arXiv:2602.07087v1 [physics.chem-ph] 6 Feb 2026

面向现实世界分子物理性质的电子信息增强粗粒化分子表示学习

Gyoung S. Na*韩国化学技术研究院 大田，韩国 ngs0@krict.re.kr

Chanyoung Park*韩国科学技术院 大田，韩国 cy.park@kaist.ac.kr

摘要

为加速数据驱动化学的发展，人们设计了多种用于分子结构的表示学习方法。然而，现有方法的表征能力本质上局限于原子级信息，无法充分描述现实世界中的分子物理性质。尽管电子级信息能够提供超越原子级信息的化合物基础知识，但获取现实分子中的电子级信息在计算上并不实际，有时甚至不可行。我们提出了一种学习电子信息分子表示的方法：将易于获取的小分子电子级信息迁移至我们关注的大分子，无需额外计算成本。该方法在包含实验观测分子物理性质的大量基准数据集上取得了最先进的预测准确率。HEDMoL 的源代码可在 https://github.com/ngs00/HEDMoL 获取。

CCS概念

• **计算方法** → **机器学习**；•**应用计算** 化学；物理学。

关键词

表示学习、化学、图神经网络

ACM参考文献格式：

Gyoung S. Na 和 Chanyoung Park。2025年。基于电子信息的粗粒化分子表示学习：面向现实世界分子物理性质。在《第*31*届*ACM SIGKDD*知识发现与数据挖掘大会论文集》第*1*卷（*KDD '25*），*2025*年*8*月*3*日至7日，加拿大安大略省多伦多。ACM，纽约，美国纽约州，12页。https://doi.org/10.1145/3690624.3709270

1 引言

图神经网络（GNN）已成功应用于通过采用分子结构的图表示来预测分子的物理和化学性质 [60]。通常，分子结构被表示为一个属性图$G = (\mathbf{V}, \mathbf{U}, \mathbf{X}, \mathbf{E})$，其中V是节点集合（即原子），U是边集合（即化学键），$\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times d}$ 是一个 d维

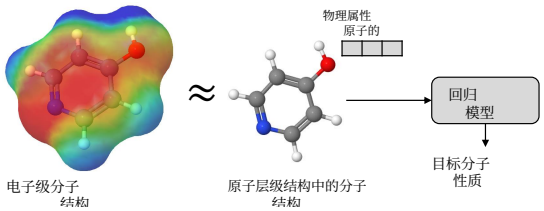


图 1：现有基于图神经网络的方法在分子表示学习中的基本假设，以及其对原子级分子结构进行预测的过程。

节点特征矩阵，而$\mathbf{E} \in \mathbb{R}^{|\mathcal{U}| \times l}$ 是一个 l维边特征矩阵 [57]。已提出多种图神经网络，用于基于分子结构的图表示学习潜在的分子嵌入 [14, 19, 45, 62]。

尽管基于图神经网络的方法在化学数据上取得了诸多成功，但其表征能力本质上受到限制，因为它们仅以原子级 分子结构 [7, 14, 19]来表示分子。现有基于图神经网络的方法假设，包含各原子物理属性的原子级分子结构足以近似电子级分子结构，如图 1 所示。然而，将分子简化为原子级结构的过程中，其电子密度不可避免地会发生扭曲 [33, 38]。本文认为，由于分子的物理和化学特性本质上源自分子的电子密度 [38]，，基于图神经网络的方法可以通过考察原子级之外的电子级分子结构而获益。一种直接且简单的解决方案是，利用密度泛函理论（DFT） [38]和后哈特里–福克方法 [11]等量子力学计算方法计算分子的电子结构，并将电子级信息提供给图神经网络。然而，这种方案并不实用；对于现实世界中的复杂分子，有时甚至无法实现，因为量子力学计算方法具有立方级或更高的时间复杂度 [38]，且电子结构优化可能会陷入局部收敛 [50]。

我们提出层次化电子衍生分子学习（HED- MoL），无需进行昂贵的量子力学计算，即可从输入的原子级分子结构中学习电子信息分子表示。其核心思想是：通过扩展构成输入大分子的小分子的电子级信息，估算以原子级分子结构表示的输入大分子的电子级信息。由于小分子的电子级信息可从公共数据库中轻松获取，因此我们能够减轻

*通讯作者

出版权已授权给 ACM。ACM 承认本贡献已由国家政府的雇员、承包商或附属人员撰写或合著。因此，政府保留非独占、免版税的权利，可仅为政府用途出版或复制本文，或允许他人这样做。请向所有者/作者申请许可。*KDD '25*，*2025*年*8*月*3*日至7日，加拿大安大略省多伦多
© 2025年版权归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 979-8-4007-1245-6/25/08 https://doi.org/10.1145/3690624.3709270

大型分子所需的昂贵量子力学计算带来的计算负担。更具体地说，HEDMoL 同时利用输入的原子级分子信息以及通过以下三个步骤估算的电子级信息，学习电子信息分子表示：(1) HEDMoL 基于图分解算法，将输入的原子级分子结构分解为小型子结构。(2) HEDMoL 根据分子距离，将存储在公共计算数据库中的电子级属性转移到各个分解后的子结构上，构建电子信息增强的分子图；我们将这一过程称为知识扩展。(3) HEDMoL 通过对原子级分子图和电子信息增强的分子图的潜在嵌入进行层次化表示学习，生成最终的分子表示。

值得注意的是，在本研究中，我们重点评估预测模型在实验数据集上的预测能力，而非模拟数据集（例如 QM9 [40] 和 Pub- ChemQC [35] 数据集）。尽管模拟数据集有助于分析原子几何结构与分子性质之间的大致关系，但由于它们未能充分模拟量子力学中的不确定性和复杂构型，因此不适合用于评估预测模型对现实世界分子物理性质的预测能力 [48, 53]。我们从物理化学、毒性和药代动力学应用领域收集了八个通过实验生成的分子数据集，用于评估预测模型对复杂现实世界分子物理性质的预测能力。HEDMoL 在预测分子经实验观察到的物理和化学性质方面达到了最先进的准确率。此外，在小型训练数据集上的各种回归任务中，HEDMoL 的表现优于最先进的图神经网络；这也是化学应用中机器学习面临的主要挑战之一 [4, 23]。

2 相关工作

2.1 分子上的图神经网络

图神经网络已被广泛研究用于处理关系数据 [60]。图卷积网络（GCN） [27] 是一种通过泛化卷积算子来学习关系数据潜在表示的神经网络。图注意力网络（GAT） [54]基于自注意力机制整合来自相邻节点的消息，以学习潜在节点嵌入。除GCN和GAT外，还基于先进的消息传递机制提出了多种图神经网络，例如自适应滤波器 [52],Transformer机制 [47], 和特征级线性调制 [2]。

二维分子图上的图神经网络已成功应用于预测分子的物理和化学性质 [57, 58]。基于消息传递神经网络（MPNN）的通用框架被提出，用于学习原子结构上的量子力学 [19]。有向消息传递神经网络（DMPNN）是MPNN的一种变体，用于在二维分子图上学习定向消息传播[66]。多种图神经网络被提出用于二维分子图上的表示学习 [7, 62]。

基于三维分子图的图神经网络也已在分子科学中得到研究。SchNet [45] 是一种图卷积网络，通过分子原子级表示学习量子力学。DimeNet [18]和 PhysChem [67]旨在通过传播局部和全局原子几何结构来生成分子嵌入。除此之外，人们还开发了各种基于三维分子图的图神经网络 [6, 14]。然而，尽管现有图神经网络在多个计算数据集上表现出最先进的性能，但它们在现实世界分子物理性质研究中的适用性仍然有限，因为受原子位置的不确定性和测量设备局限性的影响，难以准确确定现实世界分子的三维原子坐标 [28, 51]。

2.2 分子数据集上的迁移学习

化学应用中的机器学习通常受训练数据匮乏的困扰，因为开展化学实验以收集训练数据既昂贵又耗时[21, 46]。为克服训练数据不足的问题，迁移学习在物理学和化学领域受到了极大关注 [4, 23]。若干迁移学习方法已通过将在大规模源计算数据集上训练的嵌入网络迁移到目标实验数据集，成功应用于多种物理与化学应用 [4, 23]。尽管如此，现有迁移学习方法的预测能力本质上仍受到限制，因为分子结构优化所需的量子力学计算具有很高的时间复杂度，源计算数据集无法覆盖真实世界实验数据集中的大多数大分子 [15, 38]。

除传统迁移学习方法外，人们还提出了多种分子表示学习方法，通过将分解后的子结构的知识迁移到整个分子中，以学习信息丰富的分子表示 [26, 56, 69]。然而，现有方法的表征能力本质上仅限于原子级信息，因为它们忽略了分子性质与分子电子密度之间的基本关系 [26, 56, 68, 69]。此外，一些现有迁移学习方法需要额外信息，而这些信息在真实世界的分子中并不可得 [56]。

3 提出的方法：HEDMoL

3.1 问题重构

人们提出了多种方法，试图通过不同途径学习富有信息量的分子表征，例如分子几何结构、 [28],基于片段的学习 [26, 69],以及领域知识整合 [56]。然而，这些方法的表征能力从根本上受限于原子级分子结构，因为它们忽略了量子力学原理：分子的物理和化学性质本质上源自其电子密度 [15, 38]。

从物理角度来看，获取分子性质y 的计算过程被定义为一个复合函数，如下所示：

$$y = (g \circ h)(\mathcal{E}), \tag{1}$$

其中 E 是分子的电子密度， h 是一个物理信息函数，用于从 E 生成数值表示，而 g 是一个根据电子级描述符 h(E) 计算物理和化学性质的函数。由于计算电子密度 E 的计算方法相对于原子数量具有三次或更高的时间复杂度，现有

