arXiv:2603.01274v1 [cs.LG] 1 Mar 2026

GlassMol：基于概念瓶颈模型的可解释分子性质预测

Oscar Rivera✉[1][0009-0005-5742-1740],Ziqing Wang[1][0009-0004-8940-0461] Matthieu Dagommer[2][0009-0006-3260-6455],Abhishek Pandey[2][0009-0008-7905-5712],和 Kaize Ding[1][0000-0001-6684-6752]

[1] 西北大学，埃文斯顿，IL 60201，美国
oscarrivera2028@u.northwestern.edu[2] 艾伯维，
美国伊利诺伊州芝加哥，IL 60064

摘要。 机器学习加速了分子性质预测，但最先进的大语言模型和图神经网络仍像黑箱一样运作。在药物发现领域，安全至关重要，这种不透明性可能掩盖虚假相关性，并排斥人类专业知识。现有可解释性方法面临**有效性与可信度** 之间的权衡：解释可能无法反映模型的真实推理过程、降低性能，或缺乏领域依据。概念瓶颈模型（CBM）通过在读出前将输入映射到人类可解释概念，为这一问题提供了解决方案，确保解释从根本上忠实反映决策过程。然而，将概念瓶颈模型应用于化学领域面临三项挑战：相关性差距（从庞大的描述符空间中选出任务相关概念）、标注差距（为分子数据获取概念监督）以及容量差距（瓶颈约束导致性能下降）。我们提出 GlassMol，这是一种模型无关的概念瓶颈模型，通过自动化概念整理和大语言模型引导的概念选择来弥合这些差距。在十三个基准测试上的实验表明，GlassMol 的表现通常与黑箱基线持平或更优，这说明可解释性并不以牺牲性能为代价，也对普遍认为二者存在权衡的观点提出了挑战。代码见 https://github.com/walleio/GlassMol。

1 引言

机器学习驱动的分子性质预测旨在利用计算的高通量特性，为湿实验室化学家识别有前景的研究方向，从而提高整个研究流程的效率 [10,22,38]。通过从穷尽实验测试成本高昂的庞大化学数据库中学习，机器学习模型优先筛选有前景的化合物以进行湿实验室验证，从而降低成本并缩短开发周期。这种计算方法已在发现新型抗生素、优化分子性质以及实现从头药物设计方面展现出切实影响 [10,22,41,2]。

最成功的方法包括图神经网络（GNN） [37,19,4,7]和大语言模型（LLM） [42,43]。从设计上看，图神经网络依赖图结构数据，能够在节点（即原子）之间进行消息传递，因此非常适合处理本质上具有图结构的分子信息 [6]。尽管大语言模型无法利用图结构，但在自然语言分子任务中，它们展现出了强大的编码和推理能力 [48]。这两种范式都具有出色的预测

性能，但二者都是黑箱：图神经网络生成难以解释的高维图嵌入，而大语言模型则依赖数十亿个不透明参数运行 [47,46]。在药物发现领域，监管批准和患者安全要求提供严格的论证，因此这种不透明性带来了重大挑战。

对可解释性的需求不仅限于满足监管要求，还延伸到科学有效性的根本问题。如果不了解模型预测毒性的原因，研究人员就无法验证这些预测是否反映了真实的构效关系，还是仅仅源于虚假相关性。模型可能会因为与实际机制无关的原因，正确地将某个分子标记为具有肝毒性，这使得研究人员难以据此指导先导化合物优化。现有的可解释性方法，包括注意力可视化、基于梯度的归因和子图探索，试图弥补这一空白 [46,44]。然而，这些事后方法面临着根本性的**有效性—可信度** 权衡：解释可能会降低预测准确率，可能无法忠实反映模型的实际推理过程，或者可能缺乏化学知识的支撑 [36]。

