


FG-BERT: 基于官能团的通用和自监督分子表征学习框架, 用于性质预测

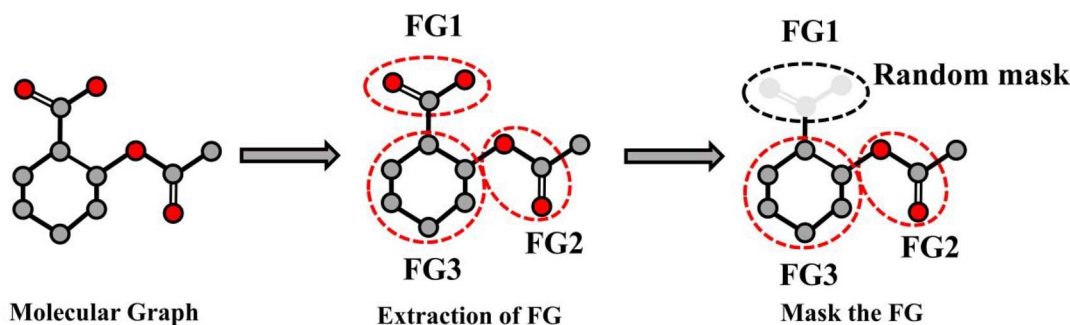
李彪顺、林木杰、陈铁根和王玲 

通讯作者: 王玲王玲, 发酵与酶工程广东省重点实验室, 教育部合成生物学与医学国际联合研究实验室, 广东省生物医药工程技术研究中心, 华南理工大学生物与生物工程学院, 广州 510006。电话: 020-39380602; 电子邮件: lingwang@scut.edu.cn

摘要

基于人工智能的分子特性预测在生物活性分子和功能材料等分子设计中发挥着关键作用。在这项研究中, 我们基于 145 万个未标记的类药物分子, 提出了一种自监督预训练深度学习 (DL) 框架, 称为来自变压器的功能组双向编码器表征 (FG-BERT), 以从功能组中学习有意义的分子表征。预训练的 FG-BERT 框架可进行微调, 以预测分子特性。与最先进的 (SOTA) 机器学习和 DL 方法相比, 我们在 44 个基准数据集上证明了 FG-BERT 在涉及物理化学、生物物理学和生理学的任务中评估分子特性的高性能。此外, FG-BERT 还利用注意力机制关注对目标特性至关重要的 FG 特征, 从而为下游训练任务提供了出色的可解释性。总之, FG-BERT 不需要任何人工制作的特征作为输入, 而且具有出色的可解释性, 为各种分子 (尤其是药物) 发现任务提供了一个开发 SOTA 模型的开箱即用的框架。

图表摘要



关键词: 分子性质预测; FG-BERT; 分子表征; 深度学习; 自监督学习

引言

准确预测分子性质对于功能分子的设计和发现具有重要意义, 尤其是对于药物分子的发现, 因为它可以在药物发现过程的早期阶段用于快速识别具有理想性质的活性分子和/或过滤掉不合适的分子[1]。通常, 分子表征是分子性质预测的基础; 因此, 如何获得有效的分子表征?

表征是分子特性预测领域亟待解决的一个重要问题。目前的分子表征可分为五类: 分子描述符、指纹、图形、分子串和分子图像。基于这些预定义的分子表征, 从业人员可以利用机器学习 (ML) 和深度学习 (DL) 建立定量结构-活性/性质关系 (QSAR/QSPR) 模型, 用于预测分子性质。

李彪顺是华南理工大学的一名研究生。他目前的研究兴趣包括机器学习、预训练和人工智能辅助药物发现 (AIDD)。

林慕洁是华南理工大学的一名本科生。她的研究兴趣包括机器学习和生物信息学。

陈铁根是中国科学院上海药物研究所中山药物研究所药物化学专业的主要研究人员。他目前的研究重点是有机合成、新型抗病毒药物的发现和开发。

王玲，华南理工大学副教授。他于 2014 年获得中山大学药学院博士学位。他的研究重点是计算机辅助药物设计（CADD）、人工智能辅助药物发现（AIDD）和药物化学。

收到：2023 年 6 月 30 日。**已修订：****修订：**2023 年 9 月 25 日。**接受：****2023 年 10 月 14 日接受：**2023 年 10 月 14 日

© 作者 2023。牛津大学出版社出版。保留所有权利。如需授权，请发送电子邮件至：journals.permissions@oup.com

一般来说, 基于传统 ML 的 QSAR/QSPR 模型的准确性在很大程度上取决于如何选择合适的分子代表 [2]。也就是说, 传统的 ML 方法要求化学家手动制定一套规则, 将分子的相关结构信息、药效特征和/或理化性质编码成固定长度的向量 [3], 如常用的分子指纹和描述符。然而, 由于设计和选择过程耗时且容易出错, 分子描述符的可扩展性和通用性较差, 这反过来又导致基于描述符的 ML 模型在分子性质预测领域存在一些缺陷。基于 DL 的分子性质预测模型不同于传统的 ML 模型, 因为它们不需要从大量预定义的可计算分子描述符中人工选择与任务属性相关的最重要描述符 [4]。目前, 基于 DL 的模型通常可以使用分子图 [5]、SMILES 序列 [6, 7] 和分子图像 [8] 作为输入特征来构建。其中, 由于分子是天然的图结构, 图神经网络 (GNN) 模型已成为分子性质预测领域的研究热点。近年来, 许多基于图的有监督 DL 模型 [9-14] 被开发出来, 并在分子性质任务中表现出了相当可观的性能。然而, 这种有监督的 GNN 模型的准确性取决于数据量的大小, 在小数据集上甚至不如基于描述符/指纹的传统 ML 模型 [15]。此外, GNN 容易出现过平滑问题, 而且模型的层数通常在 2-4 层之间, 这也限制了模型提取分子特征的能力 [6]。

最近, 有人提出了预训练模型来解决上述问题。它们可以通过设置特定的预训练策略 (如对比学习 [16] 和掩蔽语言学习 [17]), 从大量无标记数据中学习有用的分子表征, 然后将知识迁移到分子性质预测的下游任务中。与传统的有监督 ML 和 DL 模型相比, 一些预训练模型已经开发出来, 并在分子性质预测领域取得了良好的性能 [6, 18-21]。具体来说, 他们通过构建自己的预训练任务对模型进行预训练, 然后针对分子性质预测对模型进行微调。例如, K-BERT 通过利用基于原子特征预测、分子特征预测和比较学习的三个预训练任务, 可以像化学家一样从 SMILES 中提取化学信息 [6]。Mole-BERT 采用 VQ-VAE 变体编码器作为上下文感知消歧器, 将原子编码为有意义的离散值, 从而扩大了原子词汇量, 减轻了原子与稀有原子之间的显著数量差异 [21]。它通过随机屏蔽原子离散值并预训练 GNN 来预测这些离散值。对于图级预训练, Mole-BERT [21] 提出了三元掩码比较学习 (Ternary Mask Comparison Learning) 来模拟分子间的异质语义相似性, 这对分子检索特别有效, 并能以完全数据驱动的方式匹配或超越最先进的 (SOTA) 方法。然而, 现有的预训练模型并不关注分子结构中重要的官能团 (FG)

信息。众所周知, 分子结构决定了分子的各种性质, 而分子中的官能团作为分子的重要组成部分, 其结构往往与分子的性质密切相关 [22-25]。

在这项研究中, 我们开发了一种新的遮蔽 Chemical 语言预训练框架 (命名为功能组双向

FG-BERT (图 1) 用于从大规模无标记分子语料库中学习化学语义和结构信息。FG-BERT 有两个重要改进: (1) 它掩盖了分子中的 FG, 以高精度执行大规模预训练恢复预测; (2) 它利用自监督预训练学习框架, 从 145 万个具有不同生物活性的类药物分子中学习有用的分子表征。与 SOTA ML 和 DL 方法相比, 大量实验结果表明, FG-BERT 在多个基准数据集上的各种分子特性预测任务中都具有很高的准确性。此外, FG-BERT 还能通过注意力机制 (AMs) 自动学习关注与目标特性相关的 FGs, 为进一步的分子分析、设计和优化提供有价值的线索。