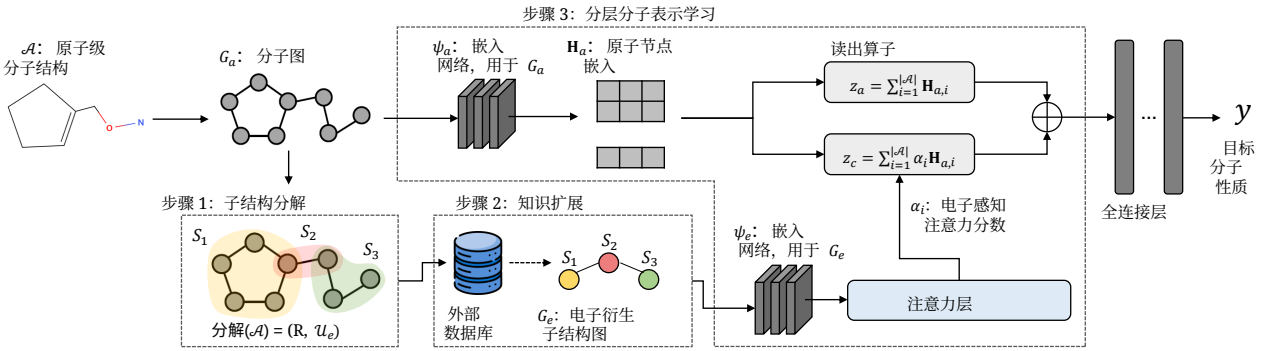


图 2：HEDMoL 用于预测输入原子级分子结构 $\mathcal{A}$ 的目标分子性质 y的整体表示学习与预测流程。在示例输入分子中，$\mathcal{R}$ = {S_1, S_2, S_3} 是一组分解得到的原子级子结构，$\mathcal{U}_e$ 是{S_1, S_2, S_3}之间的边集合。

基于图神经网络的方法基本上假设 h(E)可以由原子级分子结构 A 充分近似，以避免电子结构计算不切实际的时间复杂度。然而，在近似过程 A ≈ h(E) [33, 38]中，信息损失不可避免。

为了高效地利用电子级信息进行分子表示学习，我们将针对原始分子的预测问题重新表述为针对分解后的原子级子结构 S_1、S_2、……、S_K 的问题，其中 K 表示分解后的子结构数量。重新表述后的预测问题如下：

$$y = (g \circ h)(\mathcal{E}_1, \mathcal{E}_2, ..., \mathcal{E}_k), \tag{2}$$

其中，E_k 是小型子结构 S_k 的电子密度，可从公共化学数据库 [22, 40] 中获取。根据公式（2），通过从公共化学数据库中获取 E_k，我们无需进行昂贵的量子力学计算，即可在分子表示学习中利用电子级信息。如何将关于小型原子结构的理论计算信息扩展到现实世界中的复杂分子，是计算化学领域长期以来面临的挑战 [10]。

3.2 HEDMoL 的整体架构

图 2 展示了 HEDMoL 的整体表示学习和预测过程，该过程包括以下三个步骤：(1) 子结构分解；(2) 知识扩展；(3) 分层分子表示学习。HEDMoL 的每个步骤总结如下。**(步骤 1)** 输入的原子级分子结构 A 被分解为一组原子级子结构 R = {S_1、…、S_K}。每个分解子结构 S_k 表示 E_k的原子级描述符。**(步骤 2)** HEDMoL 基于外部计算数据库中的分子距离度量，为 E_k 赋予电子级属性。分解后的子结构被转换为电子衍生子结构图 G_e = (R、U_e、$\mathbf{X}_e$)，其中 R = {S_1、……、S_K} 是节点集合，U_e是边集合，$\mathbf{X}_e$是包含所赋电子级属性的特征矩阵。**(步骤 3)**HEDMoL 通过用于分子图 G_a和电子衍生子结构图 G_e的基于 GNN 的嵌入网络 ψ_a和 ψ_e，分别计算原子级和电子级分子嵌入 $\mathbf{z}_a$ 和 $\mathbf{z}_e$。对于计算得到的 $\mathbf{z}_a$和 $\mathbf{z}_e$，HEDMoL 通过拼接在不同层级生成的分子嵌入 $\mathbf{z}_a$ 和 $\mathbf{z}_e$，生成最终的分子嵌入。最后，HEDMoL 根据最终的分子嵌入预测目标分子性质 y。

3.3 子结构分解

HEDMoL 的子结构分解是一个数据预处理步骤，用于将输入的原子级分子结构转换为一个元组 (R，U$_e$)。从物理意义上讲，每个分解子结构 S_k 都表示公式（2）中分解后的电子密度 E_k 的原子级表征。在分解过程中，我们施加约束 A = $S_1 \cup ... \cup S_K$，以防止子结构分解过程中发生信息损失。在接下来的知识扩展步骤中，分解后的子结构将被转换为 G_e，其中包含关于 S_1、S_2、……、S_K 的电子级信息。

在子结构分解的各种实现方法中，例如谱聚类 [37]，BRICS 分解 [31], 和连接树算法 [25], ，我们采用了连接树算法，因为它具有以下三项优势：(1) 连接树算法无需针对每个输入图进行超参数调优。(2) 它能为分子结构 [25]提供稳健的图分解和节点聚类结果。(3) 它满足 HEDMoL 在子结构分解中的约束 A = $S_1 \cup ... \cup S_K$。因此，HEDMoL 基于连接树算法，将输入的原子级分子结构 A 分解为一组原子级子结构 R = {S_1, ..., S_K}，其中 S_k 定义为 A 的整个连接树中的一个顶点团。

3.4 知识扩展

为避免电子结构计算中的计算瓶颈，HEDMoL 的知识扩展步骤旨在从外部计算数据库 D_s 中，将易于获取的电子级信息传递给每个分解后的子结构 S_k ∈ R，而不是计算电子结构。也就是说，分解结果 (R，U$_e$) 通过知识扩展步骤转换为电子衍生子结构图 G_e。在 HEDMoL 的实现中，我们使用 QM9 数据集 [40]作为知识扩展步骤的外部计算数据库。

在知识扩展步骤中，HEDMoL 通过计算分子距离，将每个 S_k 与 D_s 中的一个小分子进行匹配。然后，HEDMoL 将所选小分子的电子级属性赋给 S_k的特征向量。形式上，针对 S_k 和 D_s 的信息检索过程可表示为：

$$\mathbf{X}_{e,k} = \mathbf{q}_{\mathrm{idx}_k}, \tag{3}$$

其中，$\mathbf{X}_{e,\,k}$ 是$\mathbf{X}_e$的第 k行向量。$\mathbf{q}_{\mathrm{idx}_k} \in \mathbb{R}^q$ 是 D_s 中某个小分子的预计算电子级属性，其中 idx_k 是 D_s 中所选小分子的索引，该分子与 S_k最相似，idx_k 的正式计算方式如下：

$$\mathrm{idx}_k = \underset{i\in\{1,2,\ldots,|\mathcal{D}_s|\}}{\arg\min} \; ||\pi(S_k) - \pi(G_{s,i})||_2, \tag{4}$$

其中，$G_{s,\,i}$ 是 D_s 中第 i 个分子的原子级图，π是两个分子图之间的距离度量。我们基于一种名为图上的几何散射（GeoScattering）的稳健高效无监督图嵌入方法实现了π [16]。HEDMoL 对 π 不同实现方式的预测能力在附录4.6中进行了实验评估。

知识扩展步骤之后，通过收集$\mathbf{X}_{e,\,k}$来构建$\mathbf{X}_e$，从而生成输入分子的 $G_e = (\mathrm{R}, \mathrm{U}_e, \mathbf{X}_e)$。因此，HEDMoL 生成 G_e，其中包含输入分子分解子结构的电子级信息，且无需通过昂贵的量子力学计算来优化电子结构。在下一步中，G_e将用于电子级和原子级的层次化表示学习。

3.5 层次表示学习 ning

HED-MoL 中层次表示学习步骤的目的是根据 G_a 和 G_e 生成潜在嵌入向量 $\mathbf{z}$，分别在原子级和电子级表示分子结构。HEDMoL 的层次表示学习遵循量子力学计算过程中依据基本电子密度 [15, 38] 计算分子性质的物理原理。

对于层次表示学习，HEDMoL 采用图神经网络来实现 ψ_a和 ψ_e，从而在图结构数据上进行排列不变表示学习。原子级和电子级节点嵌入矩阵 $\mathbf{H}_{\psi_a\ G_a}$ $_{e} = \psi_e\ G_e$ $_{a}$ $_{e}$ $_{a}$ = $_{e}$ 和 $\mathbf{H}$ 分别通过 $\mathbf{H}(\,)$ 和 $\mathbf{H}(\,)$ 计算得到。对于计算得到的 $\mathbf{H}$，电子态向量 $\mathbf{z}_e$ 通过 $\mathbf{z}_e = \sum_{i=1}^{|\mathcal{R}|} \mathbf{H}_{e,i}$ 计算得到，其中 $\mathbf{H}_{e,i}$ 是 $\mathbf{H}_e$ 的第 i 行向量。在对G_e进行嵌入后，通过在给定电子态向量 $\mathbf{z}_e$ 下考虑原子贡献，计算得到受电子级信息调控的分子嵌入 $\mathbf{z}_c$，如下所示：

$$\mathbf{z}_c = \sum_{i=1}^{|\mathcal{V}|} \frac{\exp(f_a(\mathbf{H}_{a,i} \oplus \mathbf{z}_e))}{\sum_{j=1}^{|\mathcal{V}|} \exp(f_a(\mathbf{H}_{a,j} \oplus \mathbf{z}_e))} \mathbf{H}_{a,i}, \tag{5}$$

其中，f_a 是一个可训练神经网络，用于计算在$\mathbf{z}_e$条件下第 i个原子的电子感知注意力系数 α_i。因此，HEDMoL 无需进行昂贵的量子力学计算，即可计算出受输入分子底层电子密度条件化的分子表征$\mathbf{z}_c$。

HEDMoL 将不同视图的图级嵌入拼接起来，生成输入分子的分子嵌入$\mathbf{z}$ ，即$\mathbf{z}$ =$\mathbf{z}_a \oplus \mathbf{z}_c$，其中$\mathbf{z}_a = \sum_{i=1}^{|\mathcal{V}|} \mathbf{H}_{a,\,i}$是 G_a 的图级嵌入向量。从概念上说，分子嵌入$\mathbf{z}$同时包含原子级和电子级分子表征。最后，HEDMoL 将$\mathbf{z}$输入可训练的全连接层 f_d，以预测目标分子性质 $y = f_d(\mathbf{z})$。