概念瓶颈模型（CBMs）通过将可解释性直接嵌入模型架构，为此提供了一种有原则的替代方案 [26,1,32]。CBM 首先预测人类可以理解的中间概念，然后仅使用这些概念进行最终预测。这种设计确保了解释天然忠实于决策过程，因为每一次预测都必须经过可解释的概念层。然而，将 CBM 应用于分子性质预测会带来特定于领域的挑战：

❶ 如何选择概念？（相关性差距）：与计算机视觉不同，计算机视觉中的概念（例如翅膀的颜色）直观易懂，而化学描述符空间包含数百种理化性质。确定与下游任务（例如毒性）相关的特定子集并非易事，而且往往超出专家的手动处理能力。❷ 数据在哪里？（标注差距）：标准分子数据集只提供结构和最终标签。训练概念瓶颈模型需要对中间概念进行监督，而这些概念通常缺失，形成了入门障碍。❸ 性能与透明度是否互不相容？（容量差距）：人们普遍担心，施加由人类可解释概念构成的瓶颈可能会限制模型的表达能力，导致其性能不如端到端黑盒模型。

为应对这些挑战，我们提出了 GlassMol，这是一种面向分子性质预测、与模型无关的概念瓶颈模型适配方法。尽管 GlassMol 在形式上与模型无关，我们仍通过实验验证了它与图神经网络和大语言模型主干的结合效果，因为它们与化学任务相关且适用 [29,9,37]。GlassMol 通过 RDKit 计算化学描述符，弥合了标注差距；通过大语言模型引导的概念选择，应对了相关性差距；并通过实证表明，可解释性不必以牺牲性能为代价。总之，本文的贡献如下：

– **方法**。一种模型无关的概念瓶颈模型框架及其对应的开源代码，可用于具有连续概念标签的分子性质预测任务。– **评估**。在十三个基准数据集上进行了系统评估，结果表明该方法的性能与黑箱基线相当或更优，并以实证方式应对了容量差距。– **验证**。通过案例研究，将概念归因与成熟的结构重要性方法进行比较，从而对可解释性进行有依据的验证。

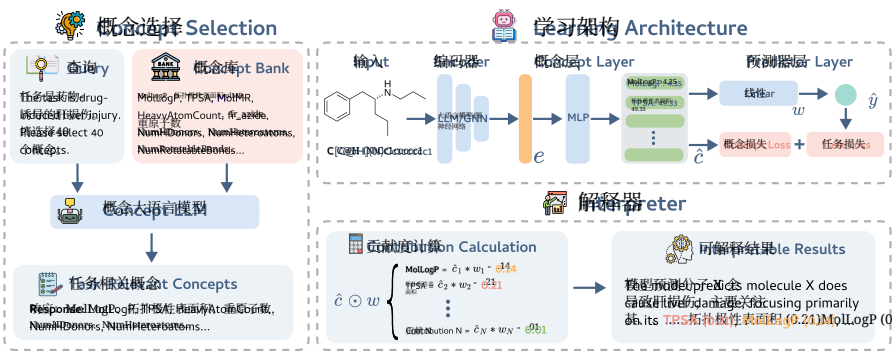


图1：**所提出的 GlassMol 框架**。编码器的输入是原始数据，可以是 SMILES 字符串或分子图。由大语言模型或图神经网络生成的嵌入表示被投影到由大语言模型选出的一组任务相关概念上。随后，预测得到的概念值被传入线性层，以产生最终预测。每个概念的重要性由其对该最终输出的贡献幅度决定，该贡献计算为概念值与其权重的乘积。

2 相关工作

分子表示学习。图神经网络 [21] 已广泛应用于二维和三维分子任务，通过监督和无监督方法捕捉拓扑特征 [12,9]。基于大语言模型的方法，包括化学预训练和各种提示方式，其表现可与图神经网络方法相媲美 [37,19]。此外，近期研究致力于融合图神经网络与大语言模型的嵌入表示，以最大限度地提升信息表达能力 [29,8]。尽管这两种学习策略在分子任务上均展现出最先进的性能，但它们并非以可解释性优先为设计理念，因此限制了与领域专家的协作 [30,33]。事实上，图神经网络和大语言模型本质上都缺乏可解释性：图神经网络将分子结构编码为难以拆解的稠密向量表示，而大语言模型则将知识分散在数十亿个参数中，且没有明确的推理轨迹 [47,45]。