材料和方法

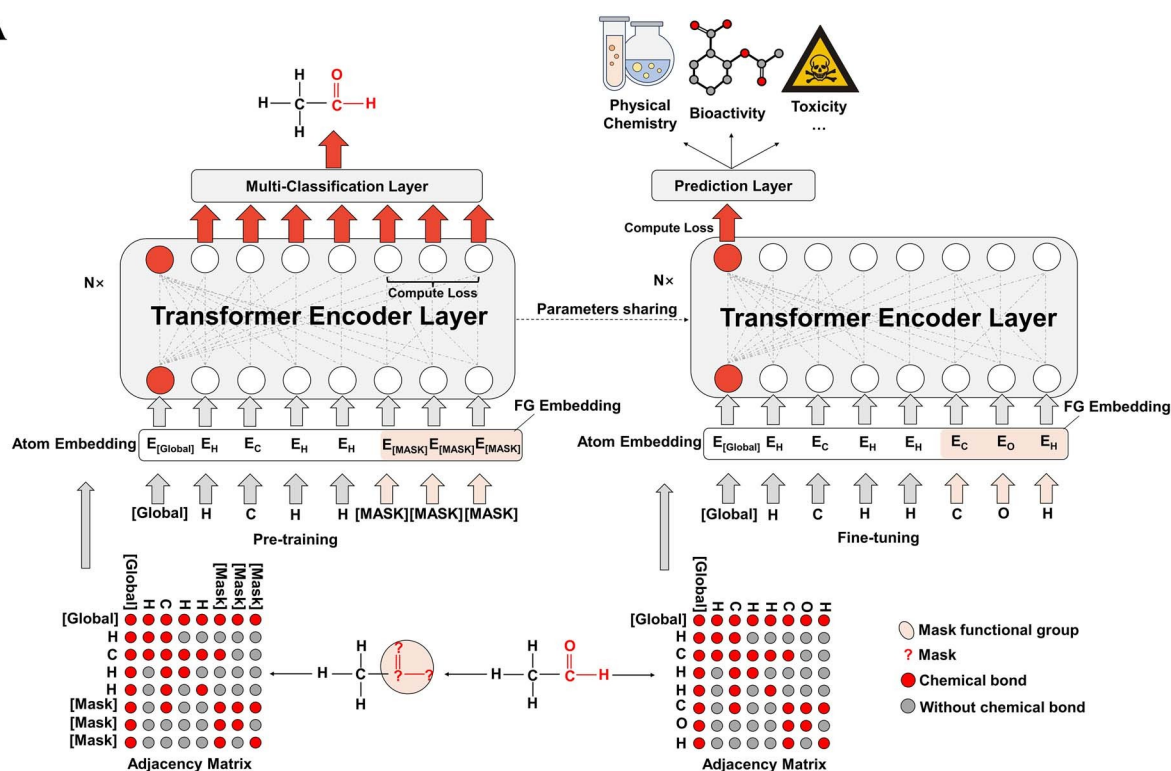
收集分子语料库和基准数据集

从 ChEMBL (第 30 版) [26] 收集初始无标记分子数据集 (~213 万个分子), 然后通过类药物属性的三个原则进行筛选 (图 1C): 分子量 ≤ 500 、ClogP ≤ 5 、氢键供体数 ≤ 5 。最终, 获得了 145 万个分子的分子语料库, 并按 9:1 的比例随机分成训练集和测试集, 在预训练过程中进行大规模掩蔽恢复预测。在微调过程中, 我们使用三个基准数据集广泛评估了 FG-BERT 预训练框架的性能。首先, 我们收集了 15 个与药物发现相关的常用公共数据集 (表 S1), 用于评估 FG-BERT 的性能, 其中包括四个物理化学数据集 (ESOL、FreeSolv、Lipo 和 CEP) [27-30]、两个生物活性和生物物理数据集 (HIV、疟疾、MUV 和 BACE) [31-34]、两个量子化学数据集 (QM7 和 QM8) [35] 和四个生理学和毒性数据集 (BBBP、Tox21、SIDER、ToxCast 和 ClinTox) [35-39]。其次, 我们还使用 15 个 ADMET (吸收、分布、代谢、排泄和毒性) 数据集测试了 FG-BERT 的性能 [6] (表 S2)。最后, 我们使用 14 个乳腺细胞系表型筛选数据集 [40] (表 S3) 来评估 FG-BERT 的预测能力。

FG-BERT 框架

FG-BERT 是在 BERT 模型 [41] 的基础上设计的, 该模型有两个预训练任务, 即屏蔽语言模型 (MLM) 和下一句预测任务。在 NLP 中, 句子是连续的, 而分子图与文本的不同之处在于, 分子中的 FG 和原子是通过相互连接的化学键而非连续顺序联系在一起的; 因此, FG-BERT 不需要关于 FG 和原子二进制值的额外信息。在自然语言处理中, 每个单词都可能与其他单词相关, 因此有必要关注所有单词。然而, 在分子图中, 原子和 FG 主要与通过化学键连接的相邻原子或 FG 相关。因此, 与 BERT 不同, 我们主要关注化学键连接的定位, 因此原子和 FG 之间的信息交互仅通过化学键。值得注意的是, FG-BERT 可以克服 BERT 中由于重连接机制而产生的过平滑问题, 并有足够的能力提取分子图中的深层模式。如图 1A 所示, 我们利用分子的邻接矩阵来控制分子间的信息交换。我们添加了一个 GLOBAL 节点, 用于与所有原子连接和 FG 连接交换信息。

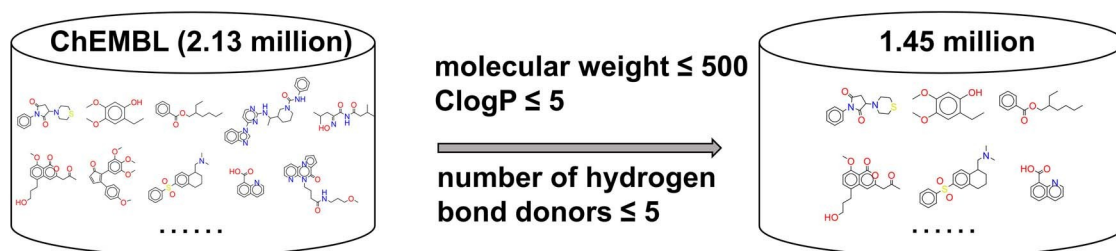
A



B

Name	Layers	Heads	Embedding size	FFN size	Learning rate	Dropout	Model Params
FG-BERT	6	4	256	512	10^{-4}	0.1	~3.2M

C



D

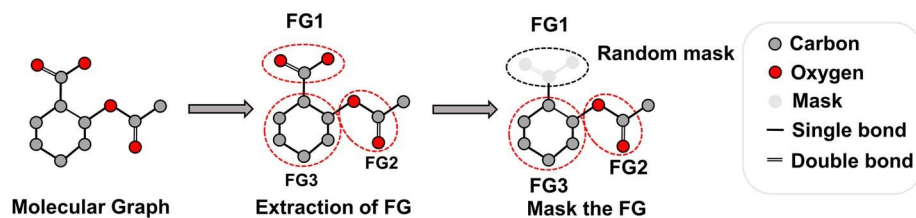


图 1: FG-BERT 示意图。FG-BERT 示意图。(A) FG-BERT 框架及相应的预训练和微调过程。(B) FG-BERT 预训练模型的超参数。(C) 分子语料筛选过程。(D) FGs 屏蔽过程。FG: 官能团。

其输出可视为解决下游分类或回归任务的最终分子表征[18]。

由于 GLOBAL 节点与所有节点相连，因此也在一定程度上考虑了长距离依赖问题。与 BERT 不同，我们的预训练任务直接屏蔽语言建模，即屏蔽 FGs 预测，然后学习分子中的化学信息并提取分子表征，以微调下游任务。

FG-BERT 框架由三部分组成：嵌入层、转换层和预训练/预测头（图 1A）。对于输入模型的每个分子，我们都会添加一个连接所有 FG 和原子的超级节点，并根据化学键关系将 SMILES 格式的分子转换为分子图。

在嵌入层面，原子列表 $w = (a_1, a_2, a_3, \dots, a_N)$ 通过嵌入矩阵 $D \in R^{V \times d_{\text{model}}}$ 嵌入到分布空间 $x = (x_1, x_2, x_3, \dots, x_N)$, $x_i \in R^{d_{\text{model}}}$ ，其中 V 是分布空间。

是词汇量的大小， d_{model} 是图 1B 中定义的嵌入大小。分子通过嵌入层后，会得到一个嵌入向量，然后将其传输到转换层。在转换层中，每个节点使用 AMs 聚合来自相邻节点的信息，单个节点的信息传递过程如下所述

$$q_i = W x_{qi} \quad (1)$$

$$k = W x_k \quad (2)$$

$$i \quad i$$

$$v_i = W x_{vi} \quad (3)$$

$$s_{i,j} = \frac{\text{dot } q_i, k_j}{\sqrt{d_{\text{model}}}}, j \in N_i \quad (4)$$

$$a_{i,j} = \text{softmax } s_{i,j} = \frac{e^{s_{i,j}}}{\sum_{j \in N_i} e^{s_{i,j}}} \quad (5)$$

$$M_i = \sum_{j \in N_i} a_{i,j} v_{i,j} \quad (6)$$

x_i 是输入节点 i 的表示， W , W_{qk} , W_v 是所有节点共享的可学习矩阵， W , W_{qk} , $W_v \in R^{d_k \times d_{\text{model}}}$, $d_k = d_{\text{model}}/H$, H 是模型的头数，如图 1B 所定义， N_i 表示节点 i 的所有邻居。

我们采用了多头 AM，上述过程分别独立执行 H 次，然后将结果拼接在一起，并根据以下公式进行线性变换

$$M_i = W_0 \text{concat } m_i^1, m_i^2, \dots, m_i^K \quad (7)$$

$W_0 \in R^{d_{\text{model}} \times d_{\text{model}}}$ 也是所有节点共享的可学习矩阵。

为了解决普通 GNN 中的过平滑问题，我们使用与转换器相同的残差连接和层归一化机制，如下式所示：

$$h_i = \text{layernorm}(x_i + M_i) \quad (8)$$

为了增强模型的表现力，我们将多头注意力子层的输出 h_i 传递给前馈子层，并根据以下原则使用前馈网络（FFN）

$$o_i = \text{layernorm FFN } h_i^{\phi} + h_i^{\phi} \quad (10)$$

其中， $W_1 \in R^{d_{\text{hidden}} \times d_{\text{model}}}$, $W_2 \in R^{d_{\text{model}} \times d_{\text{hidden}}}$, d_{hidden} 等于图 1B 中定义的 FFN 大小，gelu 表示激活函数，称为高斯误差线性单元。

根据图 1B 中的参数层，FG-BERT 中的转换器层被多次执行。预训练头与预测头不同，但它们都由两层 FFN 组成，分类任务使用交叉熵作为损失函数，回归任务使用均方根误差（RMSE）作为损失函数。预训练模型使用的激活函数是 GELU，下游分类和回归任务使用的激活函数是 LeakyReLU。最终，FG-BERT 模型的参数总数约为 320 万个。

FG-BERT 的输入分子表示法

为了方便地在分子图中表示原子和 FG，我们根据各种原子在预训练分子语料库中出现的频率构建了原子字典。对 FG-BERT 使用的预训练数据集进行统计后发现，有 14 种原子出现的频率超过 1000 次，用相应的元素符号表示，而其他原子出现的频率则低于 1000 次，我们将其统称为 [UNK]。