3.6 表示学习中的基于能量的物理一致性正则化

在 HEDMoL 的分子表示学习中，我们要求输入分子的潜在原子级和电子级嵌入表示相同的势能。势能是描述原子系统的一种普适物理量 [5, 38, 44]。从物理角度看，分子的原子级和电子级描述符应具有相同的势能，因为它们本质上源自相同的电子密度。为实现稳健的分子表示学习，我们在 HEDMoL 的训练过程中引入以下两项约束。

$$E_{p,k} + \epsilon_k = E_{a,k} = E_{e,k}, \forall k = 1, 2, ..., |\mathcal{R}|, \tag{6}$$

其中，$E_{p,\,k}$ 是与第 k个分解子结构 S_k相对应的小分子的计算物理能量，$\epsilon_k \sim \mathrm{N}(0,\ \sigma_k)$ 是一个服从正态分布 $\mathrm{N}(0,\ \sigma_k)$ 的独立同分布正态随机变量，$E_{a,\,k}$ 是根据 S_k中各节点的节点嵌入$\mathbf{H}_a$ 预测的能量，$E_{e,\,k}$是根据第 k个子结构嵌入$\mathbf{H}_{e,\,k}$预测的能量。在式（6）中，ϵ_k表示将 D_s 中小分子的物理能量迁移至 R 中分解子结构时产生的近似误差。

从物理角度来看，原子系统 A 的能量可表示为 A 中各原子的多体势能之和：[9]

$$E = \sum_{i_1,i_2}^{|\mathcal{A}|} V_2(\mathcal{H}_{i_1}, \mathcal{H}_{i_2}) + \sum_{i_1,i_2,i_3}^{|\mathcal{A}|} V_3(\mathcal{H}_{i_1}, \mathcal{H}_{i_2}, \mathcal{H}_{i_3}) + \cdots + \sum_{i_1,\ldots,i_{|\mathcal{A}|}}^{|\mathcal{A}|} V_{|\mathcal{A}|}(\mathcal{H}_{i_1}, ..., \mathcal{H}_{i_{|\mathcal{A}|}}), \tag{7}$$

其中，$|\mathrm{A}|$ 是原子系统 A 中原子的数量，V_i 是i体势函数，而 H_i 是第 i个原子周围的局部环境。然而，由于计算复杂度高得难以承受，计算多体势能并不切实际。在机器学习领域，各种实验表明，消息传递机制是一种根据粒子的物理相互作用预测物理性质的高效方法 [34, 65]。为克服难以承受的计算复杂度，我们遵循先前的研究，使用可训练的消息传递函数来近似式（7）中的多体势能。

我们使用图自注意力机制 [54]，基于 S_k中的原子间注意力得分，计算 S_k的多体势能。在形式上，在 HEDMoL 的训练过程中，S_k 的多体势能基于可训练能量函数 f_e 和消息传递函数 g_e 计算如下：

$$E_{a,k} = f_e(g_e(S_k)), \tag{8}$$

其中，原子级子结构嵌入 $g_e(S_k)$ 定义为

$$g_e(S_k) = \frac{1}{|S_k|} \sum_{i \in S_k} \left(\mathbf{WH}_{a,i} + \sum_{j \in N_i \cap S_k} \beta_{i,j} \mathbf{VH}_{a,j} \right), \tag{9}$$

$|S_k|$ 是子结构 S_k中的原子数，$\mathbf{W}$ 和$\mathbf{V}$ 是 g_e的可训练权重矩阵，$\mathbf{N}_i$ 是与第 i个原子相连的原子的索引集合，而$\beta_{i,j}$ 是第 i个原子与第 j个原子之间的注意力分数。同样，从电子级信息中获取的 S_k的多体势能$E_{e,k}$ 由可训练能量函数 f_e 计算如下：

$$E_{e,k} = f_e(\mathbf{H}_{e,k}) \tag{10}$$

根据式（8）和式（10），为了实现 HEDMoL 的稳健分子表示学习，我们为训练数据集中第 n个分子定义两个正则化项$\Omega_{a,n}$和$\Omega_{e,n}$，如下所示：

$$\Omega_{a,n} = \sum_{k=1}^{|\mathcal{R}_n|} \max\{|E_{p,k} - f_e(g_e(S_k))| - \alpha, 0\}, \tag{11}$$

$$\Omega_{e,n} = \sum_{k=1}^{|\mathcal{R}_n|} \max\{|E_{p,k} - f_e(\mathbf{H}_{e,k}))| - \alpha, 0\}, \tag{12}$$

其中，R_n 是第 n个分子的分解子结构集合，$\alpha \geq 0$ 是一个超参数，用于容许由分解子结构与外部计算数据库中匹配的小分子之间的结构差异所引起的不可控能量差异。超参数 α 是一个受物理边界约束的变量 [48],，小型有机分子之间的能量差异通常在 0.1 至 0.3 电子伏特[15, 48]的范围内。我们在第 4.5 节分析了 α取不同值时的超参数影响。最后，我们优化 HEDMoL 的模型参数，以最小化以下损失函数 L：

$$L = \sum_{n=1}^{|\mathcal{D}|} L_p(y_n, f_d(\mathbf{z}_n)) + \lambda \left(\sum_{n=1}^{|\mathcal{D}|} \Omega_{a,n} + \Omega_{e,n} \right), \tag{13}$$

其中，L_p 是预测损失，$\lambda \geq 0$ 是一个超参数，用于控制在 HEDMoL 的训练过程中正则化项 $\Omega_{a、n}$ 和 $\Omega_{e、n}$ 的影响。第 4.5 节还提供了不同 λ 取值下的超参数分析结果。

4 实验

为了评估 HEDMoL 的预测能力，我们将 HEDMoL 在包含实验观测分子物理性质的各种基准分子数据集上的预测准确率，与最先进方法的预测准确率进行了比较。在实验中，我们重点评估 HEDMoL 在实验数据集上的预测能力，而非计算数据集，原因有以下两点：(1) 在计算数据集（例如 QM9 数据集）上的评估并不公平，因为 HEDMoL 在知识扩展步骤中利用了外部计算数据库。(2) 与计算数据集相比，包含原子系统不确定性的实验数据集更接近现实世界分子物理性质 [70]。

数据集。我们从 MoleculeNet [61] 和 ChEMBL [32]等公共化学数据库中收集了通过实验生成的分子数据集。为展示 HEDMoL 在广泛科学应用中的有效性，我们选取了八个基准数据集

表1：包含原子级分子结构及其实验观测目标性质的基准分子数据集的特征。

应用类别	数据集	目标分子性质	# of Data
物理化学	Lipop [32]	亲脂性	4,200
	ESOL [12]	水溶性	1,128
	吸收、分布、代谢、排泄和毒性 [61]	水溶性	4,801
毒性	IGC50 [59]	梨形四膜虫毒性	1,791
	LC50 [59]	黑头呆鱼毒性	822
	LD50 [59]	大鼠经口毒性	7,412
药代动力学	LMC-H [32]	人微粒体清除率	5,347
	LMC-R [32]	大鼠微粒体清除率	2,165

这些数据集涵盖物理化学、毒性和药代动力学等多个应用领域。表 1 展示了这些基准分子数据集的统计信息和目标分子性质。

竞争方法。我们将HEDMoL的预测能力与基于树的集成方法 [8]以及最先进的图神经网络 [3, 6, 7, 14, 17, 19, 45, 47, 52, 62, 64, 66, 67]进行了比较。尽管三维图神经网络不适用于实验分子数据集，因为三维原子坐标不可用，我们仍通过基于力场和半经验计算 [41]生成三维分子结构来衡量其预测性能。需要注意的是，基于力场和半经验的计算方法效率高但不准确。此外，我们构建了XGB-Mor、XGB-FC和XGB-MK，它们分别针对输入的Morgan（Mor） [42],功能类（FC） [42], 和MACCS键（MK） [49] 指纹预测目标分子性质。附录A简要描述了所比较的方法。

实现。我们将分子结构转换为属性图 $G = (\mathbf{V}, \mathbf{U}, \mathbf{X}, \mathbf{E})$，其中 V 是节点（即原子）的集合，U 是边（即化学键）的集合，$\mathbf{X} \in \mathbb{R}^{|\mathcal{V}| \times d}$是一个 d维节点特征矩阵，而$\mathbf{E} \in \mathbb{R}^{|\mathcal{U}| \times l}$ 是一个 l维边特征矩阵。我们使用预定义的 200 维原子嵌入 [20] ，其原子特征由局部分子环境确定，用于定义输入节点特征矩阵 $\mathbf{X}$。为指定边特征，我们采用了常见的实现方式，将原子间的边特征定义为知名 RDKit 库[1]提供的成键类型的 22 维独热编码。我们使用 PyTorch Geometric[2] 提供的连接树算法来实现 HEDMoL 中的子结构分解步骤。

我们采用网格搜索来优化竞争对手图神经网络和 HEDMoL 的训练超参数，例如批大小和权重正则化系数。对于所有数据集，HEDMoL 训练时的初始学习率均固定为 5e-4。在 HEDMoL 的实现中，对于所有数据集的电子衍生子结构图，我们都使用 GIN 作为基于GNN的嵌入网络。但对于原子级分子图，我们使用 EGC 或 SchNet 作为基于GNN的嵌入网络。所有竞争对手图神经网络均由以下层堆叠构成：一个用于节点特征嵌入的全连接层、两个带层归一化的节点聚合层 [63],，以及两个用于预测的全连接层。所有竞争对手图神经网络的隐藏通道数均固定为 256。

[1]https://www.rdkit.org [2] https://pytorch-geometric.readthedocs.io

表 2：在基准分子数据集上测得的 R^2-分数。输入类型指输入分子所需的数据格式。每个基准数据集的最高 R^2-分数以粗体标出，R^2-分数的标准差列于括号中。N/R 和 N/A 分别表示负的 R^2-分数（表明回归失败）以及与内存溢出或数值稳定性问题相关的执行失败。