药物发现中的可解释性。为解决这种不透明性，附加方法通常通过数据操纵、输入探测、子图探索和不确定性量化来评估模型行为。尽管无需重新训练，这些事后方法仍面临“可信度”问题：它们只是近似模型的推理过程，而不是将其揭示出来，因此可能学到错误的概念 [20,18,5]。相比之下，自解释方法被视为化学领域可解释性的黄金标准，因为它们易于人类理解，并且已整合到模型中 [35]。然而，现有的自解释方法通常只能生成基于子图的解释，对于非专业人士而言仍然难以理解 [31,17,3]。

概念瓶颈模型。概念瓶颈模型通过在预测之前将潜在特征投影到语义概念空间中，为自我解释提供了一种途径 [26,24] ，并已在计算机视觉和基于文本的任务中展现出其有效性 [23,25,39]。目前已提出多种架构：将隐藏状态投影到预定义概念的线性层 [39], ，以及神经元级概念表示 [16]。在图像领域，大语言模型已被用于构建概念库 [34]。然而，这些方法

尚未系统地应用于分子性质预测，而相关性差距和标注差距带来了挑战。GlassMol在这些基础上构建，并遵循 [11] 的指导，以避免概念预测不准确性。

3 提出的方法

本节介绍 GlassMol，这是一种模型无关框架，旨在提升分子性质预测的透明度。尽管在形式上与模型无关，我们仍在大语言模型和图神经网络上对该方法进行了实验验证，因为它们与化学任务密切相关且适用性强 [33,29,30]。通过将高维分子嵌入表示显式地与人类可理解的化学概念对齐，GlassMol 弥合了黑箱性能与科学可解释性之间的差距。

3.1 问题形式化

令 $\mathcal{D} = \{(x_i,\ y_i)\}_{i=1}^{N}$ 表示一个分子数据集，其中 $x_i \in \mathcal{X}$ 是分子输入（例如图或 SMILES），$y_i \in \mathcal{Y}$ 是目标性质。标准黑箱模型学习直接映射 $F:\ \mathcal{X} \to \mathcal{Y}$。相比之下，概念瓶颈模型引入中间概念空间 $\mathcal{C} \subseteq \mathbb{R}^K$，将预测分解为一个结构化过程：

$$\hat{y}_i = h_\psi(\underbrace{g_\phi(f_\theta(x_i))}_{\hat{\mathbf{c}}_i}) = \sigma(\mathbf{w}^\top \hat{\mathbf{c}}_i + b) \tag{1}$$

其中，f_θ 提取潜在嵌入，g_ϕ 将其映射为可解释的概念得分 $\hat{\mathbf{c}}_i \in \mathbb{R}^K$，而 h_ψ 通过由权重 $\mathbf{w}$参数化的线性组合生成最终预测。这种线性性确保决策逻辑透明。

挑战。 训练此类系统需要对中间概念 $\hat{\mathbf{c}}_i$进行监督，但分子数据集缺乏真实概念标注$\mathbf{c}^*$，从而产生标注差距。此外，化学描述符空间十分庞大，使得识别任务相关的 K 概念子集成为一项非平凡的相关性差距。GlassMol 通过自动化领域推理构建替代真实值$\mathbf{c}^*$，从而同时弥合这两项差距。