此外，为了便于后续下游

任务，我们在分子图中添加了一个超级节点，表示为

用[GLOBAL]表示。我们用 [MASK] 表示屏蔽的 FG（图 1D）。

由于一个 FG 通常由多个原子组成，当一个 FG 被屏蔽时，通常会有多个原子被[MASK]。因此，我们构建一个随机选择的 FG 列表，并使用

该列表用于识别需要屏蔽的 FG，而未被选中的 FG 则保持不变。因此，词典包括

下列代币：[H]、[C]、[N]、[O]、[S]、[F]、[Cl]、[Br]、[P]、[I]、[Na]、

[B]、[Se]、[Si]、[UNK]、[MASK] 和 [GLOBAL]。本研究使用

RDKit 软件 (<http://www.rdkit.org/>) 生成 FGs。

子层连接和层归一化机制

$$\text{FFN } h_i^{\phi} = W_2 \text{gelu } W_1 h_i^{\phi} + b_1^{\phi} + b_2^{\phi} \quad (9)$$

本研究提出的预训练策略与 BERT 非常相似。首先，我们根据预定义的 FGs 列表（图 2）遍历所有分子，从而得到与分子语料库中每个分子相关的 FGs 列表。与 BERT 一样，我们会随机选择分子中 15% 的 FGs 进行屏蔽。与 BERT 不同的是，我们放弃了 FG 替换的操作，因为分子中的 FG 替换可能会导致许多不一致的化学规则发生，因此我们有 90% 的概率会在所选的 15%FG 中被掩蔽。在预训练过程中，我们只计算被掩蔽 FG 的损失，其他 10% 的概率保持不变。

训练方案、超参数优化和评估

预培训阶段

在预训练阶段，首先使用 RDKit 软件将 SMILES 格式的每个分子转换为二维（2D）无向图，然后在每个分子图中添加一个额外的超级节点。根据 FG-BERT 预训练策略，随机屏蔽 FG，最后将分子图传入 FG-BERT 模型以预测屏蔽的 FG。对于只有几个 FG 的分子，我们至少屏蔽一个 FG。同时，虽然分子语料中不包含 FG 的分子很少，但我们也不会去碰它们。我们的模型采用批量梯度下降算法进行训练

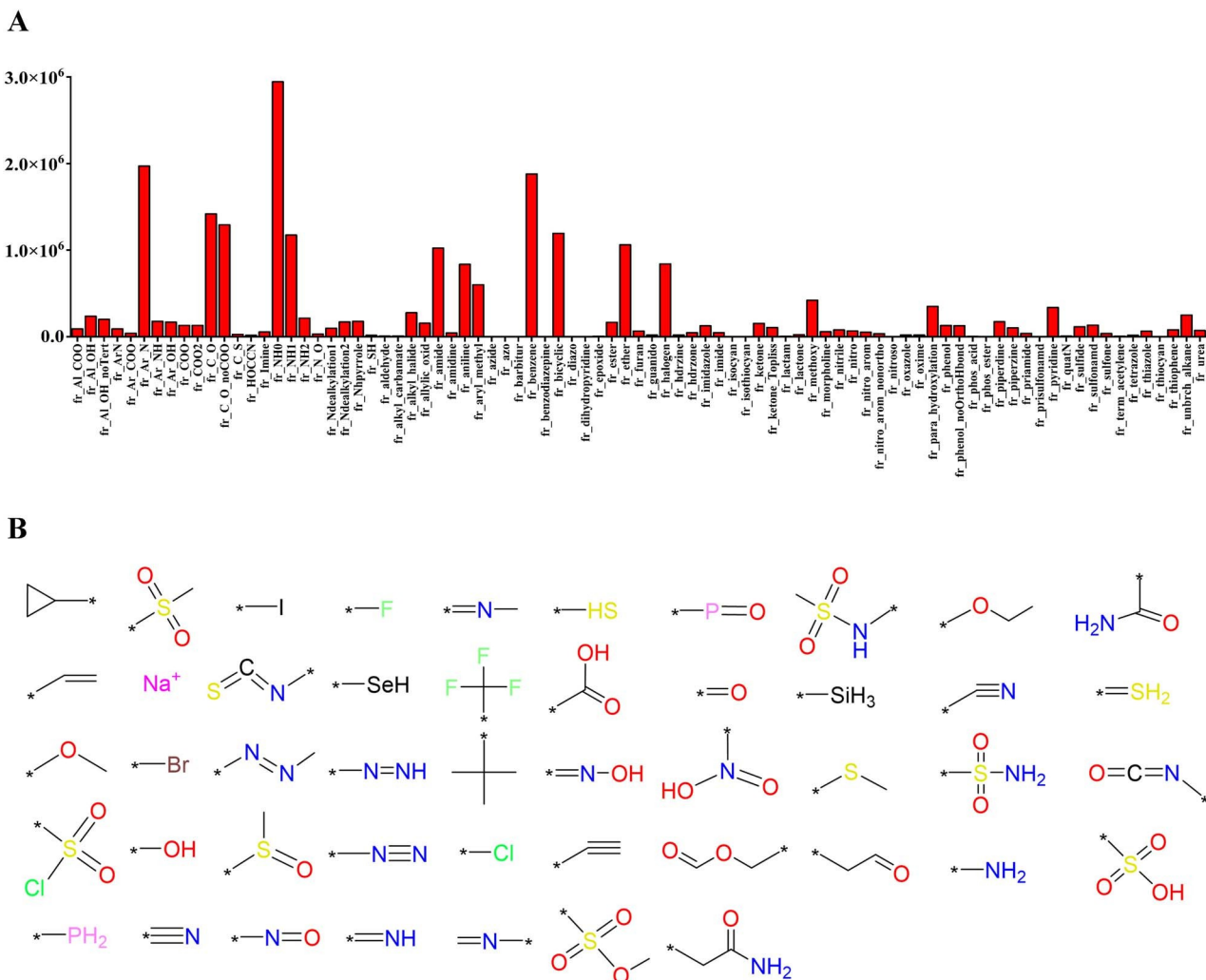


图 2. (A) FGs 数量统计。(B) FGs 列表。有关预训练语料库和 FGs 列表中 FGs 数量的详细统计信息, 请访问 <https://github.com/idrugLab/FG-BERT/tree/main/FGs>。FGs, 功能词组。

和 Adam 优化器 [42]。预训练模型的学习率设定为 10^{-4} , 批量大小设定为 16。为了评估 FG-BERT 预训练的性能, 采用了预训练屏蔽策略来屏蔽测试集中的分子, 然后计算恢复率作为评估指标。交叉熵损失函数用于计算 FG 的损失。FG 通常由多个原子组成, 因此只能完全恢复 FG 的正确性, 即如果所有原子都被正确预测, 则损失最小且为零。如果被屏蔽的 FG 只能部分正确恢复, 则会产生损失。模型通过多轮训练和参数更新使损失最小化, 从而使模型更准确地完成预测任务。因此, 我们使用多个原子的交叉熵损失之和作为 FG 屏蔽恢复预训练的损失度量。

微调阶段

在微调阶段, 我们将预训练模型从其预训练头中移除, 并在变压器编码器上添加一个与超级节点相对应的两层全连接神经网络, 称为预测头。为了避免过拟合, 采用了 dropout 策略 [43], 由于 dropout 对具体下游任务的影响大, 因此根据不同的下游任务选择不同的 dropout 值。根据之前 MG-BERT 模型的结果 [18], 我们

同时, 在 $[0, 0.5]$ 的范围内设置辍学值, 以做出最优选择。此外, 亚当优化器被用作每个任务的微调优化器, 其超参数扫描限制如下: 批量大小: {8, 16, 32, 48, 64}, 学习率: $[\ln(3e-5), \ln(15e-5)]$, 头数: 为 {4, 8, 64}: $[\ln(3e-5), \ln(15e-5)]$, 头部数量: {4, 8}。各下游任务的超参数详见表 S4。为了与现有方法进行公平比较, 我们选择了相同的数据、数据分割方法和配比以及评估指标。例如, 所有三个基准数据集都以 8:1:1 的比例分为训练集、验证集和测试集。使用 Hyperopt-Python 软件包 [44] 对 hyperparameters 进行贝叶斯优化, 其中包括多头关注数、辍学率、批量大小和学习率。验证集的结果用于确定最终模型的超参数, 以提高泛化能力, 而 FG-BERT 模型的最终性能则由测试集的结果来表示。此外, FG-BERT 采用早期停止来避免训练过程中的过拟合, 容差值设为 30, 最大历时设为 200。

下游微分类和回归任务的最终评价指标分别是 RMSE 和接收者操作特征曲线下面积 (ROC-AUC)。此外, 为了减少随机误差并确保结果的可靠性, 我们基于以下指标对 FG-BERT 模型进行了评估

在每个数据集的 10 个不同随机种子上，计算评估指标的平均值和标准偏差，以表示最终结果。FG-BERT DL 框架由 Tensorflow 软件开发，所有的 FG-BERT 预训练和微调都是在 SCUTGrid (SCUT 超级计算平台) 的 GPU [NVIDIA Corporation GV100GL (Tesla V100 PCIe 32 GB)] 和 CPU [Intel(R) Xeon(R) Silver 4216 CPU@2.10 GHz] 上进行训练的。

结果与讨论

FG-BERT 模型中 FG 的统计和屏蔽

众所周知，分子的性质与其结构密切相关，而 FGs 是分子中重要的亚结构。各种 FG 的存在是分子化合物的一个显著特征[45]。我们统计了用于预训练的分子语料库数据集 (1 456 893 个分子) 中的 FGs 数量。如图 2A 所示，大量 FGs 广泛存在于类药物小分子中，这表明 FGs 对小分子药物的性质起着至关重要的作用。