输入类型	方法	Lipop	ESOL	吸收、分布、代谢、排泄和毒性	IGC50	LC50	LD50	LMC-H	LMC-R
分子指纹	XGB-Mor [42]	0.531 (0.024)	0.659 (0.045)	0.717 (0.021)	0.621 (0.040)	0.390 (0.133)	0.497（0.016）	0.505（0.018）	**0.617（0.058）**
	XGB-FC [42]	0.578（0.018）	0.686（0.052）	0.720 (0.009)	0.628 (0.023)	0.501 (0.052)	0.519 (0.025)	0.503 (0.007)	**0.612 (0.015)**
	XGB-MK [49]	0.542 (0.041)	0.764 (0.047)	0.761 (0.020)	0.680 (0.037)	0.486 (0.112)	0.526 (0.021)	0.471 (0.019)	0.591 (0.033)
三维图	SchNet [45]	0.667 (0.021)	0.881 (0.026)	0.834 (0.012)	0.765 (0.034)	0.467 (0.025)	0.527 (0.062)	0.456 (0.024)	0.573 (0.043)
	DimeNet [17]	N/R	0.878 (0.025)	N/R	0.779 (0.019)	N/A	0.541 (0.045)	0.352 (0.101)	N/A
	PhysChem [67]	0.694 (0.024)	0.848 (0.032)	N/A	0.814 (0.017)	N/A	0.511 (0.053)	N/A	N/A
	M3GNet [6]	N/A	0.857 (0.025)	N/A	0.697 (0.029)	N/A	0.531 (0.034)	N/A	N/A
	FAENet [14]	0.670 (0.036)	0.869 (0.013)	0.788 (0.020)	0.708 (0.015)	0.528 (0.094)	0.474 (0.020)	0.437 (0.025)	0.528 (0.035)
二维图	GATv2 [3]	0.677 (0.053)	0.891 (0.020)	0.828 (0.014)	0.795 (0.013)	0.502 (0.063)	0.498 (0.030)	0.424 (0.027)	0.560 (0.036)
	GIN [64]	0.702 (0.031)	0.897 (0.022)	0.833 (0.017)	0.799 (0.021)	0.543 (0.080)	0.515 (0.044)	0.443 (0.027)	0.568 (0.020)
	EGC [52]	0.708 (0.043)	0.896 (0.017)	0.838 (0.012)	0.808 (0.029)	0.575 (0.045)	0.497 (0.034)	0.441 (0.023)	0.566 (0.017)
	消息传递神经网络 [19]	0.711 (0.022)	0.894 (0.023)	0.830 (0.014)	0.797 (0.018)	0.532 (0.064)	0.469 (0.040)	0.449 (0.057)	0.564 (0.031)
	UniMP [47]	0.702 (0.030)	0.886 (0.025)	0.833 (0.014)	0.793 (0.027)	0.504 (0.031)	0.470 (0.025)	0.422 (0.061)	0.579 (0.036)
	FiLM [2]	0.703 (0.048)	0.894 (0.031)	0.836 (0.014)	0.783 (0.046)	0.526 (0.042)	0.475 (0.032)	0.421 (0.050)	0.568 (0.032)
	MEGNet [7]	0.604 (0.023)	0.889 (0.027)	0.826 (0.038)	0.754 (0.026)	0.574 (0.122)	0.505 (0.027)	0.422 (0.032)	0.607 (0.041)
	DMPNN [66]	0.716 (0.037)	0.879 (0.013)	0.820 (0.018)	0.787 (0.008)	0.566 (0.098)	0.521 (0.011)	0.494 (0.011)	0.605 (0.043)
	AttFP [62]	0.710 (0.021)	**0.909 (0.018)**	0.841 (0.017)	0.807 (0.013)	**0.642 (0.079)**	0.513 (0.016)	0.456 (0.031)	0.588 (0.032)
	HEDMoL	**0.759 (0.043)**	**0.914 (0.016)**	**0.865 (0.014)**	**0.840 (0.010)**	**0.663 (0.053)**	**0.572 (0.035)**	**0.551 (0.008)**	**0.639 (0.035)**

表 3：各基准分子数据集上 HEDMoL 的超参数设置。B 和 γ 分别为随机梯度下降优化器的批大小和权重正则化系数。l_a 和 l_e 表示原子级分子图 G_a 和电子衍生子结构图 G_e的图嵌入维度。

数据集	ψ_a	ψ_e	B	γ	l_a	l_e	α	λ
Lipop	EGC	GIN	64	5e-6	32	32	0.2	1.0
ESOL	EGC	GIN	64	5e-6	16	16	0.2	1.0
吸收、分布、代谢、排泄和毒性	EGC	GIN	64	5e-6	16	16	0.2	1.0
IGC50	EGC	GIN	32	5e-6	16	16	0.2	0.4
LC50	SchNet	GIN	32	5e-6	16	16	0.1	0.2
LD50	SchNet	GIN	64	1e-6	16	16	0.2	1.0
LMC-H	SchNet	GIN	128	5e-6	16	16	0.2	1.0
LMC-R	SchNet	GIN	32	5e-6	32	32	0.2	1.0

HEDMoL 使用了 GIN、EGC 和 SchNet，其架构与竞争对手图神经网络相同，但为了在模型参数数量相近的情况下进行公平比较，节点聚合层与全连接层之间的隐藏通道数固定为 128。表 3 展示了 HEDMoL 在各数据集上的超参数设置。所有实验均在配备 Intel i9-12900K CPU、128G 内存和 NVIDIA GeForce RTX 3090 Ti 的机器上进行。

4.1 实验采集基准数据集上的预测

我们基于五折交叉验证测量了最先进的竞争方法和 HED-MoL 的 R^2分数，以确保测试数据集涵盖原始数据集中的所有分子，因为预测模型在真实世界化学数据上的预测准确率对训练/测试集划分 [39]十分敏感。我们报告了测试集上测得的 R^2分数的均值及标准差

数据集。表 2 展示了八个基准分子数据集上测得的 R^2分数和标准差，HEDMoL 在所有基准数据集上均取得了最高的 R^2分数。HED-MoL 在 Lipop、吸收、分布、代谢、排泄和毒性、IGC50、LC50、半数致死剂量、LMC-H 和 LMC-R 数据集上显著优于竞争方法。特别是，HEDMoL 的 R^2分数高于单独的 EGC 和 GIN，它们被用作 HED-MoL 的嵌入网络。附录 D 所示的平均绝对误差实验结果与表 2 一致。实验结果表明，通过在 G_a 和 G_e上进行分层表征学习，HEDMoL 学到了比单独使用图神经网络更具泛化性且信息更丰富的分子表征。

在实验中，图神经网络在 Lipop、ESOL 和吸收、分布、代谢、排泄和毒性数据集上预测物理化学性质的任务中，表现优于基于 XGB 的模型。然而，在包含相对较大分子的 LMC-H 和 LMC-R 数据集上，简单的基于 XGB 的模型取得了高于最先进图神经网络的 R^2分数。这一结果与现有研究 [30] 中关于图神经网络在大型原子系统上存在过拟合问题的实验观察一致。如本实验所示，基于指纹的模型和图神经网络应用于分子数据时各有局限。基于指纹的方法无法获取输入分子的物理信息，因为分子指纹旨在描述原子之间的连接关系，而非表征每个原子的物理属性。相比之下，尽管图神经网络能够从包含每个原子物理属性的输入分子图中提取物理信息，但在大型原子系统中很容易过拟合 [30]。然而，HEDMoL 利用带有电子属性的分子图克服了这两种局限，而这些电子属性具有良好的外推能力 [10, 15, 38]。这些结果表明，HEDMoL 能够提供更准确且

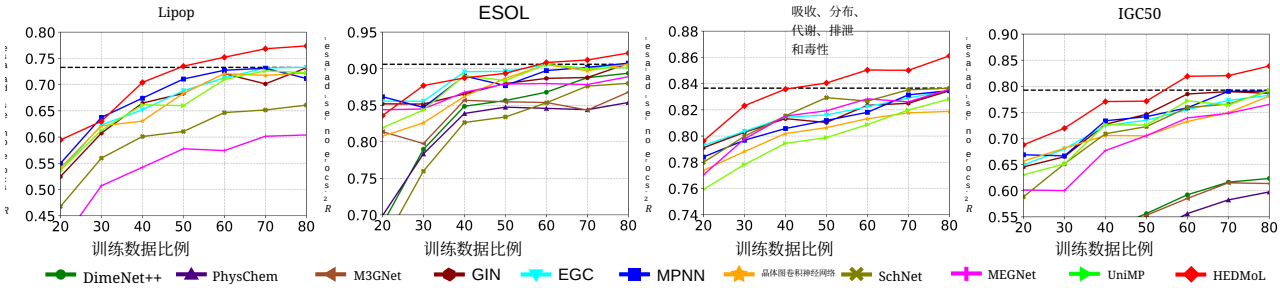


图3：不同训练数据量下的 R^2分数。黑色虚线表示最佳竞争方法在80%训练数据上的 R^2分数。

稳健的现实世界分子物理性质预测结果，且无需进行昂贵的量子力学计算。

4.2 小规模训练数据集上的预测

由于开展化学实验以获取实验标注数据既耗时又费力，训练数据的缺乏是化学应用中机器学习面临的主要挑战之一 [4, 23]。尽管人们提出了各种方法来克服化学应用中训练数据不足的问题，但现有方法仍然需要人工设计的领域知识 [24], 、对数据分布的严格假设 [29], 或密度泛函理论计算 [13]。另一方面，如第3.4节的知识扩展步骤所述，HEDMoL 本身具备将有关小分子的电子层级知识扩展到未见过的大分子的能力，这有利于基于小规模训练数据集构建准确的预测模型。在本实验中，我们比较了不同训练数据集规模下竞争方法与 HEDMoL 的 R^2-分数，以评估 HEDMoL 在小规模训练数据集上的预测能力。