3.2 通过大语言模型进行自动化概念整理

为弥合标注差距与相关性差距，我们提出一个两阶段整理流程，利用计算化学确保精度，并利用大语言模型进行推理。

阶段 1：真实值生成。 首先，为解决标注差距，我们利用 RDKit [27]作为计算预言机。对于每个分子 $x_i \in \mathcal{D}$，我们计算一套全面的 $M = 200$ 物理化学描述符，记为全局概念池 $\mathcal{C}_{pool} = \{c^{(1)}, \ldots, c^{(M)}\}$。这些描述符包括拓扑性质、化学片段、分子性质、表面积估算值、杂环估算值、三维结构性质、Lipinski 参数以及药物相似性定量估算（QED）。该步骤将原始数据集转换为 $\mathcal{D}' = \{(x_i, y_i, \mathbf{c}_i^{pool})\}$，从而在无需人工劳动的情况下提供科学准确的真值。

阶段2：任务感知概念选择。使用整个描述符池 $\mathcal{C}_{pool}(M = 200)$并非最优（见图5a），因为不相关概念会引入噪声并削弱可解释性。为解决这一相关性差距，我们采用 ChatGPT（GPT-4）[15] 作为语义筛选器。形式上，令 T 为下游的文本描述

任务（例如，“预测药物性肝损伤”）。我们将大语言模型建模为选择算子 $\mathcal{S}_{LLM}$，用以识别前K 个最任务相关概念的索引:

$$\mathcal{I}_{task} = \mathcal{S}_{LLM}(T, \mathcal{C}_{pool}) \subset \{1, \ldots, M\}, \quad |\mathcal{I}_{task}| = K \tag{2}$$

随后，通过汇集选定的描述符，构建分子 x_i 的最终任务专属概念向量 $\mathbf{c}_i^* \in \mathbb{R}^K$： $\mathbf{c}_i^* = [c_i^{(j)}]_{j \in \mathcal{I}_{task}}$。这一过程生成了紧凑且具有科学依据的监督目标。

3.3 GlassMol架构

GlassMol架构将分子数据逐步从原始输入转换为潜在嵌入，再转换为可解释的概念得分，最终生成任务预测。如图 **1** 所示，这一过程包含三个紧密集成的模块。

潜在特征提取。 这一过程始于**骨干编码器** f_θ，它充当特征提取引擎。
GlassMol的一项关键设计原则是能够适应不同的分子模态。对于图结构输入，我们采用基于GINEConv的图神经网络 [13] 来捕获拓扑特征。对于序列形式的SMILES输入，我们采用专用于化学领域的SMILY-APE大语言模型 [28]。无论选择哪种主干，该模块都会将原始分子 x 映射为潜在嵌入$\mathbf{e} = f_\theta(x) \in \mathbb{R}^d$。

概念投影。为了弥合这些抽象表示与人类理解之间的差距，**概念投影器** g_ϕ 会转换嵌入表示$\mathbf{e}$。与将嵌入表示直接映射为标签的标准分类器不同，该模块是一个多层感知机，它将潜在向量投影到上一节定义的、特定于任务的概念空间中，从而得到预测的概念向量:

$$\hat{\mathbf{c}} = g_\phi(\mathbf{e}) \in \mathbb{R}^K \tag{3}$$

$\hat{\mathbf{c}}$ 中的每个元素都表示模型对某种物理化学性质（例如 LogP）的估计，这要求网络明确表示分子的属性。
透明推理。最后，由**线性预测器** h_ψ作出预测。通过将这一层限制为简单的线性变换，我们确保决策逻辑完全透明:

$$\hat{y} = h_\psi(\hat{\mathbf{c}}) = \sigma(\mathbf{w}^\top \hat{\mathbf{c}} + b) \tag{4}$$

这种设计确保最终预测仅由预测的概念决定，而可解释性则源于 h_ψ的线性性。

3.4 优化与可解释推理

联合优化。为了确保 GlassMol 同时实现较高的预测准确率和忠实的概念学习，我们使用复合目标函数对整个框架进行端到端优化。损失函数定义如下:

$$\mathcal{L} = \mathcal{L}_{\text{task}}(y, \hat{y}) + \lambda \cdot \mathcal{L}_{\text{concept}}(\mathbf{c}^*, \hat{\mathbf{c}}) \tag{5}$$

