子图和分子图像可以提供不同层次的化学和几何信息，我们捕捉它们之间的一致性来捕捉分子的高阶语义特征。与之前的分子对比学习模型[14]相比，我们的方法从不同角度研究分子特征，并充分利用两种模式的信息。我们的贡献可总结如下：

- 我们率先提出了一种专用于分子特性预测的多模态预训练模型，并将分子图和分子图像作为两种模态对该模型进行训练。
- 为确保保留分子语义特征，我们采用了三种不同的图像增强策略。通过这种方法，我们最大限度地提高了分子图和图像模式之间的一致性，从而实现了一种更强大、更普遍适用的表示方法学习。
- 我们的模型在 Molecu- leNet 基准组和 ADMET 基准组上进行了评估。实验结果表明，我们的模型不仅能熟练提取分子的结构属性，还能捕捉到更难以捉摸的高阶语义信息。此外，它在结果上还超越了依赖分子图和分子图像的最先进方法。

II. 相关工作

A. 基于图形的分子表示法

目前，基于图的分子表征学习是最主要的方法。GROVER [15] 将 GNN 和变换器巧妙地结合在一起，通过预测上下文属性和主题信息生成节点嵌入。与此同时，对比学习也被广泛应用于分子表征学习领域。相反，MolCLR[14]采用了一种与众不同的方法，对分子图应用随机增强操作，包括节点屏蔽、边缘删除和子图移除。KANO [16] 利用已知边缘图指导分子图增强。然而，这些操作不可避免地会改变分子结构，从而与既定的化学原理相冲突。

B. 基于图像的分子表示法

ADMET-CNN [11] 成功建立了一个基于二维分子图像的卷积神经网络（CNN）模型，在预测 ADMET 属性方面取得了卓越成果。基于分子图像的表征学习方法必须将数据样本转换到欧几里得空间。然而，由于缺乏与原子和键相关的属性，使用分子图像直接预测属性的效果并不理想。

C. 基于多模态的分子表示法

GeomGCL [17] 设计了一个几何信息传递网络，利用二维和三维双视图，自适应地利用二维和三维图形中的大量信息。MoleculeSTM [18] 从 PubChem 收集了大量描述性分子文本，在这些叙述和相应的分子图之间架起了一座坚实的一致性桥梁。不过，据我们所知，目前还没有一种多模态表征学习方法能将分子结构特征与图像信息结合起来。

III. 方法

本文提出了一个多模态分子预训练框架（MolIG），用

于基于图像和图形结构预测分子结构，其架构如图 1 所示。该框架由五个部分组成：图形编码器、图像编码器、图形非线性投影、图像非线性投影和对比学习。我们将首先详细介绍

网络结构，然后解释下游任务的推理。

A. 图形和图像编码器

本节将介绍 分子图和分子图像的编码器。

1) 图编码器： 分子图 G 可定义为 $G = (V, E)$，其中每个节点 $v \in V$ 和每条边 $e_{uv} \in E$ 表示原子 u 和 v 之间的化学键。

$$a^{(k)} = AGGREGATE\left(h^{(k)}_{(k-1)} : u \in N(v)\right), \quad (1)$$

$$h_v^{(k)} = COMBINE\left(h^{(k)}_{v}{}^{(k-1)}, a^{(k)}\right) \quad (2)$$

$N(v)$ 表示节点 v 所有邻居的集合，h^k 表示原子 v 在第 k 层的表示形式。之后 k 次迭代，h^k 就能捕捉到其 k-跳邻域。AGGREGATION 函数整理来自 v 邻近节点的信息，COMBINE 函数更新聚合特征。初始代表 h^0 由节点特征 x_v 初始化。值得注意的是，在本研究中，我们只使用了两种原子属性，即原子类型和手性。同样，对于化学键，我们使用键类型和方向作为初始特征。

为进一步提取图层特征 h_G，读出操作会整合图 G 中所有节点的特征，如公式 (3) 所示：

$$h_G = READOUT\left(h^{(k)}_{v} : v \in G\right), \quad (3)$$

在这项工作中，我们采用图形同构网络（GIN）[19] 作为 GNN 编码器。然而，由于分子图不同于其他类型的图结构，边缘信息对下游任务有很大影响，而原始的 GIN 并没有考虑到这一点。为此，我们效仿 Hu 等人的研究 [20]，扩展了节点聚合功能。到 $h_v^{(k)} = \sum_{u \in N(v)} \sigma\left(h^{(k-1)} + e_{uv}\right)$ 它同时考虑了同时获得节点和边缘信息。这里，$\sigma(\cdot)$ 是一个非线性激活函数。读出操作是一个平均池化函数，用于获得每个分子的图级表示。GIN 网络有 500 万个参数。

2) 图像编码器： 对于分子图像，我们采用 ResNet [21] 作为编码器，从每个分子中提取特征

B. 图形和图像投影头

投影头是 MolIG 模型的关键组件，它将从图形编码器和图像编码器提取的高维特征向量映射到低维空间。具体来说，投影头可被视为多层感知器（MLP），其目的是在低维空间中学习更具区分性的表征，以进行对比学习。我们沿用了 [23] 中提到的非线性 MLP，在隐层中加入了 Relu 激活函数。考虑到分子图和分子图像之间的巨大差异，我们部署了两个相同的、但不共享参数的投影头。这