图3展示了不同训练数据集规模下的 R^2-分数。我们在 Lipop、ESOL、吸收、分布、代谢、排泄和毒性以及 IGC50 数据集上测量了 R^2-分数，在这些数据集中，HEDMoL 和大多数竞争图神经网络取得了大于 0.6 的 R^2-分数。由于基于 XGB 的模型在小规模训练数据集上表现失败，我们没有测量其 R^2-分数。显然，我们观察到，随着训练数据集规模的增加，预测准确率趋于提高。然而，在 Lipop、吸收、分布、代谢、排泄和毒性以及 IGC50 数据集的所有训练数据集规模下，HEDMoL 的 R^2-分数均高于竞争图神经网络。此外，在 Lipop、吸收、分布、代谢、排泄和毒性以及 IGC50 数据集分别使用 50%、40% 和 60% 的训练数据时，HEDMoL 已经取得了与各自最佳竞争方法相当的 R^2-分数，即达到 80% 训练数据对应的水平（黑色虚线）。这些小规模训练数据集上的结果表明，HEDMoL 在真实世界化学应用中具有实际潜力，而这些应用通常受到训练数据不足的困扰。

4.3 不同外部计算数据库覆盖范围下的表征能力

如第3.4节所述，HEDMoL 本质上是一种表示学习方法，基于不同分子尺度计算所得信息之间的知识迁移。在本实验中，我们评估了 HEDMoL 对包含不同分子尺度分子的外部计算数据库的预测能力。为此，

表 4：HEDMoL 在不同分子尺度外部数据集上的测试 R^2分数。C 表示 QM9 数据集子集中分子的最大原子数。HEDMoL 的知识扩展是通过使用 $C = 6$的 QM9 子集实现的。

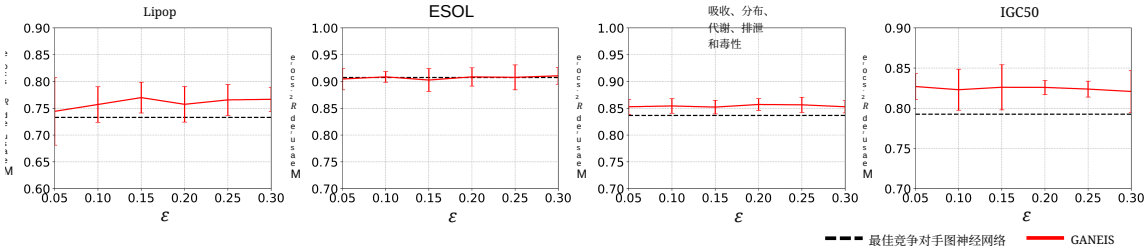
数据集	$C = 3$	$C = 4$	$C = 5$	$C = 6$	$C = 7$	$C = 8$
Lipop	0.732 (0.037)	0.723 (0.053)	0.735 (0.047)	0.736（0.030）	0.738（0.040）	0.736（0.037）
ESOL	0.914（0.018）	0.911（0.016）	0.915（0.015）	0.915 (0.016)	0.916 (0.015)	0.914 (0.019)
吸收、分布、代谢、排泄和毒性	0.867 (0.007)	0.862 (0.015)	0.868 (0.012)	0.861 (0.010)	0.867 (0.012)	0.865 (0.014)
IGC50	0.829 (0.017)	0.833 (0.012)	0.828 (0.016)	0.835 (0.017)	0.825 (0.018)	0.831 (0.016)

我们根据原子数量从 QM9 数据集中筛选分子，生成了不同分子尺度的外部计算数据库。在本实验中，我们生成了六个外部计算数据库，其中的分子由三个至 C 个原子组成，分别为 $C = \{3、4、5、6、7、8\}$。请注意，HEDMoL 的知识扩展通过 $C = 6$ 实现。表 4 展示了 HEDMoL 在不同分子尺度的外部计算数据库上的测得 R^2-分数。

在本实验中，尽管 $C = \{7, 8\}$ 的 QM9 子集包含更多分子，但 HEDMoL 的 R^2分数对于不同分子尺度下外部数据集的变化仍表现稳健，因为 HEDMoL 知识扩展中的连接树算法可以将输入的大分子分解为微小的子结构。因此，QM9 的 $C = \{3、4、5、6\}$ 子集中的小分子足以近似大多数分解后的子结构。同样，对于 $C = \{7, 8\}$，R^2分数的变化也很小，因为即使在包含较大分子的子集中，大多数分解后的子结构也会被含有 3–6 个原子的小分子取代。这一实验结果表明，为 HEDMoL 准备外部数据集并不困难，因为只需包含 3–6 个原子的小分子的外部数据集，就足以训练 HEDMoL。

4.4 消融研究

在本实验中，我们对 HEDMoL 进行了消融研究，以考察各个组件对 HEDMoL 预测能力的影响。由于移除 HEDMoL 中所有提出的组件后，它分别等同于嵌入网络 ψ_a或 ψ_e，因此我们比较了 EGC 和 GIN 的 R^2-分数。表 5 展示了 HEDMoL 的实验结果。为实现不含 HRL 的 KE，我们使用 $\mathbf{z} = \mathbf{z}_a \oplus \mathbf{z}_e$ 作为最终的分子嵌入，而非 $\mathbf{z} = \mathbf{z}_a \oplus \mathbf{z}_c$。我们观察到，



HEDMoL 在 Lipop、ESOL、吸收、分布、代谢、排泄和毒性及 IGC50 数据集上不同 α 值对应的 R^2-分数

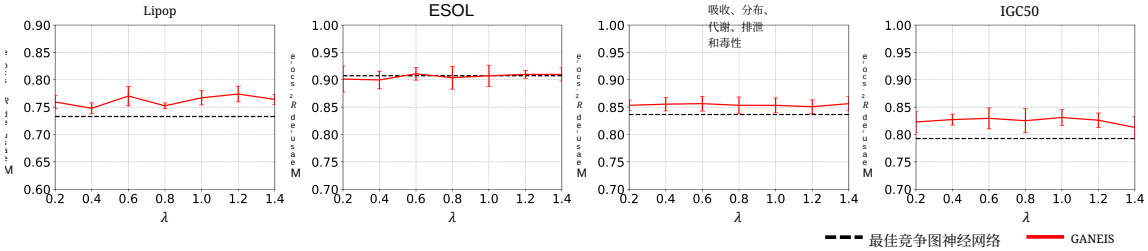


图 5： R HEDMoL 在 Lipop、ESOL、吸收、分布、代谢、排泄和毒性及 IGC50 数据集上不同取值对应的 ^{2}-分数 λ

HEDMoL 的消融研究。KE：第3.4节中的知识扩展。HRL：第3.5节中的分层分子表示学习。PCR：第3.6节中的基于能量的物理一致性正则化。

数据集	EGC	GIN	KE	知识扩展+HRL	知识扩展+PCR	HEDMoL
Lipop	0.708 (0.043)	0.702 (0.031)	0.722 (0.051)	0.726 (0.049)	0.731 (0.044)	0.759 (0.043)
ESOL	0.896 (0.017)	0.897 (0.022)	0.913 (0.015)	0.913 (0.015)	0.915 (0.017)	0.914 (0.016)
吸收、分布、代谢、排泄和毒性	0.838 (0.012)	0.833 (0.017)	0.862 (0.014)	0.864 (0.015)	0.860 (0.013)	0.865 (0.014)
IGC50	0.808 (0.029)	0.799 (0.021)	0.827 (0.018)	0.833 (0.017)	0.826 (0.014)	0.840 (0.010)
LC50	0.575 (0.045)	0.543 (0.080)	0.631 (0.091)	0.640 (0.084)	0.603 (0.061)	0.663 (0.053)

如表 5 所示，在所有数据集上实施 KE 后，预测模型的+-分数显著提高。这一结果明确表明，我们的知识扩展方法能够将电子级属性传播至原子级分子结构，且效果显著。尽管 KE+HRL 和 KE+PCR 在基准数据集上的提升幅度有限，但在 Lipop、IGC50 和 LC50数据集上，HEDMoL（KE+HRL+PCR）的 R^2-分数显著提高。

4.5 超参数分析

在 HEDMoL 的训练过程中，有两个超参数 α 和 λ，分别用于控制结构近似误差以及物理一致性正则化的效果。在本实验中，我们在分子数量相对较多的 Lipop、吸收、分布、代谢、排泄和毒性以及 IGC50 数据集上，测量了 HEDMoL 在 α 和 λ 不同取值下的 R^2分数。我们选择了

$$\alpha \quad \lambda \quad 0, 0\,05, 0\,1, 0\,15, 0\,2, 0\,25,$$

的超参数值，分别位于 $\{\ldots\ldots\}$ 和 $\{\ldots\ldots\}$ 中。图和展示了不同超参数取值下测得的 R^2分数。在本实验中，对于所有超参数取值和基准数据集，HEDMoL 始终取得了优于最佳竞争图神经网络的 R^2分数。特别是，HEDMoL 的表现超过了

0 3, 0 35 0 2, 0 4, 0 6, 0 8, 1 0 4 5

最佳竞争图神经网络在吸收、分布、代谢、排泄和毒性以及 IGC50 数据集上的 R^2分数标准差，且对于 α 和 λ的所有取值均如此。这些实验结果表明，HEDMoL 对超参数取值具有稳健性。

4.6 知识扩展步骤中不同图嵌入方法的预测能力

在 HEDMoL 的知识扩展实现中， $\pi(S_k, G_i^s)$有多种选择。在本实验中，我们评估了 HEDMoL 在知识扩展步骤中采用不同图嵌入方法时的预测能力。此外，我们还评估了一种基于分子 Morgan 指纹的 Tanimoto 距离 [1] 来实现 $\pi(S_k, G_i^s)$的方法，这是一种在化学信息学中寻找相似分子的常用方法。该实验评估在 Lipop、ESOL、吸收、分布、代谢、排泄和毒性以及 IGC50 数据集上进行，其中作为对比的图神经网络和 HEDMoL 的 R2 分数均高于 0.7。如表6所示，采用图嵌入方法的 HEDMoL 在所有基准数据集上的 R^2-分数相近，因为 HEDMoL 已假设知识扩展中存在匹配误差，并通过高斯误差 ϵ_k对其进行了形式化。因此，HEDMoL 对知识扩展的实现方式具有鲁棒性。