目前，我们的预定义 FG 列表包括 47 种常见 FG (图 2B)。该列表之所以小于图 2A 所示的 85 个 FG，是因为在计算 FG 数量时，RDKit 软件将连接到苯环和脂肪族链上的 FG 视为不同的 FG。例如，苯环上的羧基和甲基被视为两个不同的 FG。但是，在 FG-BERT 的预训练过程中，我们将它们视为一个唯一的 FG，因此我们预定义的 FG 列表只包含 47 个 FG，而不是 85 个 FG。此外，考虑到环状结构在小分子药物中的广泛存在，这些环状结构可能是影响药物分子性质的重要子结构/碎片。因此，本研究将环状结构也视为特殊的 FGs，并对其与其他 FGs 相同的屏蔽和恢复预训练操作 (图 1D)。

与 BERT 和 MG-BERT [18, 41] 类似，我们随机选取分子中 15% 的 FGs 进行掩蔽，具体的掩蔽过程如图 1D 所示。在此基础上，我们的 FG-BERT 专注于分子中的 FGs 信息，通过掩蔽恢复 FGs 进行预训练，学习分子中有用的语义和结构信息，然后提取分子表征，最终完成下游任务的预测。经过 20 个历时的预训练后，FG-BERT 模型恢复 FG 的准确率达到 98.70%，证明了屏蔽 FG 构建化学语言预训练模型的有效性。预训练完成后，获得的权重可用于下游的分类和回归任务。

FG-BERT 在公共基准数据集上的性能

我们利用 15 个与药物发现相关的基准数据集[20, 35]，包括 8 个分类任务和 7 个回归任务 (表 S1)，来评估 FG-BERT 预训练模型的预测能力。Xia 等人收集了 17 种预训练 DL 方

法 (表 1) 作为基线[21]。我们严格按照 Mole-BERT 的数据分割策略进行比较。例如，我们采用支架拆分法，在分类任务中使用 10 个不同的随机种子 (0-9)，在回归任务中使用 3 个不同的随机种子 (1-3)，将数据集按 8:1:1 的比例拆分为训练集、验证集和测试集。计算每个数据集的 ROC-AUC 或 RMSE 的平均值和标准偏差，作为 FG-BERT 的最终结果。表 1 和表 2 总结了详细的性能结果。

在分类任务中, FG-BERT 在八个数据集中的五个数据中表现最佳 (表 1), 包括 Tox21 (AUC = 0.784)、ToxCast (AUC = 0.663)、Sider (AUC = 0.640)、ClinTox (AUC = 0.832) 和 Bace (AUC = 0.845)。此外, FG-BERT 在这八项分类任务中总体表现最佳, AUC 值最高, 为 0.7492 (表 2)。在回归任务方面, FG-BERT 在所有四个基准数据集 (ESOL、Lipo、Malaria 和 CEP) 上都表现最佳, 平均 RMSE 值最低, 为 0.927。此外, 与其他监督学习模型相比, 我们的 FG-BERT 模型在回归任务的总体平均值上提高了 7.2%。同时, 我们还统计了 FG-BERT 模型与各基准 DL 模型在这些常用公共基准数据集上的对比情况。如表 S5 所示, 我们的 FG-BERT 模型不仅在每个基线上表现优异, 而且在所有基线上都表现优异 (表 1)。

为了进一步证明 FG-BERT 模型的全面性和适应性, 我们添加了更先进的基线模型 (表 S6), 作为与回归数据集的比较。如表 S6 所示, FG-BERT 在 FreeSolv 和 Lipo 数据集上表现最佳, 在 ESOL 数据集上排名第二。此外, 两个量子化学数据集 QM7 和 QM8 也用于测试 FG-BERT 模型的预测性能。FG-BERT 在 QM8 数据集上的表现最好, 在 QM7 数据集上的表现排名第二 (表 S6), 这意味着 FG-BERT 模型具有全面性和适应性。综上所述, 这些结果表明 FG-BERT 在预测分子性质方面优于 SOAT DL 模型。因此, 我们提出的 FG-BERT 模型在预测药物发现领域的分子性质方面具有很强的竞争力。

FG-BERT 在 ADMET 数据集上的性能

为了进一步说明 FG-BERT 在预测分子性质方面的优越性, 我们共使用了 15 个与药物发现相关的 ADMET 数据集 (表 S2) [6]。根据 Wu 等人的研究[6], 我们选择了以下先进模型作为比较基准模型, 包括两种基于图的竞争方法 (HRGCN+ 和 Attentive FP) [9, 10]、两种基于指纹的 XGBoost 模型 (XGBoost-MACCS 和 XGBoost-ECFP4) [46, 47] 和一种基于知识的预训练模型 K-BERT [6]。为了进行公平比较, 我们使用了与 K-BERT 相同的建模数据、拆分方法和数据拆分率。此外, 我们将每个数据集测试集上 10 个不同随机种子的结果取平均值, 作为模型的最终结果。FG-BERT 在这些 ADMET 数据集上的详细性能结果如图 3A 所示。如图 3B 所示, 与所有基线模型相比, FG-BERT 模型的分子性质预测性能最好, 平均 AUC 值最高, 为 0.813。K-BERT 预训练模型的性能排名第二, 其次是 HRGCN+、XGBoost-MACCS、XGBoost-ECFP4 和 Attentive FP。显然, 在这些 ADMET 数据集上, 两个基于 BERT 的预训练模型比三个非预训练模型表现更好, 这表明预训练模型在药物发现领域具有一定的优势。究其原因, 可能是基于 BERT 的预训练模型能从大量未标注数据集中学习到更准确、更有用的分子表征。同时, FG-BERT 优于 K-BERT, 主要是因为我们的模型更关注与分子性质相关的重要结

构 FG, 使得 FG-BERT 预训练模型能够从分子中提取重要的化学结构和语义信息。总之, FG-BERT 模型的出色预测能力

表 1: FG-BERT 在常用公共数据集的八项分类任务上的性能结果 (ROC-UAC)

方法	Tox21	ToxCast	Sider	临床毒理学	MUV	艾滋病毒	BBBP	贝丝	平均
InfoGraph [57]	73.3 ± 0.6	61.8 ± 0.4	58.7 ± 0.6	75.4 ± 4.3	74.4 ± 1.8	74.2 ± 0.9	68.7 ± 0.6	74.3 ± 2.6	70.10
GPT-GNN [58]	74.9 ± 0.3	62.5 ± 0.4	58.1 ± 0.3	58.3 ± 5.2	75.9 ± 2.3	65.2 ± 2.1	64.5 ± 1.4	77.9 ± 3.2	68.45
EdgePred [59]	76.0 ± 0.6	64.1 ± 0.6	60.4 ± 0.7	64.1 ± 3.7	75.1 ± 1.2	76.3 ± 1.0	67.3 ± 2.4	77.3 ± 3.5	70.08
ContextPred [60]	73.6 ± 0.3	62.6 ± 0.6	59.7 ± 1.8	74.0 ± 3.4	72.5 ± 1.5	75.6 ± 1.0	70.6 ± 1.5	78.8 ± 1.2	70.93
GraphLoG [61]	75.0 ± 0.6	63.4 ± 0.6	59.6 ± 1.9	75.7 ± 2.4	75.5 ± 1.6	76.1 ± 0.8	68.7 ± 1.6	78.6 ± 1.0	71.56
G-Contextual [62]	75.0 ± 0.6	62.8 ± 0.7	58.7 ± 1.0	60.6 ± 5.2	72.1 ± 0.7	76.3 ± 1.5	69.9 ± 2.1	79.3 ± 1.1	69.34
G-Motif [62]	73.6 ± 0.7	62.3 ± 0.6	61.0 ± 1.5	77.7 ± 2.7	73.0 ± 1.8	73.8 ± 1.2	66.9 ± 3.1	73.0 ± 3.3	70.16
AD-GCL [63]	74.9 ± 0.4	63.4 ± 0.7	61.5 ± 0.9	77.2 ± 2.7	76.3 ± 1.4	76.7 ± 1.2	70.7 ± 0.3	76.6 ± 1.5	72.16
JOAO [64]	74.8 ± 0.6	62.8 ± 0.7	60.4 ± 1.5	66.6 ± 3.1	76.6 ± 1.7	76.9 ± 0.7	66.4 ± 1.0	73.2 ± 1.6	69.71
SimGRACE [65]	74.4 ± 0.3	62.6 ± 0.7	60.2 ± 0.9	75.5 ± 2.0	75.4 ± 1.3	75.0 ± 0.6	71.2 ± 1.1	74.9 ± 2.0	71.15
GraphCL [66]	75.1 ± 0.7	63.0 ± 0.4	59.8 ± 1.3	77.5 ± 3.8	76.4 ± 0.4	75.1 ± 0.7	67.8 ± 2.4	74.6 ± 2.1	71.16
GraphMAE [67]	75.2 ± 0.9	63.6 ± 0.3	60.5 ± 1.2	76.5 ± 3.0	76.4 ± 2.0	76.8 ± 0.6	71.2 ± 1.0	78.2 ± 1.5	72.30
3D InfoMax [19]	74.5 ± 0.7	63.5 ± 0.8	56.8 ± 2.1	62.7 ± 3.3	76.2 ± 1.4	76.1 ± 1.3	69.1 ± 1.2	78.6 ± 1.9	69.69
GraphMVP [20]	74.9 ± 0.8	63.1 ± 0.2	60.2 ± 1.1	79.1 ± 2.8	77.7 ± 0.6	76.0 ± 0.1	70.8 ± 0.5	79.3 ± 1.5	72.64
MGSSL [68]	75.2 ± 0.6	63.3 ± 0.5	61.6 ± 1.0	77.1 ± 4.5	77.6 ± 0.4	75.8 ± 0.4	68.8 ± 0.6	78.8 ± 0.9	72.28
AttrMask [60]	75.1 ± 0.9	63.3 ± 0.6	60.5 ± 0.9	73.5 ± 4.3	75.8 ± 1.0	75.3 ± 1.5	65.2 ± 1.4	77.8 ± 1.8	70.81
Mole-BERT [21]	76.8 ± 0.5	64.3 ± 0.2	62.8 ± 1.1	78.9 ± 3.0	78.6 ± 1.8	78.2 ± 0.8	71.9 ± 1.6	80.8 ± 1.4	74.04
FG-BERT	78.4 ± 0.8	66.3 ± 0.8	64.0 ± 0.7	83.2 ± 1.6	75.3 ± 2.4	77.4 ± 1.0	70.2 ± 0.9	84.5 ± 1.5	74.92