这种方法有助于在以下下游任务中取得优异成绩
分子特性预测。投影头的形式化过程如下：

$$z_i^{graph} = PrejectionHead\left(h_{i}^{示意图}\right) \quad (4)$$

$$z^{image} = PrejectionHead\left(h_{image}\right) \quad (5)$$

其中，$PrejectionHead_{graph}$ 和 $PrejectionHead_{image}$ 分别表示 GNN 编码器和图像编码器的非线性投影头。

C. 对比学习框架

具体来说 鉴于 a 批 的 N 个分子 $\{M_0, M_1 \ldots M_{n-1}\}$，我们首先将它们转换成图 $\{G, G_{01}, G_2 \ldots G_{n-1}\}$ 和 图像 $\{i_0, i_1, i_2 \ldots I\}_{n-1}$ 通过 Rdkit [24]。分子图像分辨率为 224x244。来自同一分子的图形和图像构成正样本，而来自不同分子的图形则被视为负样本。基于 GNN 的图编码器对 G h^{graph}，而基于 ResNet 的图像编码器将 I_i 编码为 h^{image}。然而，由于两种模式之间存在显著差异，直接使用 h^{graph} 和 h^{image} 作为对比度损失的输入往往会导致下游任务的性能不达标。为了解决这个问题，我们为图模式和图像模式设计了独立的映射头，分别映射分子和图像。图和分子图像到同一投影空间，以便计算对比损失，即 z^{graph} 和 z^{image}。

我们的目标是最大限度地提高正样本分子在图形和图像模式下的表示相似性。因此，在图-像视图下，按照 SimCLR [23]，我们使用归一化温度标度交叉熵（NT-Xent）作为损失函数：

图像。在我们的任务中，由于 ResNet-34. 我们最终选择了它作为图像主干。它值得

注意到分子图像与 ImageNet 之间存在很大差异[22]，因此我们不使用在 ImageNet 上预先训练好的 ResNet。Resnet 网络有 2200 万个

参数。相反，我们从头开始训练图像编码器，以适应分子图像的独特属性。

我们在实验中验证了这一决定。

$$L_{i,j} = -\log \frac{\exp(\text{模拟}(z_i^{\text{示意}}, z^{\text{图像}})/\tau)}{\sum_{k=1}^{2N} 1(k \neq i)\exp(\text{模拟}(z^{\text{图}}, z^{\text{像}})/\tau)} \quad (6)$$

其中，τ 表示温度系数，我们将相似性计算为 $sim(z_i, z_j) = \frac{z_i^T z_j}{\|z_i\|_2 \|z_j\|_2}$.最终损耗在一个小批次内计算。

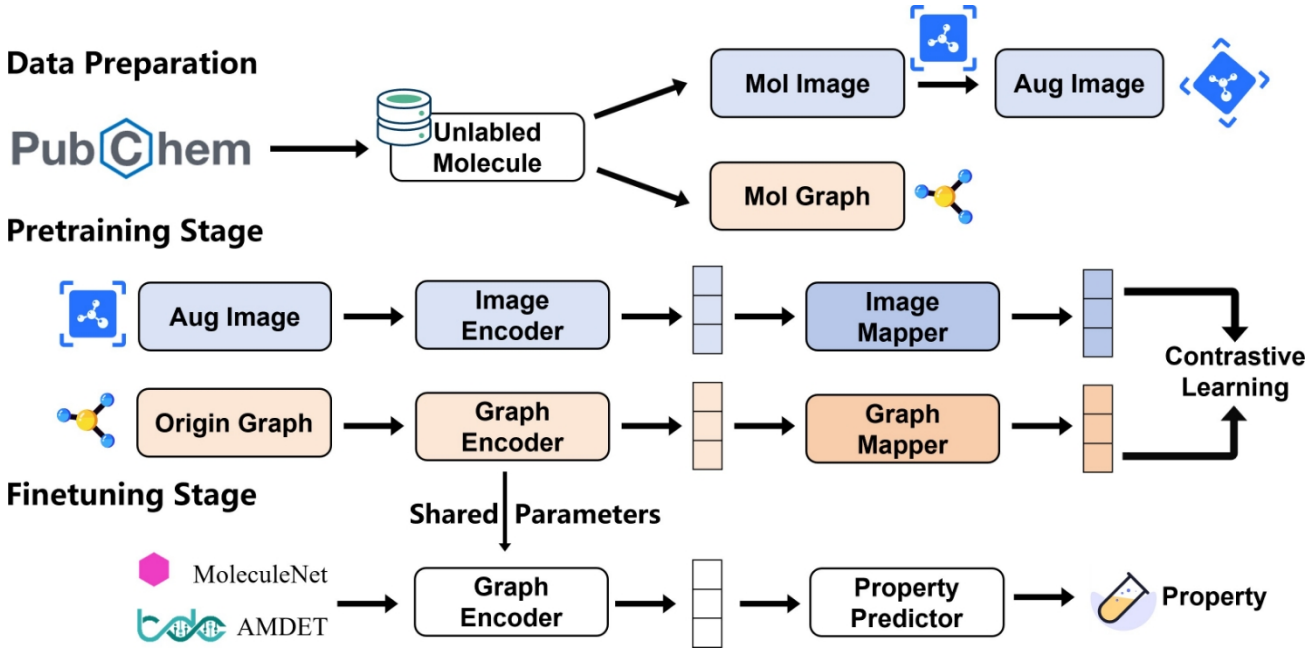

图 1.我们的方法概览：MolIG.