4.7 训练/推理执行时间

HEDMoL 在训练和推理阶段不可避免地需要额外的计算开销，以使用两个嵌入网络 ψ_a和 ψ_e执行分层表征学习。此外，计算物理一致性的正则化项也会带来额外计算。我们比较了 HEDMoL 与 EGC、GIN、消息传递神经网络和 SchNet 在 LC50、Lipop 和半数致死剂量数据集上的执行时间。消息传递神经网络和 SchNet 在基准数据集上表现出最高的预测准确率。EGC、GIN 和 SchNet 被用作 HEDMoL 的嵌入网络。LC50、Lipop 和

表6：不同知识扩展实现方式下 HEDMoL 的 R^2分数。FEATURE-G、FGSD、Graph2Vec 和 GeoScattering 是无监督图嵌入方法。

数据集	Morgan [42]	FEATHER-G [43]	FGSD [55]	Graph2Vec [36]	GeoScattering [16]
Lipop	0.734 (0.044)	0.734 (0.042)	0.741 (0.038)	0.734 (0.047)	0.736 (0.030)
ESOL	0.909 (0.018)	0.912 (0.019)	0.914 (0.014)	0.910 (0.016)	0.915 (0.016)
吸收、分布、代谢、排泄和毒性	0.865 (0.011)	0.870 (0.010)	0.864 (0.013)	0.864 (0.010)	0.861 (0.010)
IGC50	0.796 (0.018)	0.835 (0.011)	0.837 (0.021)	0.831 (0.023)	0.835 (0.017)

表7：LC50、Lipop 和 LD50 数据集上的每轮执行时间（秒/轮）。LC50、Lipop 和 LD50 数据集分别包含 822、4,200 和 7,412 个分子。

数据集	训练					推理				
	EGC	GIN	MPNN	SchNet	HEDMoL	EGC	GIN	MPNN	SchNet	HEDMoL
LC50	0.087	0.077	0.091	0.172	0.229	0.014	0.013	0.014	0.022	0.035
Lipop	0.643	0.558	0.619	1.650	1.371	0.099	0.090	0.099	0.191	0.228
LD50	0.929	0.828	0.902	2.038	2.468	0.136	0.125	0.133	0.237	0.373

半数致死剂量数据集分别包含 822、4,200 和 7,412 个分子。该实验在一台配备 Intel i9-12900K CPU、128G 内存和 NVIDIA GeForce RTX 3090 Ti 的机器上进行。

表7展示了MPNN、SchNet和HEDMoL在训练和推理过程中的执行时间。由于HEDMoL包含两个基于GNN的嵌入网络，因此其在训练和推理过程中的执行时间更长。尽管如此，HEDMoL的执行时间呈线性增长。例如，如 $1.371 \approx 0.643 + 0.558$ 和 $0.228 \approx 0.099 + 0.090$所示，HEDMoL在Lipop数据集上的执行时间与EGC和GIN的执行时间之和相近。这些实验结果表明，HEDMoL的前向过程与现有图神经网络的前向过程同样高效。

5 结论

本文提出了HEDMoL，用于学习电子衍生分子表征，以提升对真实世界分子物理学的预测能力。HEDMoL通过将外部数据库中小分子预先计算的电子级信息迁移至大型输入分子，在无需额外计算成本的情况下学习电子衍生分子表征。基于电子衍生分子表征，HEDMoL在包含实验观测分子及其性质的大规模分子数据集上实现了最先进的预测准确率。此外，即使在缺乏训练数据的情况下，HEDMoL仍取得了更高的预测准确率，而训练数据不足正是化学机器学习面临的主要挑战之一。这些结果表明，HEDMoL在真实世界化学应用中具有实际应用潜力。

致谢

本研究得到韩国化学技术研究院（编号：KK2451-10）的支持。本研究还得到了韩国政府（科学技术信息通信部）资助的韩国国家研究基金会（NRF）项目（RS-2024-00335098）的支持。

参考文献

[1] Dávid Bajusz、Anita Rácz和Károly Héberger。2015年。为何Tanimoto指数是基于指纹的相似性计算的合适选择？《化学信息学杂志》7, 1 (2015), 1–13。[2] Marc Brockschmidt。2020年。GNN-FiLM：采用特征级线性调制的图神经网络。见*ICML*。PMLR，1144–1152。[3] Shaked Brody、Uri Alon和Eran Yahav。2021年。图注意力网络有多专注？*arXiv*预印本 *arXiv:2105.14491*(2021)。[4] Chenjing Cai、Shiwei Wang、Youjun Xu、Weilin Zhang、Ke Tang、Qi Ouyang、Luhua Lai和Jianfeng Pei。2020年。面向药物发现的迁移学习。《药物化学杂志》63, 16 (2020), 8683–8694。[5] Richard Car和Mark Parrinello。1985年。分子动力学与密度泛函理论的统一方法。《物理评论快报》55, 22 (1985), 2471。[6] Chi Chen和Shyue Ping Ong。2022年。面向元素周期表的通用图深度学习原子间势。《自然·计算科学》2, 11 (2022), 718–728。[7] Chi Chen、Weike Ye、Yunxing Zuo、Chen Zheng和Shyue Ping Ong。2019年。图网络：面向分子与晶体的通用机器学习框架。《材料化学》31, 9 (2019), 3564–3572。[8] Tianqi Chen和Carlos Guestrin。2016年。XGBoost：一种可扩展的树提升系统。见*ACM SIGKDD*。785–794。[9] Gerardo Andrés Cisneros、Kjartan Thor Wikfeldt、Lars Ojamae、Jibao Lu、Yao Xu、Hedieh Torabifard、Albert P Bartók、Gábor Csányi、Valeria Molinero和Francesco Paesani。2016年。水中分子相互作用建模：从成对势到多体势能函数。《化学评论》116, 13 (2016), 7501–7528。[10] Aron J Cohen、Paula Mori-Sánchez和Weitao Yang。2012年。密度泛函理论面临的挑战。《化学评论》112, 1 (2012), 289–320。[11] Christopher J Cramer。2013年。计算化学精要：理论与模型。John Wiley & Sons。[12] John S. Delaney。2004年。ESOL：直接从分子结构估算水溶性。《化学信息与计算机科学杂志》44 (2004), 1000–1005。[13] Hieu A Doan、Garvit Agarwal、Hai Qian、Michael J Counihan、Joaquín Rodríguez-López、Jeffrey S Moore和Rajeev S Assary。2020年。量子化学引导的主动学习，以加速可持续储能材料的设计与发现。《材料化学》32, 15 (2020), 6338–6346。[14] Alexandre Agm Duval、Victor Schmidt、Alex Hernandez-Garcia、Santiago Miret、Fragkiskos D Malliaros、Yoshua Bengio和David Rolnick。2023年。FAENet：用于材料建模的帧平均等变图神经网络。见*ICML*。PMLR，9013–9033。[15] Eberhard Engel。2011年。密度泛函理论。Springer。[16] Feng Gao、Guy Wolf和Matthew Hirn。2019年。面向图数据分析的几何散射。见*ICML*。PMLR，2122–2131。[17] Johannes Gasteiger、Shankari Giri、Johannes T Margraf和Stephan Günnemann。2020年。面向非平衡分子的快速且具备不确定性感知的方向性消息传递。*arXiv*预印本 *arXiv:2011.14115*(2020)。[18] Johannes Gasteiger、Janek Groß和Stephan Günnemann。2020年。面向分子图的方向性消息传递。见*ICLR*。[19] Justin Gilmer、Samuel S Schoenholz、Patrick F Riley、Oriol Vinyals和George E Dahl。2017年。面向量子化学的神经消息传递。见*ICML*。[20] Rhys EA Goodall和Alpha A Lee。2020年。在无晶体结构条件下预测材料性质：基于化学计量的深度表示学习。《自然·通讯》11, 1 (2020), 1–9。