其中， $\mathcal{L}_{\text{task}}$ 是分类目标的二元交叉熵损失。概念监督项 $\mathcal{L}_{\text{concept}}$定义为预测的概念得分 $\hat{\mathbf{c}}$ 与通过RDKit计算得到的真值 $\mathbf{c}^*$之间的平均绝对误差（L1损失）:

$$\mathcal{L}_{\text{concept}} = \frac{1}{K} \sum_{k=1}^{K} |c_k^* - \hat{c}_k| \tag{6}$$

超参数 λ 控制任务性能与概念对齐之间的权衡。正如消融研究所示（见图5），设置 $\lambda = 1$ 通常能达到最佳平衡。此外，较低的 λ与可解释性变差和概念准确率降低有关 [11]。

可解释推理。 训练完成后，模型提供的不只是预测，还会对其决策过程给出细粒度解释。由于最终预测器 h_ψ的线性性，预测的逻辑值可以精确分解为各个概念贡献之和。对于给定分子，第 k个概念的贡献分数s_k 计算如下：

$$s_k = w_k \cdot \hat{c}_k \tag{7}$$

此处，$\mathbf{w} \in \mathbb{R}^K$ 直接量化每个概念的贡献：较大的正s_k 表明概念 k 会强烈推动模型做出正向预测，而负值则意味着抑制作用。这使领域专家能够验证模型的推理是否与既定化学原理一致。

4 实验

我们开展全面实验以验证GlassMol的有效性。这些实验旨在探究以下研究问题：

– **研究问题1（有效性）：** 该模型能否在图神经网络与大语言模型两种主干上均展现出具有竞争力的性能？ – **研究问题2（可解释性）：** GlassMol能否学习到具有化学意义的概念？ – **研究问题3（消融）：** 各模型组件对性能的关键程度如何？

4.1 实验设置

数据集。 我们在 Therapeutics Data Commons [14] 的 13 个广泛使用的基准数据集上评估所提出的框架。这些数据集分为 ADME 属性（如血脑屏障、人肠道吸收）和毒性任务（如药物性肝损伤、AMES）。对于所有数据集，我们均采用骨架划分 （80/10/10），严格评估模型对未见过的化学结构的泛化能力。

基线方法。 为评估所提出框架的有效性，我们将其与两个主要类别的基线方法进行比较：（1）图神经网络基线：包括 GINE [13]、GCN、图注意力网络和 GraphSAGE 等成熟的图神经网络；（2）大语言模型基线：我们对多种基于大语言模型的方法进行基准测试，主要包括专注于化学领域的 SMILY-APE [28]，以及通义千问和 Llama。

实现细节。 对于瓶颈层，我们使用 RDKit 生成真实概念，并使用 ChatGPT 进行任务感知的筛选。根据消融研究的指导，超参数 λ 设为 1，并选取 40 个概念。所有实验和分析均在单台 NVIDIA H100 GPU 上进行。

4.2 性能比较（研究问题1）

表1将GlassMol与黑箱基线进行比较，以考察透明度与性能之间的关系。

表 1：在 Therapeutics Data Commons 数据集上的实验结果。除不包含概念损失和概念层外，基线架构与各自的 GlassMol 架构完全相同。我们报告五次运行的曲线下面积（均值 ± 标准差）。**粗体** 表示同一架构下的最佳性能。我们方法（**`GlassMol`**）对应的列以高亮显示。