D. 微调和下游推理

在本研究中，我们采用预训练加微调的策略来完成药物性质预测任务。在预训练阶段，我们舍弃了图像编码器，只保留图形编码器作为药物分子编码器，利用预训练模型来学习药物的分子表示。随后，我们针对药物特性预测任务对预训练模型进行微调，使其更好地适应特定任务的数据分布。具体来说，我们只在图形编码器之后增加了一个全连接层 PredictionHead(w)，其中 w 代表全连接层的参数。在微调过程中，我们首先将药物的分子图（$Drug_i$）输入 GNN，以获得分子的潜在表示（h_i），然后将其转发给 PredictionHead，以进行最终的属性预测，形式化如下：

$$\hat{y} = PredictionHead\,(GNN\,(Drug))_i \qquad (7)$$

IV. 实验

A. 预训练数据集和数据扩充

对于预训练数据集，我们从 Pubchem 数据库 [25] 中的 1.14 亿个分子中随机选取 1000 万个未标记的类药物分子。然后，我们按 0.95/0.05 的比例将预训练数据集分为训练集和验证集。数据增强是提高模型泛化能力和鲁棒性的有效方法。通过在训练过程中引入更多的数据变化，模型可以更好地适应现实世界中遇到的复杂情况。数据增强已被广泛应用于计算机视觉和多模态领域。

但是，对于分子图来说，随机原子屏蔽、边缘扰动或子图采样等常见的增强技术会破坏分子结构信息，导致学习到的表征与下游任务中的数据不匹配。因此，我们在预训练阶段不对分子图进行任何增强操作。相比之下，分子图像与真实世界的图像相比更为稀疏，90% 以上的区域都是零，这意味着只有极少部分像素与下游任务真正相关[13]。有鉴于此，我们在 预训练阶段选择了三种增强策略：（1）随机水平翻转（RandomHorizontalFlip）；（2）随机灰度（RandomGrayscale）；（3）随机旋转（RandomRotation）。每种策略的执行概率为 25%。这些策略不会改变分子图像的结构，并使模型能够学习数据增强带来的不变性。

B. 分子网上的分子特性预测

*1) 数据集*为了评估 MolIG 的性能，我们在 MoleculeNet [2] 的六个分类数据集（包括 BBBP、BACE、SIDER、Tox21、HIV、ToxCast 和 ClinTox）上对预训练模型进行了微调。与随机分割相比，支架分割是一种更具挑战性和现实性的分割方法，极大地考验了模型的鲁棒性和泛化能力。我们采用支架拆分法将上述六个数据集按 0.8/0.1/0.1 的比例分为训练集/验证集/测试集。

*2) 基准：*我们将 MolIG 与最先进的分子特性预测基准模型进行了比较，其中包括 GCN [26]、GIN [19] 和 Attentive FP [4] 等监督学习模型。其余模型均为预训练模型。

表 I

在六个分类基准上测试不同模型的性能。前两个模型是监督学习方法，后八个模型是自我监督/预训 练 方 法 。每个模型的测试 ROC-AUC 的平均值和标准偏差（%）。

报告了基准。

日期集	BBBP	计算机设备行动伙伴关系	西德	Tox21	艾滋病毒	ToxCast	临床毒理学	平均值
分子	2,039	1,513	1,427	7,831	41,127	8,575	1,478	-
任务	1	1	27	12	1	617	2	-
分割	脚手架	脚手架	脚手架	脚手架	脚手架	脚手架	脚手架	-
GCN	71.8(0.1)	71.6(2.0)	53.6(3.2)	70.9(2.6)	74.0(3.0)	60.1(1.3)	62.5(2.8)	66.3
GIN	65.8(4.5)	70.1(5.4)	57.3(1.6)	74.0(0.8)	75.3(1.9)	62.2(1.9)	58.0(4.4)	66.1
细心的 FP	64.3(1.8)	78.4(0.1)	60.6(3.2)	76.1(0.5)	75.7(1.4)	63.7(0.2)	84.7(0.3)	71.9
MFBERT	71.6(3.3)	71.6(4.4)	61.1(7.4)	63.9(4.8)	71.1(2.4)	63.7(8.3)	77.9(10.2)	68.7
BARTSmiles	70.9(3.7)	83.2(3.3)	57.6(6.9)	65.1(5.0)	70.9(2.6)	64.9(8.6)	79.3(7.5)	70.3
图表	67.5(3.3)	68.7(7.8)	60.1(1.3)	74.4(0.7)	75.0(0.4)	63.0(0.4)	78.9(4.2)	69.6
GraphMVP	68.5(0.2)	76.8(1.1)	62.3(1.6)	74.5(0.4)	74.8(1.4)	62.7(0.1)	79.0(2.5)	71.2
3DInfoMax	69.1(1.0)	79.4(1.9)	53.3(3.3)	74.5(0.7)	76.1(1.3)	63.5(0.8)	59.4(3.2)	67.9
格罗夫	68.0(1.5)	79.5(1.1)	60.7(0.5)	76.3(0.6)	77.8(1.4)	63.4(0.6)	76.9(1.9)	71.8
MGSSL	69.7(0.9)	79.1(0.9)	61.8(0.8)	76.5(0.3)	78.8(1.2)	63.3(0.5)	80.7(2.1)	72.8
MolCLR	71.6(0.7)	81.9(1.5)	59.9(0.9)	75.0(0.4)	78.3(0.4)	64.7(0.1)	81.9(1.2)	73.3
鼹鼠-伯特	71.9(1.6)	80.8(1.4)	62.8(1.1)	76.8(0.5)	78.2(0.8)	64.3(0.2)	78.9(3.0)	73.3
SimSGT	72.3(0.7)	83.6(0.8)	60.6(0.5)	75.7(0.5)	77.7(0.8)	64.1(0.4)	82.0(2.6)	73.7
FG-BERT	70.2(0.9)	84.5(1.5)	64.0(0.7)	**78.4(0.8)**	77.4(1.0)	**66.3(0.8)**	83.2(1.6)	74.8
谟尔格	**73.4(0.7)**	**84.5(0.2)**	**66.1(0.5)**	75.9(0.1)	**79.8(0.6)**	64.8(0.1)	**89.1(2.0)**	**76.2**