[21] Piotr S Gromski、Alon B Henson、Jarosław M Granda和Leroy Cronin。2019年。如何利用算法与自动化探索化学空间。《自然评论：化学》3, 2 (2019), 119–128。[22] Anubhav Jain、Shyue Ping Ong、Geoffroy Hautier、Wei Chen、William Davidson Richards、Stephen Dacek、Shreyas Cholia、Dan Gunter、David Skinner、Gerbrand Ceder等。2013年。评论：材料项目：一种加速材料创新的材料基因组方法。《应用物理快报：材料》1, 1 (2013)。[23] Dipendra Jha、Kamal Choudhary、Francesca Tavazza、Wei-keng Liao、Alok Choudhary、Carelyn Campbell和Ankit Agrawal。2019年。利用深度迁移学习融合计算数据与实验数据以增强材料性质预测。《自然·通讯》10, 5316 (2019)。第1期。https://doi.org/10.1038/s41467- 019-13297-w[24] Dejun Jiang、Zhenxing Wu、Chang-Yu Hsieh、Guangyong Chen、Ben Liao、Zhe Wang、Chao Shen、Dongsheng Cao、Jian Wu和Tingjun Hou。2021年。图神经网络能否为药物发现学习更好的分子表征？基于描述符的模型与基于图的模型之比较研究。《化学信息学杂志》13, 1 (2021), 1–23。[25] Wengong Jin、Regina Barzilay和Tommi Jaakkola。2018年。用于分子图生成的连接树变分自编码器。见ICML。PMLR，2323–2332。[26] Seojin Kim、Jaehyun Nam、Junsu Kim、Hankook Lee、Sungsoo Ahn和Jinwoo Shin。2023年。基于片段的多视图分子对比学习。见ICLR“材料机器学习”研讨会。[27] Thomas N. Kipf和Max Welling。2017年。基于图卷积网络的半监督分类。ICLR (2017)。[28] Ondrej L Krivanek、Matthew F Chisholm、Valeria Nicolosi、Timothy J Pennycook、George J Corbin、Niklas Dellby、Matthew F Murfitt、Christopher S Own、Zoltan S Szilagyi、Mark P Oxley等。2010年。利用环形暗场电子显微镜进行逐原子结构与化学分析。《自然》464, 7288 (2010), 571–574。[29] Yea-Lee Lee、Hyungseok Lee、Taeshik Kim、Sejin Byun、Yong Kyu Lee、Seunghun Jang、In Chung、Hyunju Chang和Jino Im。2022年。数据驱动提升硒化锡基热电系统的热电优值。《美国化学会志》144, 30 (2022), 13748–13763。[30] Kangming Li、Brian DeCost、Kamal Choudhary、Michael Greenwood和Jason Hattrick-Simpers。2023年。对材料性质机器学习预测之稳健性与泛化能力的批判性审视。《自然合作期刊：计算材料》9, 1 (2023), 55。[31] Tairan Liu、Misagh Naderi、Chris Alvin、Supratik Mukhopadhyay和Michal Brylinski。2017年。先分解再构建：利用eMolFrag分解小分子以进行基于片段的药物设计。《化学信息与建模杂志》57, 4 (2017), 627–631。[32] David Mendez、Anna Gaulton、A Patrícia Bento、Jon Chambers、Marleen De Veij、Eloy Félix、María Paula Magariños、Juan F Mosquera、Prudence Mutowo、Michał Nowotka等。2019年。ChEMBL：迈向生物测定数据的直接存缴。《核酸研究》47, D1 (2019), D930–D940。[33] Grégoire Montavon、Matthias Rupp、Vivekanand Gobre、Alvaro Vazquez-Mayagoitia、Katja Hansen、Alexandre Tkatchenko、Klaus-Robert Müller和O Anatole Von Lilienfeld。2013年。化学化合物空间中分子电子性质的机器学习。《新物理学杂志》15, 9 (2013), 095003。[34] Gyoung S. Na和Chanyoung Park。2022年。用于神经网络外推的非线性编码。见ACM SIGKDD会议。计算机协会，1284–1294。[35] Maho Nakata和Tomomi Shimazaki。2017年。PubChemQC项目：面向数据驱动化学的大规模第一性原理电子结构数据库。《化学信息与建模杂志》57, 6 (2017), 1300–1308。[36] Annamalai Narayanan、Mahinthan Chandramohan、Rajasekar Venkatesan、Lihui Chen、Yang Liu和Shantanu Jaiswal。2017年。Graph2Vec：学习图的分布式表示。arXiv预印本 arXiv:1707.05005 (2017)。[37] Andrew Ng、Michael Jordan和Yair Weiss。2001年。论谱聚类：分析与一种算法。NeurIPS 14 (2001)。[38] Robert G Parr和Weitao Yang。1995年。分子电子结构的密度泛函理论。《物理化学年评》46, 1 (1995), 701–728。[39] Assima Rakhimbekova、Tagir N Akhmetshin、Guzel I Minibaeva、Ramil I Nugmanov、Timur R Gimadiev、Timur I Madzhidov、Igor I Baskin和Alexandre Varnek。2021年。化学反应定量构效关系建模中的交叉验证策略。《结构-活性关系与定量结构-活性关系环境研究》32, 3 (2021), 207–219。[40] Raghunathan Ramakrishnan、Pavlo O. Dral、Matthias Rupp和O. Anatole von Lilienfeld。2014年。13.4万种分子的量子化学结构与性质。《科学数据》1 (2014)。[41] Sereina Riniker和Gregory A Landrum。2015年。信息更充分的距离几何：利用已知信息改进构象生成。《化学信息与建模杂志》55, 12 (2015), 2562–2574。[42] David Rogers和Mathew Hahn。2010年。扩展连接指纹。《化学信息与建模杂志》50, 5 (2010), 742–754。[43] Benedek Rozemberczki和Rik Sarkar。2020年。图上的特征函数：物以类聚——从统计描述符到参数模型。见信息与知识管理国际会议。1325–1334。

[44] Matthieu Schapira、Maxim Totrov 和 Ruben Abagyan。1999年。小分子、肽和蛋白质结合能的预测。《分子识别杂志》12，3（1999），177–190。[45] Kristof Schütt、Pieter-Jan Kindermans、Huziel Enoc Sauceda Felix、Stefan Chmiela、Alexandre Tkatchenko 和 Klaus-Robert Müller。2017年。Schnet：一种用于量子相互作用建模的连续滤波卷积神经网络。NeurIPS 30（2017）。[46] Yuning Shen、Julia E Borowski、Melissa A Hardy、Richmond Sarpong、Abigail G Doyle 和 Tim Cernak。2021年。化学合成的自动化与计算机辅助规划。《自然综述：疾病引物》1，1（2021），1–23。[47] Yunsheng Shi、Zhengjie Huang、Shikun Feng、Hui Zhong、Wenjing Wang 和 Yu Sun。2021年。掩码标签预测：用于半监督分类的统一消息传递模型。发表于 IJCAI。[48] Eunji Sim、Suhwan Song 和 Kieron Burke。2018年。量化密度泛函理论中的密度误差。《物理化学快报杂志》9，22（2018），6385–6392。[49] Narender Singh、Rajarshi Guha、Marc A Giulianotti、Clemencia Pinilla、Richard A Houghten 和 Jose L Medina-Franco。2009年。组合库、药物、天然产物及分子库小分子资源库的化学信息学分析。《化学信息与建模杂志》49，4（2009），1010–1024。[50] Alain St-Amant 和 Dennis R Salahub。1990年。局域密度泛函理论中几何结构优化的新算法。《化学物理快报》169，5（1990），387–392。[51] Kazu Suenaga 和 Masanori Koshino。2010年。石墨烯边缘的逐原子光谱学。《自然》468，7327（2010），1088–1090。[52] Shyam A Tailor、Felix Opolka、Pietro Lio 和 Nicholas Donald Lane。2021年。我们需要各向异性图神经网络吗？发表于 ICLR。[53] Fabien Tran 和 Peter Blaha。2009年。使用半局域交换关联势精确预测半导体和绝缘体的带隙。《物理评论快报》102，22（2009），226401。[54] Petar Veličković、Guillem Cucurull、Arantxa Casanova、Adriana Romero、Pietro Lio 和 Yoshua Bengio。2020年。图注意力网络。发表于 ICLR。[55] Saurabh Verma 和 Zhi-Li Zhang。2017年。在图上寻找独特、稳定、稀疏且快速的特征学习方法。NeurIPS 30（2017）。[56] Hongwei Wang、Weijiang Li、Xiaomeng Jin、Kyunghyun Cho、Heng Ji、Jiawei Han 和 Martin D Burke。2022年。化学反应感知分子表征学习。ICLR（2022）。[57] Oliver Wieder、Stefan Kohlbacher、Mélaine Kuenemann、Arthur Garon、Pierre Ducrot、Thomas Seidel 和 Thierry Langer。2020年。图神经网络分子性质预测简明综述。《今日药物发现：技术》37（2020），1–12。[58] Daniel S Wigh、Jonathan M Goodman 和 Alexei A Lapkin。2022年。机器学习时代的分子表征综述。《Wiley跨学科综述：计算与分子科学》（2022），e1603。[59] Kedi Wu 和 Guo-Wei Wei。2018年。基于拓扑的多任务深度神经网络定量毒性预测。《化学信息与建模杂志》58，2（2018），520–531。[60] Zonghan Wu、Shirui Pan、Fengwen Chen、Guodong Long、Chengqi Zhang 和 S Yu Philip。2020年。图神经网络综合综述。《IEEE神经网络与学习系统汇刊》32，1（2020），4–24。[61] Zhenqin Wu、Bharath Ramsundar、Evan N Feinberg、Joseph Gomes、Caleb Geniesse、Aneesh S Pappu、Karl Leswing 和 Vijay Pande。2018年。MoleculeNet：分子机器学习基准。《化学科学》9，2（2018），513–530。[62] Zhaoping Xiong、Dingyan Wang、Xiaohong Liu、Feisheng Zhong、Xiaozhe Wan、Xutong Li、Zhaojun Li、Xiaomin Luo、Kaixian Chen、Hualiang Jiang 等。2019年。借助图注意力机制拓展药物发现分子表征的边界。《药物化学杂志》63，16（2019），8749–8760。[63] Jingjing Xu、Xu Sun、Zhiyuan Zhang、Guangxiang Zhao 和 Junyang Lin。2019年。理解并改进层归一化。NeurIPS 32（2019）。[64] Keyulu Xu、Weihua Hu、Jure Leskovec 和 Stefanie Jegelka。2018年。图神经网络的能力有多强？发表于 ICLR。[65] Keyulu Xu、Mozhi Zhang、Jingling Li、Simon Shaolei Du、Ken-Ichi Kawarabayashi 和 Stefanie Jegelka。2021年。神经网络如何外推：从前馈网络到图神经网络。发表于 ICLR。[66] Kevin Yang、Kyle Swanson、Wengong Jin、Connor Coley、Philipp Eiden、Hua Gao、Angel Guzman-Perez、Timothy Hopper、Brian Kelley、Miriam Mathea 等。2019年。分析用于性质预测的学习所得分子表征。《化学信息与建模杂志》59，8（2019），3370–3388。[67] Shuwen Yang、Ziyao Li、Guojie Song 和 Lingsheng Cai。2021年。融合物理与化学信息的深度分子表征学习。NeurIPS 34（2021），16346–16357。[68] Zhaoning Yu 和 Hongyang Gao。通过异质基序图神经网络进行分子表征学习。发表于 ICML。PMLR，25581–25594。[69] Zaixi Zhang、Qi Liu、Hao Wang、Chengqiang Lu 和 Chee-Kong Lee。2021年。用于分子性质预测的基于基序的图自监督学习。NeurIPS 34（2021），15870–15882。[70] Ya Zhuo、Aria Mansouri Tehrani 和 Jakoah Brgoch。2018年。通过机器学习预测无机固体的带隙。《物理化学快报杂志》9，7（2018），1668–1673。https://doi.org/10.1021/acs.jpclett.8b00124

一种竞争方法

在实验中，我们将 HEDMoL 的预测能力与一种基线树方法以及八种在化学应用中广泛使用的最先进图神经网络进行了比较。竞争方法简要介绍如下：

• **XGB-Mor**：XGBoost（XGB）是一种基于树的梯度提升模型，在各种科学应用中展现了最先进的性能。在实验评估中，我们构建了 XGB-Mor，使用原子级分子结构的 Morgan（Mor）指纹来预测目标分子性质。• **XGB-FC**：我们将 XGB 与输入分子的官能团类别（FC） [42] 指纹相结合，构建了 XGB-FC。FC 指纹根据分子的官能团子结构和原子来表示原子级分子结构。