数据集	大语言模型架构		图神经网络架构	
	基线	GlassMol（我们的方法）	基线	GlassMol（我们的方法）
—*ADME*性质—				
BBB	0.856 ± 0.010	**0.892** ± 0.023	0.840 ± 0.006	**0.867** ± 0.014
生物利用度	0.556 ± 0.054	**0.679** ± 0.010	**0。615** ± 0。044	0.572 ± 0.045
HIA	0.880 ± 0.111	**0.989** ± 0.007	0.768 ± 0.041	**0.867** ± 0.057
Pgp	0。801 ± 0.005	**0.903** ± 0.003	**0.899** ± 0.011	0.861 ± 0.014
CYP2C9	0。822 ± 0.001	**0.859** ± 0.010	0.851 ± 0.008	**0.875** ± 0.004
CYP2D6	0.761 ± 0.004	**0.793** ± 0.021	0.802 ± 0.003	**0。843** ± 0.005
CYP3A4	0.799 ± 0.002	**0.848** ± 0.002	0.840 ± 0.002	**0.868** ± 0.004
CYP2C9 _Subs	0.504 ± 0.031	**0.545** ± 0.024	0.599 ± 0.044	**0.627** ± 0.043
CYP2D6 _Subs	0.670 ± 0.014	**0.747** ± 0.024	0.665 ± 0.023	**0.695** ± 0.034
CYP3A4 _Subs	0.532 ± 0.015	**0.548** ± 0.022	0.523 ± 0.036	**0.540** ± 0.039
—毒性—				
DILI	0.814 ± 0.008	**0.851** ± 0.011	0.818 ± 0.030	**0.850** ± 0.027
hERG	0.726 ± 0.027	**0.765** ± 0.013	**0.790** ± 0.030	0.695 ± 0.029
AMES	0.792 ± 0.009	**0.842** ± 0.008	0。855 ± 0.005	0.855 ± 0.003
平均值	0.732	**0.789**	0.766	**0.778**

挑战权衡。 与常见担忧相反，我们的结果表明，可解释性并不必然以牺牲性能为代价。在大语言模型设定中，GlassMol在所有数据集上均持续优于基线，曲线下面积平均提升0.057。同样，在图神经网络设定中，我们的可解释模型在13项任务中的9项上优于黑箱基线，并在1项上持平，平均提升0.012。

任务特定增益。值得注意的是，在毒性任务（例如药物性肝损伤、AMES）中，性能提升最为显著。这表明，对化学概念进行显式建模，有助于模型聚焦于特定毒性基团（结构警示），而不是过度拟合高维潜在空间中的虚假相关性。这回答了**研究问题1**：GlassMol 表明，概念瓶颈模型架构的表现可与基线模型相媲美；大语言模型架构展现出持续的性能提升，而图神经网络架构在大多数任务上仍具竞争力。

4.3 可解释性与验证（研究问题2）

为回答**研究问题2**，我们通过潜在空间可视化和细致的案例研究，验证所学习的概念是否具有化学意义。

潜在空间解耦。 图 2 展示了倒数第二层的 t-SNE 嵌入表示。基线模型（左）生成的表示相互纠缠，没有清晰的决策边界；而 GlassMol（右）则产生了明显且彼此分离的簇。这表明，显式的概念监督有助于解耦潜在化学空间，使线性预测器能够做出稳健且有依据的决策。

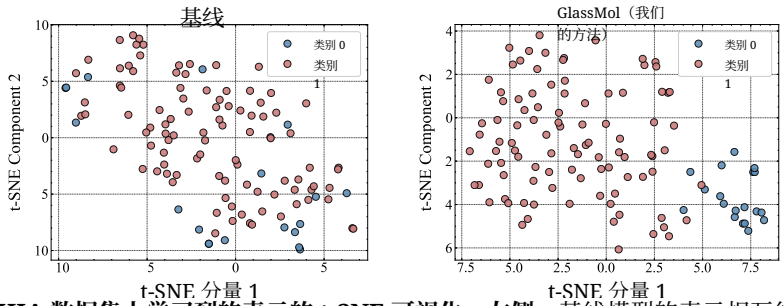


图2：**HIA 数据集上学习到的表示的 t-SNE 可视化。左侧：**基线模型的表示相互纠缠，没有清晰的决策边界。**右侧：**GlassMol（我们的方法）为两个类别生成了分离良好的簇。