- MFBERT [27]：基于变换器的化学指纹处理方法，涉及分布式计算，包括预训练和微调过程。
- BARTSmiles [28]：一种基 于 生成式掩码语言模型的自我监督策略，通过训练 BART 模型来学习有效的分子表征。
- Mole-Bert [29]：一种分子表征学习方法，可将原子编码转换为具有化学意义的离散代码，并将掩蔽原子建模与对比学习相结合。
- SimSGT [30]：一种基于掩码图建模（MGM）的预训练方法，利用简单的 GNN 作为标记器，通过重新构建子图的嵌入来学习分子表征。
- FG-BERT [31]：一种基于 BERT 的自监督方法，通过掩盖具有化学语义的官能团，利用领域知识学习分子表征。
- GraphCL [10] 通过最大化两种不同增强策略下图形的互信息来学习分子表征。
- 3DInfoMax[32]和 GraphMVP[33]引入了三维分子信息，并最大限度地以二维和三维方式表示分子。
- GROVER [15] 结合了 GNN 和 Transformer，通过两个预训练任务（即上下文预测和动机预测）学习分子表征。

- MGSSL [34] 提出了一种基于 Motif 的通用生成预训练框架，它要求 GNN 进行拓扑和标签预测。
- GraphMAE [35] 将 MAE [36] 应用于图形，通过重建节点特征来学习有效信息。
- MolCLR [14] 采用两种相同的增强策略对分子图进行对比学习。

*3) 结果*表 1 列出了 MolIG 在七个基准数据集上的平均 ROC-AUC 值和标准偏差。我们在每个数据集上独立运行 MolIG 3 次。值得注意的是，MolIG 在其中五个数据集上取得了最佳性能，并在所有数据集上达到了最高平均值。我们将 MolIG 的优异表现归功于预训练模型在整合分子图像和图结构信息方面的有效性。这两种模式共同为分子性质预测提供了丰富的背景，从而提高了预测的准确性。具体来说，与只使用单一分子图模式的 SOTA 模型 FG-BERT 相比，FG-BERT 具有显著的性能优势，在五个数据集上超过了 FG-BERT。特别是在 BBBP 数据集上，MolIG 的性能提高了 3.2%。特别是在 ClinTox 数据集上，MolIG 的性能提高了 5.9%。这证明了 MolIG 采用的多模式策略在挖掘分子数据中的潜在知识方面的优越性。

C. 基于 ADMET 的分子特性预测

*1) 数据集：*ADMET [37]是药物发现中的一个重要概念，代表了药物的吸收、分布、代谢、排泄和毒性等特性。与 MoleculeNet 基准组类似，我们在 ADMET 基准组的不同任务中对 MolIG 进行了微调。但不同的是，我们使用 ADMET 提供的分子支架划分法，将每个数据集按照 0.8/0.1/0.1 的比例划分为训练集、验证集和 测试集。

*2) 基准：*在本研究中，我们使用三种基准模型实现了 MolIG：GCN [26]、Attentive FP [4] 和 MolCLR [14]。

- GCN [26] 和 MolCLR [14] 与我们在 MoleculeNet 基准组中使用的模型一致。
- Attentive FP [4] 结合了注意力机制和分子指纹的概念，以捕捉分子中原子和化学键之间复杂的相互作用。它通过迭代更新原子和化学键的特征来生成分子的矢量表示。

*3) 结果*在表 2 中，我们报告了 MolIG 在 21 个 ADMET 基准数据集上的平均 ROC-AUC 值。与 MoleculeNet 基准数据集相比，ADMET 基准数据集更加关注生物实体中药物分子的行为特征，因此包含的分子结构相对更加复杂。实验结果表明，在 21 个数据集中，MolIG 在 17 个数据集上取得了最佳性能。除生物利用率和艾姆斯数据集外，我们的模型都优于只使用图形模式的 MolCLR。特别是在 Pgp 数据集中，我们的性能比最佳基线提高了近 3%。在 VDss 数据集中，性能提高了近 17%。在 CYP2D6 底物中，MolIG 的性能提高了近 9%。这些结果表明，通过将分子图像信息整合到模型中，MolIG在预测生物实体中更复杂的性质时具有更明显的优势。MolIG 在 ADMET 基准数据集上的出色表现证实了多模态预训练框架在分子性质预测领域的 有效性。这项研究为今后进一步探索多模态学习策略以提高药物分子性质预测的准确性提供了有益的启示。