Mole-BERT 的数据分割方法是将每个分类数据集按 8:1:1 的比例分割成训练集、验证集和测试集[21]。FG-BERT 使用相同的数据集和数据拆分方法进行公平比较。每个数据集的最佳结果以粗体标出。标准偏差位于平均值之后。所有比较结果均来自 Mole-BERT [21]。

表 2: FG-BERT 在常用公共数据集的四项回归任务上的性能结果 (RMSE)

方法	ESOL	脂肪	疟疾	CEP	平均
ContextPred [60]	1.196 ± 0.037	0.702 ± 0.020	1.101 ± 0.015	1.243 ± 0.025	1.061
JOAO [64]	1.120 ± 0.019	0.708 ± 0.007	1.145 ± 0.010	1.293 ± 0.003	1.067
GraphMVP [20]	1.064 ± 0.045	0.691 ± 0.013	1.106 ± 0.013	1.228 ± 0.001	1.022
AttrMask [60]	1.112 ± 0.048	0.730 ± 0.004	1.119 ± 0.014	1.256 ± 0.000	1.054
Mole-BERT [21]	1.015 ± 0.030	0.676 ± 0.017	1.074 ± 0.009	1.232 ± 0.009	0.999
FG-BERT	0.944 ± 0.025	0.655 ± 0.009	1.057 ± 0.006	1.051 ± 0.029	0.927

Mole-BERT 的数据分割方法是将每个回归数据集按 8:1:1 的比例分割成训练集、验证集和测试集[21]。FG-BERT 使用相同的数据集和数据分割方法进行公平比较。每个数据集的最佳结果以粗体标出。标准偏差位于平均值之后。所有比较结果均来自 Mole-BERT [21]。

使其成为预测分子 ADMET 特性的最具竞争力的方法之一。

在基于细胞的表型筛选数据集上，FG-BERT 的性能与基于图谱和指纹的高级模型进行比较

基于表型的筛选，如全细胞活力，是一种原始但不可或缺的药物筛选方法，近年来再次受到关注[48-52]。因此，我们评估了 FG-BERT 在 13 个乳腺癌细胞系和 1 个正常乳腺癌细胞系表型筛选数据集上的预测性能（表 S3）[40]。为了与其他模型进行公平比较，我们还使用了一致的数据集、分割方法以及相同的数据分割比例（训练集、验证集和测试集之间的比例为 8:1:1），然后我们计算了从 10 个不同随机种子中得到的 AUC 值的平均值，作为 FG-BERT 的最终结果。图 3C 和 D 显示了 FG-BERT 在 14 个细胞系筛选数据集上的详细性能结果。可以发现，FG-BERT 的性能全面优于这些基线模型。例如，FG-BERT 在 14 个细胞系中的 9 个（即 MDA-MB-453、SK-BR-3、T-47D、MCF-7、BT-474 和 BT-20）上表现最佳、

BT-549、MDA-MB-231 和 HBL-100），而 Attentive FP 在 2 个细胞系（HS-578T 和 Bcap37）上表现最佳，XGBoost 在 2 个细胞系（MDA-MB-361 和 MDA-MB-468）上表现最佳，GCN 在 MDA-MB-435 上表现最佳。重要的是，FG-BERT 在这些细胞系上取得了最佳的整体性能。

14 个细胞系的平均 AUC 值最高，为 0.856 (图 3D)。统计结果表明，与排名第二的 XGBoost 模型 (AUC = 0.813) 相比，FG-BERT 模型的总体准确率提高了 4.3%。这些结果充分证明了 FG-BERT 在基于细胞的表型筛选数据集上的优异表现，表明 FG-BERT 在基于细胞的表型筛选药物发现方面具有巨大潜力。

消融研究

掩盖 FG 是否有用

为了证明屏蔽 FG 预训练的有效性，我们将 FG-BERT 与使用随机屏蔽原子进行预训练的 MG-BERT [18]进行了比较。由于其原始论文使用的是随机分割法，因此我们也使用随机分割法对相同的数据集进行分割，以公平地评估模型性能。如图 4A 和 B 所示，在所有五个数据集中，FG-BERT 的预测性能均优于 MG-BERT，整体相对提高了 11.7%，其中分类和回归任务的平均性能分别提高了 1.9% 和 18.2%。毫无疑问，BERT 模型通过屏蔽 FG 在预训练中是有效的。

预培训是否确实有效

为了证明预训练确实可以提高分子性质预测任务的准确性，我们测试了预训练和未预训练的 FG-BERT 在以下条件下的性能

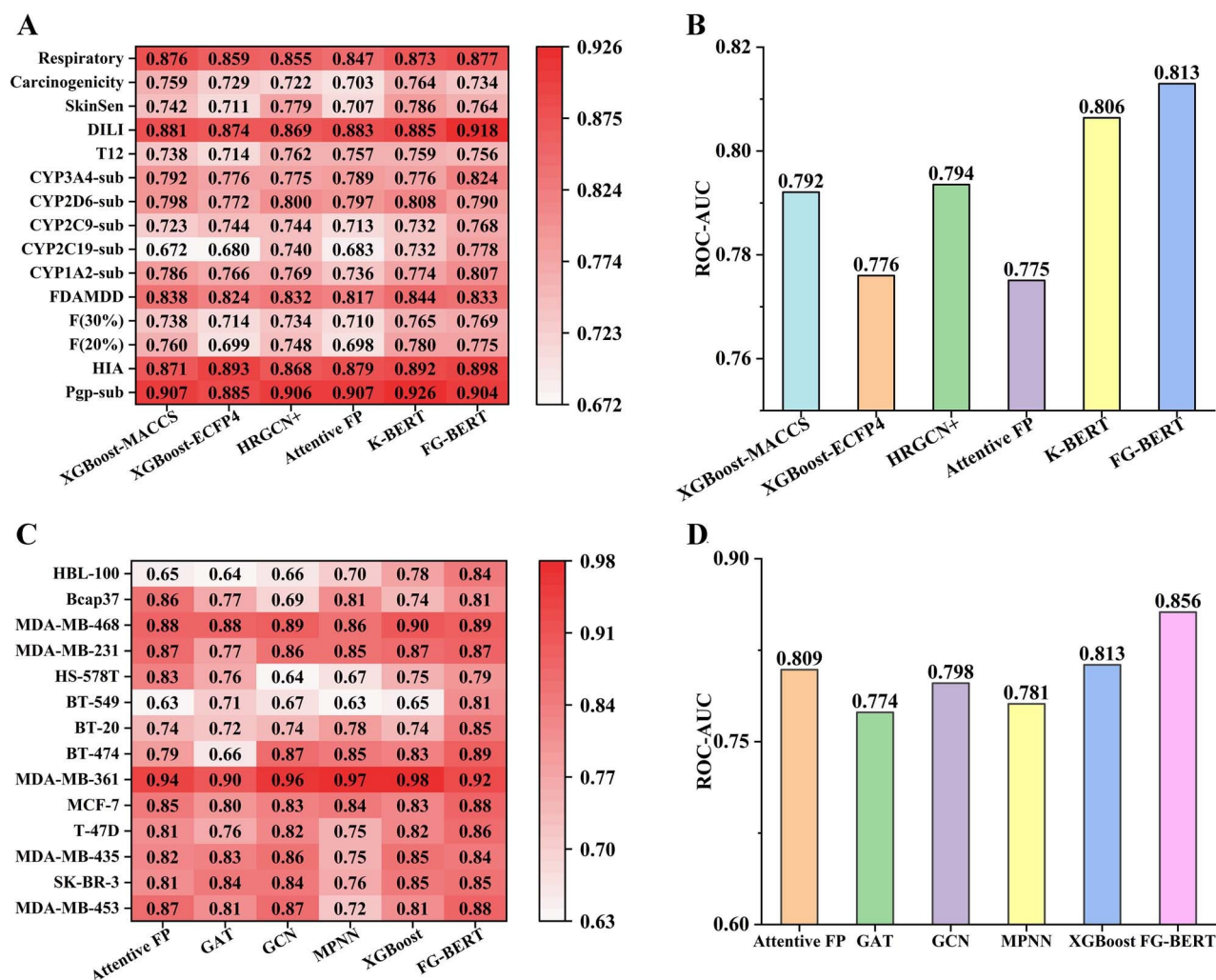


图 3.FG-BERT 与基线模型在 15 个 ADMET 数据集上的性能比较。(A) FG-BERT 和基线模型在测试集上的详细 AUC 值。(B) FG-BERT 和基线模型的平均 AUC 值。所有五个基线模型的性能均来自 Wu 等人的研究[6]。FG-BERT 与基线模型在 14 个基于乳腺细胞的芬太尼筛选数据集上的表现比较。(C) FG-BERT 和基线模型在测试集上的 AUC 值详情。(D) FG-BERT 和基线模型的平均 AUC 值。所有五个基线模型的性能均来自 He 等人的研究[40]。

同一组超参数。在非预训练条件下，模型参数使用下游任务的初始化权重进行微调。图 4C 和 D 列出了详细的比较结果。结果表明，在同一组超参数下，预训练 FG-BERT 模型的性能优于未预训练的 FG-BERT 模型，在分类任务中平均性能提高了近 7.9%，在回归任务中平均性能提高了近 9.6%。这些结果表明，预训练实际上使 FG-BERT 能够从大规模未标记的分子中捕获丰富的结构和语义信息，提取有效的分子表征，并通过简单的神经网络工程将其轻松迁移到特定的下游任务中，从而增强模型的预测能力。