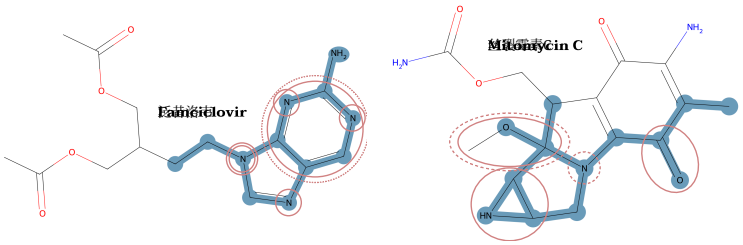


图3：**概念归因与真实结构重要性的定性比较。**蓝色高亮区域表示 TopoPool 标注的真实重要性簇。红色圆圈表示 GlassMol 识别出的子结构，其中实线表示 → 图神经网络架构，虚线表示 → 大语言模型架构。

案例研究与真值验证。 我们将分子子结构（s_k）的概念贡献与 TopoPool [40] 的结构重要性图进行比较，以验证解释的忠实性。我们从不同任务中选取了两个具有代表性的分子，以展示可解释性。请注意，高亮区域与圆圈的重叠表明模型与真值高度一致。如图 3 所示，我们分析以下分子：

– **泛昔洛韦（HIA任务）：** 基于图神经网络的 GlassMol 正确识别出苯胺、芳香氮和叔胺是贡献最大的因素，与 TopoPool 的评估一致。值得注意的是，大语言模型架构呈现出更集中的归因模式，仅突出苯胺基团，其贡献分数是第二大因素的十倍。– **丝裂霉素C（DILI任务）：** 基于图神经网络的 GlassMol 正确识别出甲氧基、羰基氧和哌嗪基团是肝损伤预测的关键驱动因素。大语言模型架构再次选择了较少的概念，但仍与 TopoPool 高度吻合。这些子结构都是已知的代谢易感性因素，证实该模型符合药物化学直觉。