D. 消融研究

这项消融研究旨在评估预训练策略和数据增强策略能否帮助模型在下游任务中取得更好的性能。如图 2 所示，在所有模型架构中，同时采用预训练和数据增强策略的 MolIG（以灰色条表示）表现最佳。没有预训练的模型（用 "w/o pretrain "表示）在所有任务中表现最差。没有数据扩增的预训练策略（以 "w/o img aug "表示）表现次之，但与 "w/o img aug "相比，性能有显著提高。

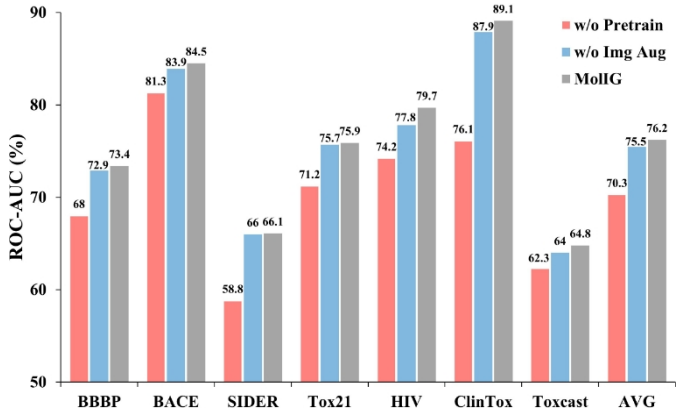

图 2.消融研究：预训练和数据增强策略对 MolIG 在 MoleculeNet 中六个分类数据集上的性能的影响。

模型，而不进行任何预训练。排除这两个部分中的任何一个都很容易导致性能下降。

与完全不进行预训练的模型相比，MolIG 在 ClinTox 数据集上的 ROC-AUC 指标提高了 13.0%，在 SIDER 数据集上提高了 7.8%，在所有六个数据集上平均提高了 5.9%。w/o img aug "策略虽然没有使用数据增强，但仍有明显改善，这表明我们的预训练策略非常有效。MolIG 在所有六个数据集上都有所提高，这表明数据增强策略进一步增强了模型的鲁棒性和泛化能力。

总之，MolIG的多模态预训练策略可以从大量无标注的分子数据中学习高级语义信息，使模型学习到更多的判别表征。同时，数据增强策略在不改变语义信息的情况下，通过扰动图像模态增强了鲁棒性和泛化能力，并有效地应用于分子性质预测的相关生物学任务。

E. 参数实验

MolIG 模型中的一个关键超参数是对比学习中使用的温度缩放，它可以调整模型对正负样本相似度得分的敏感度。如表 3 所示，为了研究不同温度系数对下游任务执行效果的影响，我们在模型的预训练阶段采用了三种不同的温度系数，即 0.05、0.1 和 0.5，并在 MoleculeNet 上的六个分类数据集上进行了实验，报告了这三种温度系数下的平均 ROC-AUC 值。此外，图 3 说明了不同温度系数在这些数据集上的具体表现。

从表 3 和图 3 中可以看出，当温度系数为 0.1 时，模型在下游任务中 的性能达到顶峰。我们认为，较低的温度系数有助于模型更有效地区分负样本。但是，当

表 II
不同模型在 ADMET 基准组上的测试性能。报告了每个基准的平均结果。

类别	数据集	公制	GCN	专心FP	MolCLR	谟尔格
吸收	Caco2	MAE	0.599	0.401	0.434	**0.341**
	HIA	ROC-AUC	0.936	**0.974**	0.956	0.966
	Pgp	ROC-AUC	0.895	0.892	0.861	**0.921**
	生物利用率	ROC-AUC	0.566	0.632	**0.738**	0.725
	亲油性	MAE	0.541	0.572	0.496	**0.461**
	可溶性	MAE	0.907	0.776	0.776	**0.753**
分发	PPBR	MAE	10.194	9.373	9.196	**8.499**
	VDss	斯皮尔曼	0.457	0.241	0.547	**0.717**
代谢	抑制 CYP2D6	PR-AUC	0.616	0.646	0.686	**0.693**
	抑制 CYP3A4	PR-AUC	0.840	**0.851**	0.840	0.841
	CYP2C9 抑制	PR-AUC	0.735	0.749	0.781	**0.802**
	CYP2D6 底物	PR-AUC	0.617	0.547	0.486	**0.702**
	CYP3A4 底物	ROC-AUC	0.590	0.576	0.543	**0.655**
	CYP2C9 底物	PR-AUC	0.344	0.375	0.353	**0.400**
	半衰期	斯皮尔曼	0.239	0.085	0.133	**0.357**
	排泄物清除微粒体	斯皮尔曼	0.532	0.365	0.525	**0.561**
	清除肝细胞	斯皮尔曼	0.366	0.289	0.386	**0.387**
毒性	hERG	ROC-AUC	0.738	0.825	0.821	**0.831**
	Ames	ROC-AUC	0.818	0.814	**0.887**	0.883
	迪利	ROC-AUC	0.856	0.886	0.871	**0.921**
	半数致死剂量	MAE	0.649	0.678	0.435	**0.428**