过滤预训练数据集是否有效

我们进一步测试了基于从 ChEMBL 收集的未过滤分子数据集构建的 FG-BERT 模型的性能，以获得六项分类任务的权重。图 5 显示，基于过滤数据集构建的 FG-BERT 模型的性能

测性能优于基于未过滤数据集的 FG-BERT 模型。这可能是因为

基于 "类药物三原则" 过滤后的语料库包含了更高质量的分子, 这使得 FG-BERT 模型更容易识别类药物分子, 从而提取出更好的分子表征, 这也使得模型在下游预测任务中更加准确。

不同FG 掩蔽率的影响

我们还对预训练模型的 FGs 屏蔽率进行了消融实验。预训练设定了 0.15、0.2、0.3 和 0.5 四种掩蔽率, 并使用公共常用数据集进行微调。值得注意的是, 由于 HIV 数据集太大, 需要大量计算时间, 因此没有采用。我们使用随机支架拆分数据集进行下游任务, 并计算了基于三个随机种子的评价指标平均值作为最终结果。如图 4E 和 F 所示, FG-BERT 模型在 FGs 屏蔽率为 0.15 时达到了平均最优性能, 这与 KPGT [53] 在屏蔽率为 0.5 和 ATMOL [54] 在屏蔽率为 0.25 时获得的最佳结果不同。最近在 CV 领域进行的一项研究表明, KPGT 的高掩蔽率和 ATMOL [54] 的 0.25 掩蔽率所获得的最佳结果是不同的。

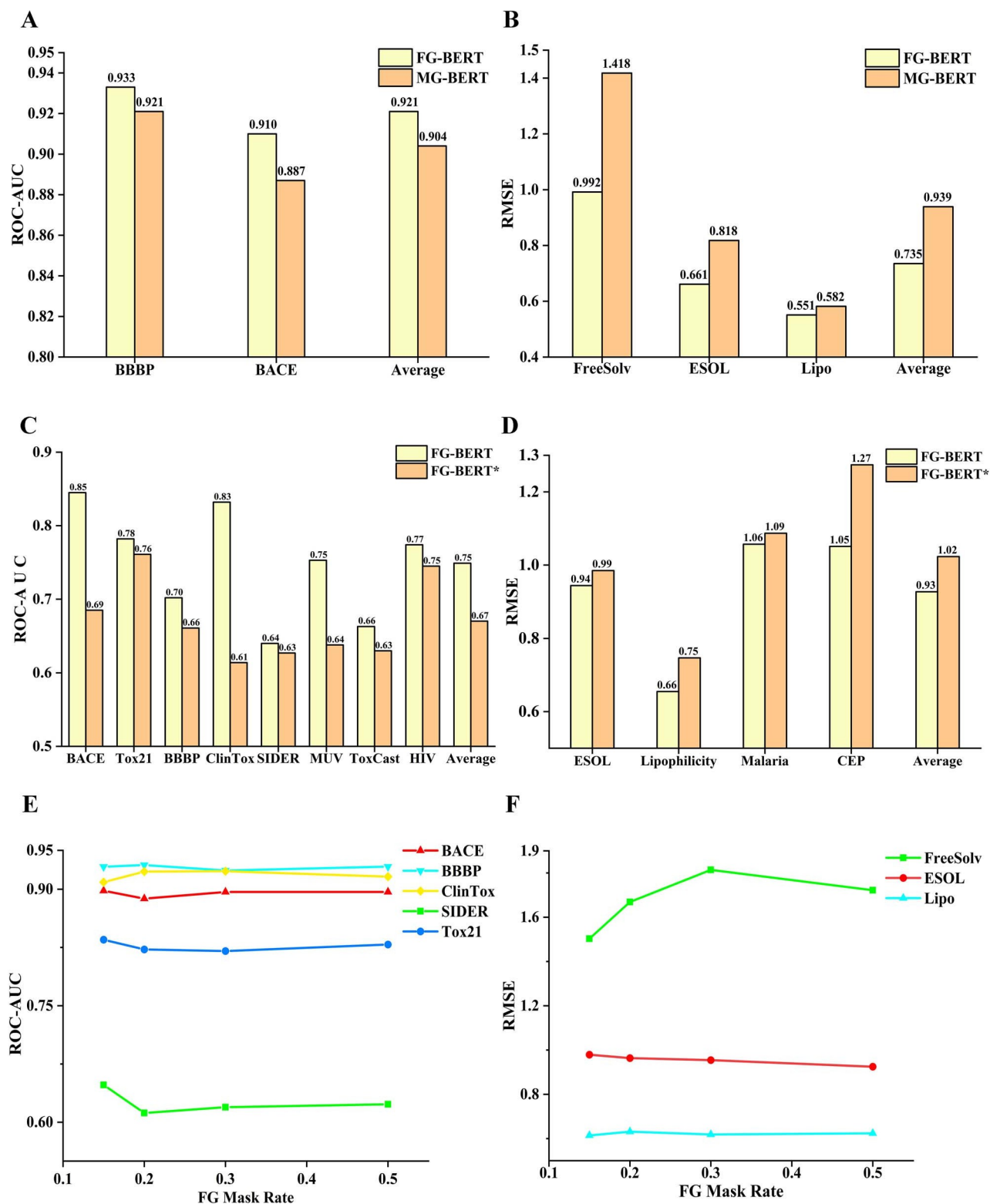


图 4. FG-BERT 的消融实验。FG-BERT 和 MG-BERT 在分类任务 (A) 和回归任务 (B) 上的性能比较。数据来自 MG-BERT [18]。经过预训练的 FG-BERT 与未经过预训练的 FG-BERT 在分类任务 (C) 和回归任务 (D) 上的性能比较。FG-BERT* 表示未经过预训练的 FG-BERT。FG 屏蔽率对下游分类任务 (E) 和回归任务 (D) 的影响 (F)。

这与我们的结论相矛盾。原因可能是高掩蔽率会使模型难以学习语义。

这反过来又会使模型提取的分子表征质量不高，影响模型的性能。

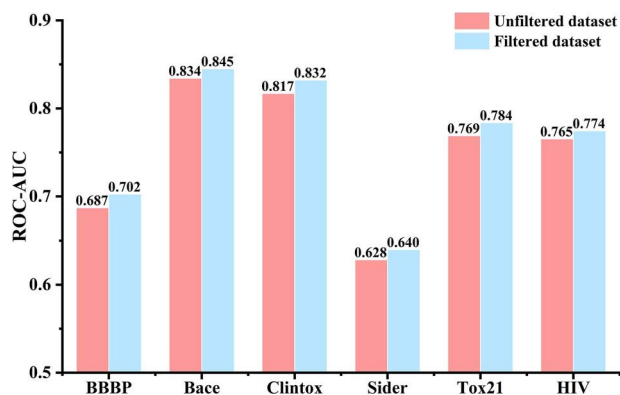


图 5 有无过滤预训练语料的 FG-BERT 模型结果比较。

FG-BERT 的解释

全面了解分子结构与其性质之间的关系对于分析和进一步优化先导化合物至关重要，这就需要进一步探索 FG-BERT 模型的可解释性。目前，FG-BERT 可以通过 AM 聚合所有原子和 FG 表征的信息，生成整个分子的表征，从而揭示这种关系，同时可以生成注意权重，用于表示原子和/或 FG 在分子表征中的重要性，因此可以将其视为目标性质相关性的衡量标准。

首先，利用基于血脑屏障渗透性（BBBP）数据集的 FG-BERT 模型来分析模型的可解释性。由于血脑屏障（BBB）阻止了大多数药物和激素的进入，因此准确预测分子的 BBBP 对于开发治疗中枢神经系统疾病的药物至关重要。通常情况下，疏水性分子由于极性低和 ClogP 高而更有可能穿过 BBB，而亲水性分子则相反。如图 6 所示，我们很容易就能直观地看出相关分子的关注权重，红色越深表示这些分子的权重越高，反之亦然。以一个可渗透分子为例（图 6A），分子中的苯环和环己烷（极性最低的高疏水 FGs）对 BBB 的影响最大。我们使用 ChemBioDraw（版本 14.0.0.117）进一步量化了这些 FGs 的 ClogP 值。不难看出，分子左侧部分（图 6A）的 ClogP 值为 0.547，而右侧部分的 ClogP 值为 5.291，这与 FG-BERT 模型的预测一致。同时，对于防渗分子，FG-BERT 倾向于将注意力集中在分子左侧的氨基和羟基上（图 6B），这两个基团提供了大部分极性以阻止分子穿过 BBB。此外，左侧 FG 的 ClogP 值

为 -0.575，表明红色部分

分子的亲水性更强，在通过 BBB 时可能会遇到困难。我们的 FG-BERT 模型中红色部分标注的高关注度与非活性预测结果一致。很明显，我们的 FG-BERT 模型中标注为红色的高关注度 FG 与不渗透分子的预测结果一致。