• **XGB-MK**：我们还基于 MACCS 键（MK）指纹构建了 XGB-MK。MACCS 键是最常用的分子表征之一，它使用 166位二进制模式对原子级分子结构进行编码。• **SchNet** [45]：这是一种卷积神经网络，通过分子中的量子相互作用学习分子表示。它已被广泛用作各种化学应用中的基线模型 [45]。

• **DimeNet** [17]：DimeNet 及其变体采用方向性消息传递方案，以学习物理化学中的方向信息。我们使用了原始 DimeNet 的一个改进模型，称为 DimeNet++。
• **PhysChem** [67]：这是一种神经网络架构，通过融合原子间几何结构信息以及经由化学键进行消息传递的信息来学习分子表示。

• **M3GNet** [6]：三体相互作用图神经网络（M3GNet）是一种基于三体原子间相互作用来学习分子表示的神经网络，其建模超越了传统的单原子属性和原子间二体相互作用。

• **FAENet** [14]：帧平均等变图神经网络（FAENet）是一种简单快速、针对随机帧平均进行优化的图神经网络。FAENet 可以灵活地处理原子相对位置来学习分子表示，而无需满足保持对称性的要求。
• **GATv2** [3]：图注意力网络（GAT）旨在基于节点之间的自注意力机制学习节点和图嵌入 [54]。我们采用了第二版图注意力网络（GATv2），在全面的实验中，其性能优于标准 GAT。• **GIN** [64]：图同构网络（GIN）是一种基于图同构检验来学习图表示的框架。

• **EGC** [52]：高效图卷积（EGC）是一种各向同性图神经网络，在节点聚合阶段基于自适应滤波器和聚合融合。EGC 在基准数据集上的表现优于图注意力网络等常见的各向异性图神经网络。
• **MPNN** [19]：消息传递神经网络是用于在量子化学中学习分子表示的节点和边卷积方法的统一框架。• **UniMP** [47]：统一消息传递（UniMP）是一种基于Transformer的图神经网络。UniMP 基于Transformer架构，在训练和推理阶段同时引入特征传播和标签传播，展现出最先进的预测能力。

• **FiLM** [2]：特征级线性调制（FiLM）旨在有效地在图中传播信息，并对传递的信息进行逐特征调制。

• **MEGNet** [7]：材料图网络（MEGNet）旨在预测分子和晶体结构的物理和化学性质。它通过原子间的消息传递过程传播原子结构的全局状态，显著提高了预测准确率。• **DMPNN** [66]：有向消息传递神经网络（D-MPNN）是针对有向分子图对原始消息传递神经网络的扩展。DMPNN 通过有向边传递消息，以学习原子间的方向信息。

• **AttFP** [62]：AttFP（最初名为 AttentiveFP）采用图自注意力机制来学习类药分子的分子表示。AttFP 旨在学习非局部分子内相互作用，从而提取信息丰富的分子表示。

B 实现细节

在 HEDMoL 的实现中，我们从 QM9 数据集中选取由六个或更少原子组成的分子，得到包含 685 个分子的子集。这是因为我们旨在基于小分子进行知识扩展，因为小分子的电子层级计算误差较小。在知识扩展中，迁移的子结构特征 $\mathbf{X}_{e,\,k}$ 定义为一个子结构特征矩阵 $\mathbf{X}_e$，该矩阵是一个包含 12 维物理计算的电子属性的向量：偶极矩、各向同性极化率、最高占据分子轨道、最低未占分子轨道、HOMO-LUMO能隙、电子空间范围、振动能、0 K 和 298.15 K 下的内能、焓、自由能以及热容。

C HEDMoL架构采用连接树和BRICS分解算法的预测准确率

我们使用HEDMoL的一其中[子级]分解过程中的连接树算法被替换为BRICS分解 [31]。BRICS分解是化学科学中最常用的图分解方法，可将分子图拆分为化学有效子结构。

如表 8 所示，在大多数基准数据集上，HEDMoL 的预测准确率略高于 HEDMoL-BC。具体而言，HEDMoL 在 Lipop 和 LC50数据集上的表现优于 HEDMoL-BC。这一实验结果有两个原因：（1）BRICS 分解无法将大分子拆分成足够小的子结构，因为它必须确保分解后的子结构具有化学有效性

表 8：HEDMoL 和 HEDMoL-BC 的实测 R^2 分数。

方法	Lipop	ESOL	吸收、分布、代谢、排泄和毒性	IGC50	LC50	LD50	LMC-H	LMC-R
HEDMoL-BC	0.731（0.054）	0.915（0.021）	0.862 (0.007)	0.836 (0.016)	0.620 (0.081)	N/A	0.532 (0.013)	0.622 (0.039)
HEDMoL	**0.759 (0.043)**	0.914 (0.016)	0.865 (0.014)	0.840 (0.010)	**0.663 (0.053)**	**0.572 (0.035)**	0.551（0.008）	0.639（0.035）

。（2）为了确保分解后的子结构具有化学有效性，系统会在其中虚拟添加许多氢原子，这可能会扭曲原始分子的结构信息。此外，由于在 LD50 数据集中的若干大分子上执行 BRICS 分解所需的执行时间急剧增加，我们无法在 LD50 数据集上运行 HEDMoL-BC。这一实验结果表明，将分子结构作为属性图而非化学系统进行处理是有效的。

D 评估结果：平均绝对误差

表 9 展示了竞争方法和 HEDMoL 在基准分子数据集上的平均绝对误差（平均绝对误差）。与通过 R^2-得分测得的结果一样，HEDMoL 在所有基准数据集上都表现出最佳性能，即 HEDMoL 在所有基准数据集上都取得了最低的平均绝对误差。

表 9：在基准分子数据集上测得的 R^2-得分。输入类型表示输入分子所需的数据格式。每个基准数据集中的最高 R^2-得分以粗体标出，R^2-得分的标准差列于括号中。

输入类型	方法	Lipop	ESOL	吸收、分布、代谢、排泄和毒性	IGC50	LC50	LD50	LMC-H	LMC-R
分子指纹	XGB-Mor [42]	0.608 (0.012)	0.908 (0.039)	0.855 (0.018)	0.648 (0.017)	0.876 (0.061)	0.475 (0.008)	0.358 (0.006)	0.368 (0.027)
	XGB-FC [42]	0.586 (0.010)	0.874 (0.064)	0.855 (0.017)	0.468 (0.021)	0.776 (0.020)	0.484 (0.006)	0.357 (0.005)	0.365 (0.007)
	XGB-MK [49]	0.610 (0.014)	0.740 (0.047)	0.773 (0.017)	0.444 (0.012)	0.770 (0.139)	0.491 (0.015)	0.364 (0.004)	0.373 (0.018)
三维图	SchNet [45]	0.560 (0.013)	0.554 (0.042)	0.606 (0.020)	0.390 (0.022)	0.677 (0.019)	0.483 (0.010)	0.359 (0.007)	0.410 (0.016)
	DimeNet [17]	N/A	0.562 (0.045)	N/A	0.382 (0.021)	N/A	0.468 (0.009)	0.420 (0.011)	N/A
	PhysChem [67]	0.535 (0.014)	0.592 (0.041)	N/A	0.313 (0.017)	N/A	0.478 (0.015)	N/A	N/A
	M3GNet [6]	N/A	0.585 (0.039)	N/A	0.452 (0.021)	N/A	0.481 (0.008)	N/A	N/A
	FAENet [14]	0.558 (0.012)	0.565 (0.042)	0.725 (0.018)	0.443 (0.025)	0.665 (0.021)	0.493 (0.009)	0.378 (0.015)	0.447 (0.020)
二维图	GATv2 [3]	0.463 (0.016)	0.461 (0.021)	0.609 (0.016)	0.315 (0.013)	0.680 (0.024)	0.484 (0.008)	0.374 (0.006)	0.386 (0.009)
	GIN [64]	0.437 (0.024)	0.443 (0.040)	0.575 (0.017)	0.311 (0.012)	0.663 (0.038)	0.475 (0.016)	0.365 (0.007)	0.381 (0.008)
	EGC [52]	0.436 (0.016)	0.434 (0.016)	0.590 (0.012)	0.306 (0.022)	0.644 (0.039)	0.486 (0.005)	0.367 (0.005)	0.378 (0.009)
	消息传递神经网络 [19]	0.438 (0.012)	0.453 (0.017)	0.611 (0.011)	0.312 (0.011)	0.670 (0.053)	0.500 (0.014)	0.363 (0.011)	0.382 (0.014)
	UniMP [47]	0.449 (0.020)	0.472 (0.030)	0.599 (0.021)	0.320 (0.024)	0.682 (0.035)	0.497 (0.009)	0.373 (0.011)	0.376 (0.016)
	FiLM [2]	0.454 (0.018)	0.450 (0.038)	0.599 (0.014)	0.309 (0.023)	0.695 (0.043)	0.496 (0.013)	0.372 (0.014)	0.383 (0.018)
	MEGNet [7]	0.537 (0.019)	0.474 (0.024)	0.601 (0.025)	0.355 (0.018)	0.662 (0.054)	0.478 (0.011)	0.371 (0.008)	0.366 (0.010)
	DMPNN [66]	0.435 (0.017)	0.558 (0.038)	0.595 (0.021)	0.307 (0.019)	0.658 (0.043)	0.475 (0.008)	0.352 (0.015)	0.365 (0.020)
	AttFP [62]	0.435 (0.018)	**0.435 (0.021)**	0.581 (0.018)	0.303 (0.022)	**0.623 (0.038)**	0.478 (0.016)	0.357 (0.010)	0.382 (0.018)
	HEDMoL	**0.392 (0.017)**	**0.427 (0.027)**	**0.554 (0.026)**	**0.287 (0.007)**	**0.603 (0.036)**	**0.451（0.013）**	**0.335（0.004）**	**0.344（0.016）**