表 III
温度 (τ) 对摩尔纹损失的影响：六种分类的平均性能报告数据集。

温度 (τ)	0.05	0.1	0.5
ROC-AUC (%)	74.2	76.2	73.4

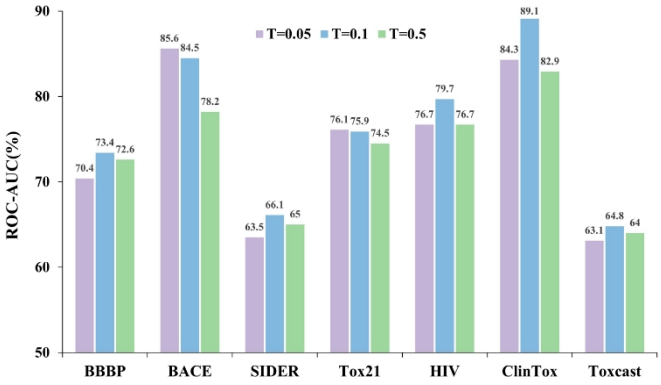

图 3.不同温度系数下 MolIG 在 MoleculeNet 六个分类数据集上的表现。

当温度系数为 0.05 时，该模型在下游任务中的平均性能比温度系数为 0.1 的模型低 2%。这可能是由于分子图像的稀疏性导致模型从图像模态向 GNN 编码器传输冗余信息。另一方面，温度系数越高，模型的训练过程就越简单，从而无法学习有意义的表征。我们观察到，当温度系数设置为 0.5 时，模型很难收敛，导致训练不稳定。此外，它还降低了不同分子之间的区分能力，导致预测结果含糊不清。这可能会严重影响其在下游任务中的表现。

F. 通过 MolIG 检索分子

为了进一步研究 MolIG 方法的表现能力，我们评估了该方法的实地性能的分子相似性搜索。在实验设计中，我们首先从 PubChem 数据库中随机选取 100,000 个分子作为查询集。对于每个查询分子，我们使用 MolIG 获取其分子表示，然后计算其与数据库中其余分子表示的余弦相似度。