β -分泌酶 1（BACE-1）是人体内参与淀粉样前体蛋白（APP

）裂解的一种酶[55]。APP 的裂解会产生 β -淀粉样蛋白，而 β -淀粉样蛋白是阿尔茨海默病（AD）的主要沉积成分之一，因此 BACE-1

被认为是治疗 AD 的重要靶点 [34]。在此，我们选择了基于 BACE 数据集的最优模型来进一步探究 FG-BERT 模型的可解释性。如图 6C 和 D 所示，从测试集中选择了两个分子（BACE_350 和 BACE_1015）进行案例研究。值得注意的是，这两个分子具有相同的支架，但 BACE_350 ($\text{pIC}_{50} = 8.22$) 是活跃的，而 BACE_1015 ($\text{pIC}_{50} = 6.35$) 是不活跃的 [56]。根据两个分子的原子注意权重的可视化（图 6C 和 D），很明显该模型捕捉到了分子中重要的 FG。我们推测 FG-BERT 关注的是 BACE_350 和 BACE_1015 之间的差异，因为 BACE_350 包含一个六元环和一个炔基取代基，而 BACE_1015 的相应部分包含一个不含炔基取代基的五元环。此外，还利用 Glide SP docking 研究了 BACE_350 和 BACE_1015 与 BACE-1 的结合模式，两个分子的二维蛋白质-配体相互作用分别如图 6C 和 D 所示。Glide 评分表明，BACE_350 (docking score = -6.786 kcal/mol) 对 BACE-1 的抑制活性优于 BACE_1015 (docking score = -3.823 kcal/mol)，这与酶抑制实验结果一致。根据对接结果，分子中突出显示的氨基可以与 ASP32 和 ASP228 形成两个关键氢键，这表明我们的模型可以自动学习分子中的关键 FGs 信息，并将其应用于分子性质预测任务中。此外，图 6E 显示了这两种分子与 BACE-1 的详细三维结合模式。值得注意的是，BACE_350 的炔基 FG 面向 BACE-1 的 S3 疏水口袋，从而产生了强烈的疏水相互作用，这也与 BACE_350 的高亮 FG 相一致。然而，由于 BACE_1015 中缺少炔基 FG，因此没有观察到这些疏水相互作用，这可能是 BACE_1015 对 BACE-1 抑制活性较差的原因。这些结果表明，我们的 FG-BERT 模型可以识别与生物活性相关的关键相互作用模式。

结论

鉴于分子性质预测领域标注数据的稀缺性，我们在本研究中提出了一种新的自我监督学习框架，称为 FG-BERT。FG-BERT 模型通过屏蔽分子图中的 FG，实现高效的恢复预训练，并从 145 万个未标记的分子中全面挖掘化学结构和语义信息，学习有用的分子表征。通过微调策略，FG-BERT 预训练模型可轻松用于下游分子性质预测任务。系统评估结果表明，FG-BERT 预训练模型具有很强的竞争力，即使与最先进的传统监督 ML 和 DL 模型以及自监督预训练模型相比也不遑多让。此外，FG-BERT 模型的高可解释性还能让用户充分理解分子结构与其特性之间的关系，从而提供宝贵的片段/FGs 信息，帮助科学家更准确地设计和识别具有所需功能和/或治疗效果分子。总之，我们预计 FG-BERT 作为一种开箱即用、有效且可解释的计算工具，可用于各种药物发现相关任务。FG-BERT 目前主要关注分子中的 FG 信息，并没有充分考虑以下因素

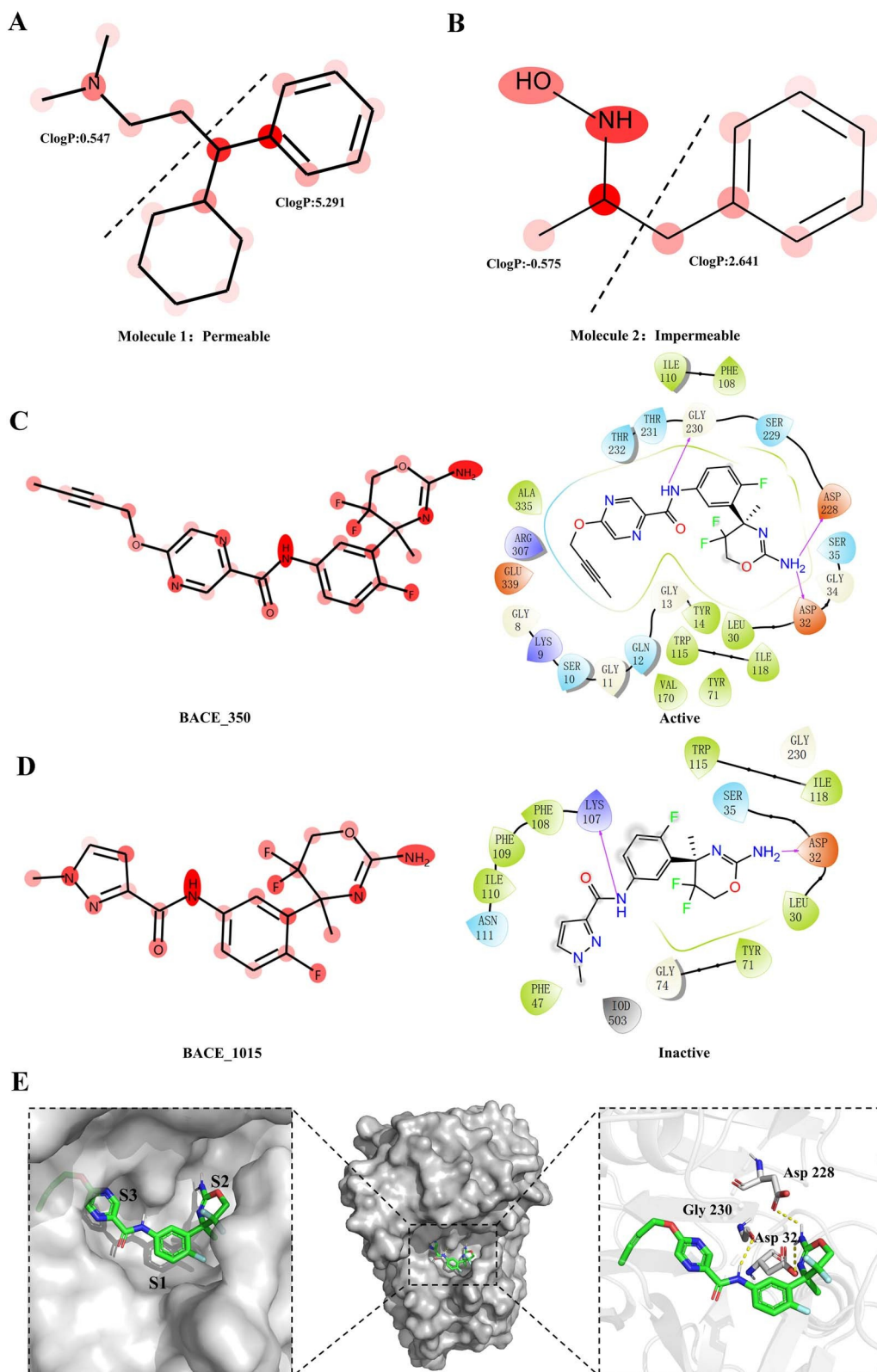


图 6.分子结构在预测过程中的重要性。颜色越深,说明结构越重要。分子 1 和分子 2 来自 BBBP (BBB 穿透) 数据集。BACE_350 和 BACE_1015 来自 BACE 数据集。

(A) 分子 1 具有渗透性,颜色较深的部分具有较高的 ClogP,表示亲脂性较强。(B) 分子 2 是不可渗透的,颜色较深的部分 ClogP 较低,表示亲脂性较弱。FG-BERT 模型捕捉到的重要部分与预测结果一致。(C)和(D)展示了针对 BACE-1 的彩色活性分子和非活性分子,以及二维蛋白质配体结合模式。Glide SP docking 生成的相互作用结合模式。(E) 结合口袋(左)、蛋白质和配体结合的实体表面表示位点(中)和预测的 BACE_350 和 BACE_1015 与 BACE-1 的三维结合模式(右)。所有图表均使用 PyMOL 软件生成 (<https://pymol.org/2/>)。

分子支架信息。通过将分子支架信息与 FG 信息相结合，可以增强模型提取分子特征的能力，从而提高其分子性质预测性能。我们将在今后的工作中进一步探讨这一问题。

要点

- 我们提出了一个名为 FG-BERT 的 DL 预训练模型来预测分子特性。
- 它利用一种自我监督的预训练学习框架，从以下数据中学习有用的分子表征
~145 万个具有不同生物特性的类药物分子活动。
- 广泛的实验结果表明，FG-BERT 与经典的 ML 方法、SOTA 预训练和基于图的 DL 方法相比，具有很强的竞争力。
- FG-BERT 的消融实验表明，该模型在通过屏蔽 FG 恢复预训练后，可以提高模型的下游任务预测性能。
- FG-BERT 模型直观易懂，可提供重要的化学片段，帮助化学家和药剂师设计或优化具有所需特性的新分子。

Conf Mach Learn 2016; 2702-11.

- Li Y, Hsieh C-Y, Lu R, *et al.* 自适应图学习方法用于自动分子相互作用和性质预测。 *Nat Mach Intell* 2022;4:645-51.
- Wu Z, Jiang D, Wang J, *et al.* 基于知识的 BERT：一种像计算化学家一样提取分子特征的方法。 *Brief BIOINFORM* 2022;23:bbac131.