根据计算出的相似性得分，实验结果展示了几个分子实例，其相似性分别为

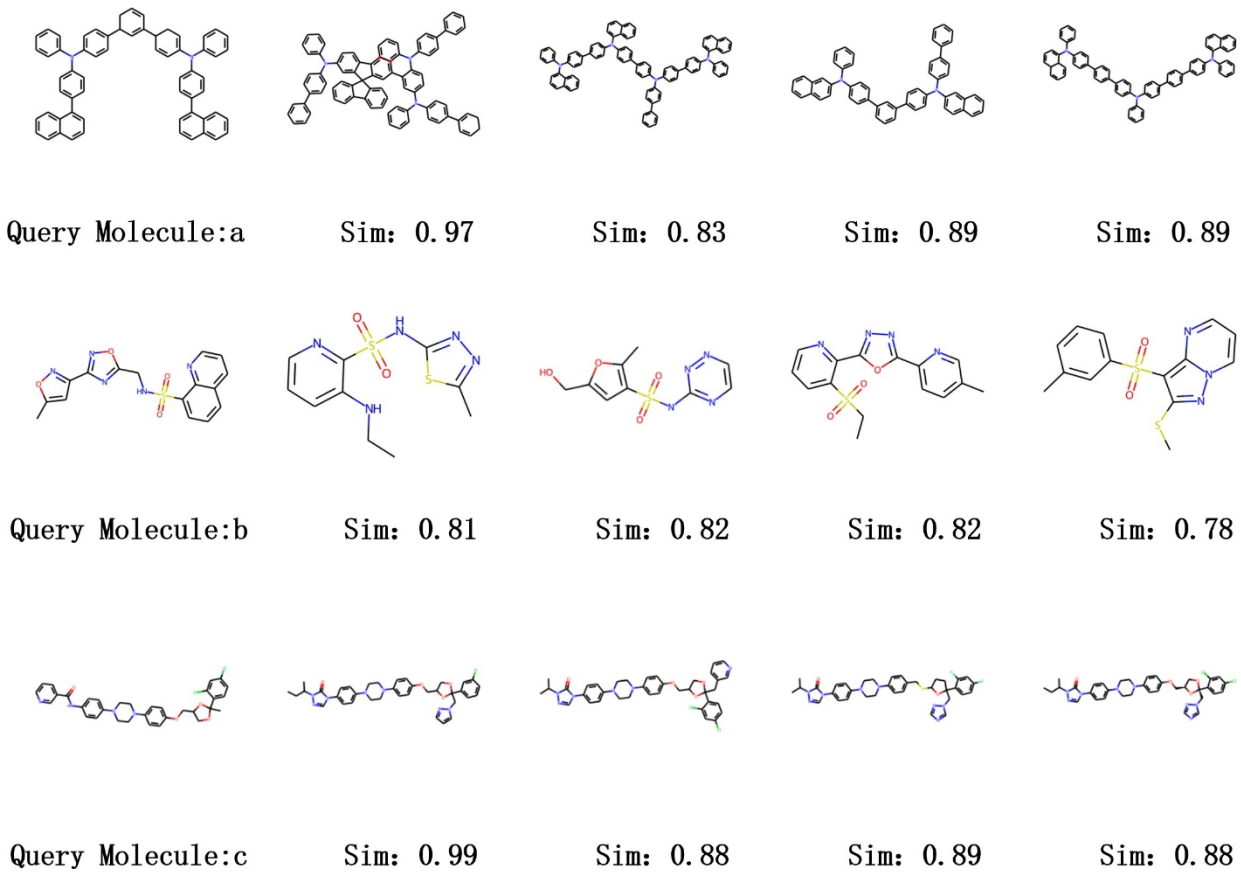

图 4.查询结果

与查询分子相似度最高的分子，见图 4.例如，查询分子 a 的结构特征是由多个苯环组成，MolIG 成功检索到了具有相似苯环排列和一个中心氮原子的分子。这一结果验证了 MolIG 在捕捉原子级结构信息方面的有效性。对于查询的分子 b，MolIG 准确地识别出了其关键官能团 O=S=O，证明了该模型识别分子官能团的能力。在分子 c 的查询实验中，MolIG 检索到的分子通常含有氟或氯等卤素元素，并表现出有毒特性，这表明 MolIG 能够从语义角度理解分子特性。总之，这些实验结果表明，MolIG 不仅能精确捕捉分子结构细节，还能理解更高阶的分子语义信息，如官能团的特征和分子的潜在属性。

V. 实施细节

我们的预训练模型由 5 层、300 个隐藏维度的图形同构网络（GIN）和一个由 5 层、300 个隐藏维度的图形同构网络（GIN）组成。

的残差卷积神经网络（ResNet-34）。我们在 1 个 RTX 3090 GPU 上使用 512 个批次对模型进行了 50 个历元的预训练。我们使用 Adam 优化器，ResNet 学习率为 0.01，GIN 学习率为 0.0005。对比损失的温度系数为
0.01.权重衰减为 1e-5。我们拍摄分辨率为 224×224 的图像。

VI. 结论

分子性质预测在药物发现中起着至关重要的作用。虽然以往的分子性质预测模型已经取得了相当大的成功，但使用单一信息模式往往会限制其预测性能。在本研究中，我们提出了一种新颖的多模态分子预训练框架 MolIG，该框架通过在预训练阶段最大化图形和图像模态特征之间的一致性来学习分子表征。在药物发现任务（如 MoleculeNet 和 ADMET）上进行的经验评估表明，MolIG 优于当前最先进的基线模型。值得注意的是，MolIG 不仅能提取分子的结构特征，还能捕捉更高阶的语义信息，这些信息可以转移到生物相关任务中

在分子特性预测中。此外，我们的模型目前只考虑了两种模式的信息。将 SMILES 和三维信息纳入该框架将是我们未来研究的主题。

致谢

这项工作得到了国 家 重点研发计划（2019YFC1711000）和新型软件技术与产业化协同创新中心的支持。

参考资料

[1] Garrett B Goh、Nathan O Hodas 和 Abhinav Vishnu。计算化学的深度学习。*计算化学杂志*》，38（16）：1291-1307，2017.

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

[3] 陈宏明、Ola Engkvist、王银海、Marcus Olivecrona 和 Thomas Blaschke。深度学习在药物发现领域的崛起。*药物 discovery today*, 23(6):1241-1250, 2018.

[4] 熊兆平、王定彦、刘晓红、钟飞生、万晓哲、李旭彤、李兆军、罗晓敏、陈开贤、蒋华良等. 以图注意机制推动药物发现的 分子表征边界. *药物化学杂志*，63（16）：8749-8760，2019.

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

[6] Maylis Orio、Dimitrios A Pantazis 和 Frank Neese。密度泛函 理论。*光合作用研究*》，102:443-453，2009 年。

[7] 吴宗汉、潘世瑞、陈凤文、龙国栋、张成琪、俞皓。图神 经 网 络综合研究。*IEEE 神经网络与学习系统事务*，32(1):4-24, 2020.

[8] Zhichun Guo, Bozhao Nan, Yijun Tian, Olaf Wiest, Chuxu Zhang, and Nitesh V Chawla. 基 于 图 的 分 子 表 征 学 习 。 *arXiv preprint arXiv:2207.04869*, 2022.

[9] Justin Gilmer、Samuel S Schoenholz、Patrick F Riley、Oriol Vinyals 和 George E Dahl。量子化学的神经信息传递。*机器学习国际会议*，第 1263-1272 页。PMLR，2017。

[10] 尤宁、陈天龙、隋永铎、陈婷、王占洋和沈洋。具有增强功能的图形对比学习。*神经信息处理系统进展*》，33:5812-5823， 2020。

[11] Tingting Shi, Yingwu Yang, Shuheng Huang, Linxin Chen, Zuyin Kuang, Yu Heng, and Hu Mei.基于分子图像的卷积神经网络用于钦佩特性预测。*化学计量学与智能实验室系统*》，194:103853，2019.

[12] Shifa Zhong, Jiajie Hu, Xiong Yu, and Huichun Zhang.分子图像-卷积神经网络（cnn）辅助 qsar 模型预测污染物对哦自由基的反应性：迁移学习、数据增强和模型解释。*化学工程 期刊*，408:127998，2021。

[13] Garrett B Goh、Charles Siegel、Abhinav Vishnu、Nathan O Hodas 和 Nathan Baker。 Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed qsar/qspr models. *arXiv preprint arXiv:1706.06689*, 2017.

[14] 王宇阳、王建仁、曹中林和阿米尔-巴拉蒂-法里马尼。通过图神经网络进行表征的分子对比学习- works.*自然机器智能*》，4（3）：279-287，2022。

[15] 于荣、边亚涛、徐廷阳、谢伟阳、魏颖、黄 文兵、黄俊洲。大规模分子数据的自监督图变换器。*神经信息处理进展 系统*，33:12559-12571，2020.

[16] Yin Fang, Qiang Zhang, Ningyu Zhang, Zhuo Chen, Xiang Zhuang, Xin Shao, Xiaohui Fan, and Huajun Chen.知识图谱增强的分子对比学习与功能提示。*自然-机器智能*》，第1-12页，2023年。

[17] Shuangli Li, Jingbo Zhou, Tong Xu, Dejing Dou, and Hui Xiong.Geomgcl：用于分子性质预测的几何图对比学习。In *Proceedings of the AAAI Conference on Artificial Intelligence, volume* 36, pages 4541-4549, 2022.

[18] 刘胜超、聂伟力、王成鹏、陆家瑞、乔卓然、刘玲、唐健、肖超伟、阿尼玛-阿南德库马尔。基于*文本检索*和编辑的多模态分子结构-文本模型。

[19] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka.图神经网络有多强大？*学习表征国际会议*，2019。

[20] 胡伟华、刘博文、约瑟夫-戈麦斯、马林卡-齐特尼克、梁珀西、维杰-潘德和尤雷-莱斯科维奇。预训练图神经网络的策略。*国际学习表征会议*，2020。

[21] 何开明、张翔宇、任少清和孙健。图像识别的深度残差学习。In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770-778, 2016.

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton.利用深度卷积神经网络进行图像分类。*ACM 通信*》，60（6）：84-90，2017 年。

[23] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton.视觉表征对比学习的简单框架。*机器学习国际会议*，第 1597-1607 页。PMLR，2020。

[24] 格雷格-兰德伦Rdkit：开源化学信息学。2006.*Google Scholar*, 2006.

[25] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2019 update: Improved access to chemical data.*核酸 research*, 47(D1):D1102-D1109, 2019.

[26] Thomas N. Kipf 和 Max Welling.用图卷积网络进行半监督分类。*国际学习表征会议（ICLR）*，2017 年。

[27] Hisham Abdel-Aty 和 Ian R Gould.用于化学指纹识别的 大规模分布式变压器训练。*化学信息与建模期刊*》，62（20）：4852-4862，2022 年。

[28] Gayane Chilingaryan、Hovhannes Tamoyan、Ani Tevosyan、Nelly Babayan、Lusine Khondkaryan、Karen Hambardzumyan、Zaven Navoyan、Hrant Khachatrian 和 Armen Aghajanyan。Bartsmiles：*ArXiv preprint arXiv:2211.16349*, 2022.

[29] Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z Li.Mole-bert：反思分子的预训练图神经网络。*第十一届学习表征国际会议*，2022 年。

[30] 刘志远、史耀瑞、张安、张恩智、川口健二、王翔、蔡达生。重新思考分子屏蔽图建模中的标记器和解码器。*神经信息处理系统进展*》，36，2024。

[31] Biaoshun Li、Mujie Lin、Tiegen Chen 和 Ling Wang。 Fg-bert：基于官能团的 分子表征学习框架，用于性质预测。*生物信息学简报*，24（6）：bbad398，2023.

[32] Hannes Sta¨rk、Dominique Beaini、Gabriele Corso、Prudencio Tossou、Christian Dallago、Stephan Gu¨nnemann、Pietro Lio`。3d infomax 改进了用于分子特性预测的 gnns。*国际机器学习会议*，第 20479-20502 页。PMLR, 2022.

[33] 刘胜超、王汉臣、刘伟阳、Joan Lasenby、郭宏宇和唐健。用三维几何预训练分子图表示。*国际学习表征会议*，2022。

[34] Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee.分子性质预判定的基于动机的图自监督学习。*神经信息处理系统进展*》，34:15870- 15882, 2021.

[35] Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang.Graphmae：自监督屏蔽图自编码器。*第 28 届 ACM SIGKDD 知识发现与数据挖掘大会论文集*》，第 594-604 页，2022 年。

[36] 何开明、陈新磊、谢赛宁、李阳浩、Piotr Dolla´r 和罗斯-吉尔希克。遮蔽式自动编码器是可扩展的视觉学习器。在

IEEE/CVF 计算机视觉和 模式识别会议论文集》，第 16000-16009 页，2022 年。

[37] Kexin Huang、Tianfan Fu、Wenhao Gao、Yue Zhao、Yusuf Roohani、Jure Leskovec、Connor W Coley、Cao Xiao、Jimeng Sun 和 Marinka Zitnik。治疗数据共享：*ArXiv preprint arXiv:2102.09548*, 2021.