补充数据

补充数据可在线查阅：<http://bib.oxfordjournals.org/>。

资金

本研究得到广东省自然科学基金（2023B1515020042）和国家自然科学基金（81973241）的部分资助。

数据可用性

本研究使用的数据集和 FG-BERT 的源代码可在 <https://github.com/idrugLab/FG-BERT> 网站上公开获取。

参考文献

- Song CM, Lim SJ, Tong JC. 计算机辅助药物设计的最新进展。 *Brief BIOINFORM* 2009; 10:579-91.
- Eklund M, Norinder U, Boyer S, Carlsson L. 在 QSAR 中选择特征选择和学习算法。 *J CHEM Inf Model* 2014;54:837-43.
- Phillips JC, Gibson WB, Yam J, *et al.* *Food CHEM Toxicol* 1990; 28:375-94.
- Dai H, Dai B, Song L. 结构化数据潜变量模型的判别嵌入。 *Int*

7. Wang S, Guo Y, Wang Y, *et al.* SMILES-BERT: 用于分子性质预测的大规模非渗透预训练。In: 第10届ACM生物信息学、计算生物学和健康信息学国际会议论文集, 纽约州尼亚加拉瀑布, 2019年9月7-10日; 429-36.
8. Zeng X, Xiang H, Yu L, *et al.* 使用自监督图像表征学习框架准确预测分子性质和药物靶点。 *Nat Mach Intell* 2022; 1-13.
9. Xiong Z, Wang D, Liu X, *et al.* 利用图注意 机制推动药物发现的分子表示界限。 *J Med CHEM* 2019; **63**:8749-60.
10. Wu Z, Jiang D, Hsieh C-Y, *et al.* 双曲线关系图卷积网络加: 一种简单而高效的QSAR建模方法。 *Brief BIOINFORM* 2021; **22**:bbab112.
11. Cai H, Zhang H, Zhao D, *et al.* FP-GNN: 用于增强分子性质预测的多功能深度学习架构。 *Brief BIOINFORM* 2022; **23**:bbac408.
12. Wu J, Xiao Y, Lin M, *et al.* DeepCancerMap: a versatile deep learning platform for target-and cell-based anticancer drug discovery. *Eur J Med CHEM* 2023; **255**:115401.
13. Ai D, Wu J, Cai H, *et al.* 多任务 FP-GNN 框架实现了选择性 PARP 抑制剂的准确预测。 *Front Pharmacol* 2022; **13**:971369.
14. Zhu W, Zhang Y, Zhao D, *et al.* HiGNN: a hierarchical informative graph neural network for molecular property prediction equipped with feature-wise attention. *J CHEM Inf Model* 2023; **63**: 43-55.
15. Jiang D, Wu Z, Hsieh C-Y, *et al.* 基于描述符和基于图的模型 的比较研究。 *J CHEM* 2021; **13**:1-23.
16. Liu X, Zhang F, Hou Z, *et al.* 自监督学习: 生成性还是对比性。 *IEEE Trans Knowl Data Eng* 2021; **35**:857-76.
17. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need[J]. *神经信息处理系统进展* 2017; 30.
18. Zhang X-C, Wu C-K, Yang Z-J, *et al.* MG-BERT: 利用未渗透原子表征学习进行分子性质预测。 *Brief BIOINFORM* 2021; **22**:bbab152.
19. Stärk H, Beaini D, Corso G, *et al.* *Int Conf Mach Learn* 2022; 20479-502.
20. Liu S, Wang H, Liu W, *et al.* ICLR, ArXiv Prepr. ArXiv211007728. 2021.
21. Xia J, Zhao C, Hu B, *et al.* Mole-BERT: 重新思考分子的预训练图神经网络。 *Elev Int Conf Learn* 2023.
22. Ertl P, Altmann E, McKenna JM. 生物活性分子中最常见的官能团及其受欢迎程度的演变。 *J Med CHEM* 2020; **63**:8408-18.
23. Wadhwa K, Hennissen J, Shetty S, *et al.* 各种官能团的取代对TEMPO类似物抑制苯乙烯聚合效率的影响 *J POLYM Res* 2017; **24**:1-8.
24. Assad H, Kumar A. Understanding functional group effect on corrosion inhibition efficiency of selected organic compounds. *J Mol Liq* 2021; **344**:117755.
25. Iqbal J, Vogt M, Bajorath J. 从分子图像中学习官能团化学从而准确预测活性悬崖。 *Artif Intell Life Sci* 2021; **1**:100022.
26. Gaulton A, Bellis LJ, Bento AP, *et al.* ChEMBL: 用于药物发现的大规模生物活性数据库。 *Nucleic Acids Res* 2012; **40**:D1100-7.

27. Delaney JS. ESOL: 直接从分子结构估算水溶性. *J CHEM Inf Comput Sci* 2004;44:1000-5.
28. Mobley DL, Guthrie JP. FreeSolv: 含输入文件的实验和计算水合自由能数据库. *J Comput Aided Mol Des* 2014;28:711-20.
29. Mendez D, Gaulton A, Bento AP, et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* 2019;47:D930-40.
30. Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, et al. Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J Phys CHEM Lett* 2011;2:2241-51.
31. 艾滋病抗病毒筛选数据. In: NIH/NCI (ed). 2017.
32. Gamo F-J, Sanz LM, Vidal J 等. 抗疟先导物鉴定的数千个化学起点. *自然* 2010; 465: 305-10.
33. Rohrer SG, Baumann K. 基于 PubChem 生物活性数据的虚拟筛选最大无偏验证 (MUV) 数据集. *J CHEM Inf Model* 2009; 49:169-84.
34. Subramanian G, Ramsundar B, Pande V, et al. 使用基于配体的方法对 β -分泌酶1 (BACE-1) 抑制剂进行计算建模. *J CHEM Inf Model* 2016;56:1936-49.
35. Martins IF, Teixeira AL, Pinheiro L, et al. *J CHEM Inf Model* 2012;52:1686-97.
36. Tox21 数据挑战. 美国国立卫生研究院 2017 年.
37. Kuhn M, Letunic I, Jensen LJ, et al. The SIDER database of drugs and side effects. *Nucleic Acids Res* 2016;44:D1075-9.
38. Wu Z, Ramsundar B, Feinberg EN, et al. MoleculeNet: a benchmark for molecular machine learning. *CHEM Sci* 2018; 9: 513-30.
39. Gayvert KM, Madhukar NS, Elemento O. 预测临床试验成功与失败的数据驱动方法. *Cell CHEM Biol* 2016;23:1294-301.
40. He S, Zhao D, Ling Y, et al. 机器学习可准确快速预测抗乳腺癌细胞的活性分子. *Front Pharmacol* 2021;12:3766.
41. Devlin J, Chang M-W, Lee K, et al. BERT: 用于语言理解的深度双向变换器预训练. ArXiv Prepr. ArXiv1810.04805, 2018.
42. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: 学习表征国际会议, 美国加利福尼亚州圣地亚哥, 2015 年. OpenReview.net. 国际表征学习会议, 美国加利福尼亚州拉霍亚。
43. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: 防止神经网络过度拟合的简单方法. *J Mach Learn Res* 2014;15:1929-58.
44. Bergstra J, Yamins D, Cox DD. Hyperopt: 分布式非同步超参数优化. *Astrophys Source Code Libr* 2022.
45. Ji Z, Shi R, Lu J, et al. ReLMole: 基于两级图相似性的分子表征学习. *J CHEM Inf Model* 2022;62: 5361-72.
46. Durant JL, Leland BA, Henry DR, et al. *J CHEM Inf Comput Sci* 2002; 42: 1273-80.
47. Rogers D, Hahn M. Extended-connectivity fingerprints. *J CHEM Inf Model* 2010;50:742-54.
48. Luo Y, Zeng R, Guo Q, et al. 基于细胞的计算生物活性预测模型和生物测定. *Org Biomol CHEM* 2019;17:1519-30.
49. Guo Q, Luo Y, Zhai S, et al. 吡唑并[3, 4-b]吡啶-6-酮衍生物作为一类新型抗癌药物的发现、生物学评价、构效关系及作用机制. *Org BIOMOL CHEM* 2019;17:6201-14.
50. Moffat JG, Vincent F, Lee JA, et al. 表型药物发现的机遇与挑战: 行业视角. *Nat Rev Drug Discov* 2017;16:531-43.
51. Malandraki-Miller S, Riley PR. 利用人工智能加强表型药物发现. *Drug Discov Today* 2021;26: 887-901.
52. Berg EL. 表型药物发现的未来. *细胞化学生物学* 2021; 28:424-30.
53. Li H, Zhao D, Zeng J. KPGT: 用于分子性质预测的图转换器知识引导预训练. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Washington, D.C., USA. 2022; 857-67. ACM.
54. 用于预测分子性质的注意-明智掩蔽图控制学习. *Brief Bioinform* 2022;23:bbac303.
55. Hunt CE, Turner AJ. 细胞生物学、调节和抑制 β -分泌酶 (BACE-1) [J]. *febs j* 2009;276(7):1845-59.
56. Malamas MS, Erdei J, Gunawan I, et al. *J Med CHEM* 2009;52:6314-23.
57. Sun F-Y, Hoffmann J, Verma V, et al. Infograph: unsupervised and semi-supervised graph-level representation learning via mutual information maximization. ArXiv Prepr ArXiv1908.01000. 2019.
58. Hu Z, Dong Y, Wang K, et al. GPT-GNN: 图神经网络的生成预训练. In: *第 26 届 ACM SIGKDD 知识发现与数据挖掘国际会议论文集*, 美国加利福尼亚州虚拟活动. 2020; 1857-67. ACM.
59. Hamilton W, Ying Z, Leskovec J. 大型图上的归纳表征学习. *Adv Neural Inf Process Syst* 2017;30.
60. Hu W, Liu B, Gomes J, et al. 预训练图神经网络的策略. ArXiv Prepr. ArXiv1905.12265 2019.
61. 具有局部和全局结构的自监督图级表征学习. *Int Conf Mach Learn* 2021;11548-58.
62. Rong Y, Bian Y, Xu T, et al. 大规模分子数据的自监督图变换器. *Adv Neural Inf Process Syst* 2020; 33: 12559-71.
63. Suresh S, Li P, Hao C, et al. Adversarial graph augmentation to improve graph contrastive learning. *Adv Neural Inf Process Syst* 2021;34:15920-33.
64. You Y, Chen T, Shen Y, et al. *Int Conf Mach Learn* 2021;12121-32.
65. Xia J, Wu L, Chen J, et al. SimGRACE: 无需数据增强的图对比学习简单框架. *Proc ACM Web Confs* 2022;2022:1070-9.
66. You Y, Chen T, Sui Y, et al. *Adv Neural Inf Process Syst* 2020;33:5812-23.
67. Hou Z, Liu X, Cen Y, et al. Graphmae: self-supervised masked graph autoencoders. *第 28 届 ACM SIGKDD 知识发现与数据挖掘大会论文集*。2022; 594-604.
68. Zhang Z, Liu Q, Wang H, et al. Motif-based graph self-supervised