这些结果回答了**研究问题2**，证实 GlassMol 学习到了忠实且具有化学依据的概念。

4.4 消融研究（RQ3）

我们探究各个组件的贡献，以回答**RQ3**。图 4 展示了全面的结果，涵盖三个关键维度。

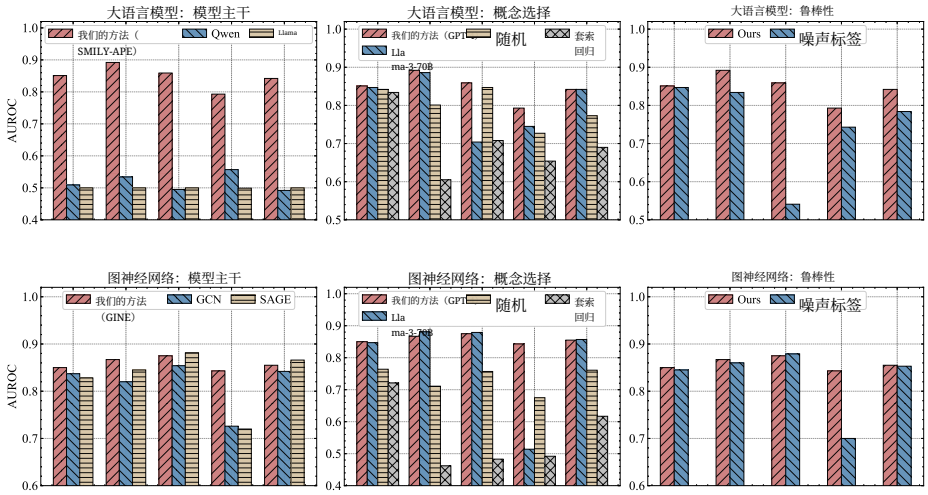


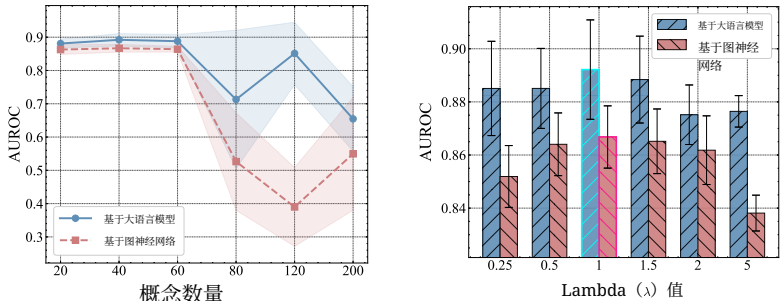
图 4：**针对大语言模型（上排）和图神经网络（下排）架构中关键模型组件的消融实验。左：**模型骨干网络对比。对于大语言模型，我们的特定任务小模型（红色）显著优于通用大模型（蓝色/米色）。对于图神经网络，我们基于 GINE 的架构仍具有竞争力。**中：**概念选择策略评估。GPT-4 选出的概念（红色）表现最佳，开源 Llama-3-70B 选择器（蓝色）则是专有模型的有力替代方案。**右：**鲁棒性分析表明，即使概念标签带有噪声（蓝色），我们的方法仍能保持较高性能。

模型主干（左侧面板）。我们探究架构和模型规模对性能的影响。

– 大语言模型：将默认的化学专用小型模型（SMILY-APE，34.7M）替换为通义千问-7B 和 Llama-8B 等通用模型，会导致性能大幅下降（红色柱与蓝色/米色柱对比）。这凸显了**领域专用预训练对于**分子概念学习**至关重要**。– 图神经网络：对于图编码器，我们基于 GINE 的架构仍具有竞争力。尽管 SAGE 在 2 个数据集上的表现略优于 GINE，但 GINE 在各方面的表现最为稳定，因此适合作为默认主干。

概念选择策略（中间面板）。我们探究了概念选择策略，以确定是否有必要使用大型大语言模型选择器。结果表明，使用 GPT-4 选出的化学相关概念（红色）显著优于随机选择（米色）。基于 Lasso 的稀疏学习（灰色）表现不佳，这表明端到端概念发现难以收敛。值得注意的是，开源的 Llama-3- 70B 选择器（蓝色）取得了与 GPT-4 相当的性能。这一发现对实际应用至关重要，因为它使 GlassMol 能够完全在本地硬件上运行。

鲁棒性（右图）。现实世界的数据往往并不完美。我们通过随机扰动真实概念，测试 GlassMol 在标签含噪情况下的表现。值得注意的是，模型性能几乎没有下降（红色与蓝色柱形），这表明该架构能够抵御标注噪声。这说明模型能够从不完善的监督中提取有意义的模式，而不是死记硬背精确的概念值。



(a) 概念数量的影响（K）　　(b) 对权重参数 λ 的敏感性

图5：**超参数敏感性分析**。(a) 性能与概念数量（K）的关系。准确率迅速提升，并在 $K = 40$附近趋于饱和，从而在表示能力与复杂度之间取得平衡。(b) 性能与损失权重 λ的关系。均衡的权重 $\lambda = 1$可获得最优结果，而极端取值则会降低性能。

超参数敏感性。我们分析了两个关键超参数的影响：概念数量 K 和损失平衡权重 λ。图5a显示，随着 K 从10增至40，模型性能显著提升，这表明要捕捉复杂的化学性质，需要足够数量的概念。然而，超过 $K = 40$ 后性能进入平台期，这说明继续增加概念的收益递减，还可能引入噪声。对于损失权重，图5b表明，$\lambda = 1$ 可实现最佳权衡。较小的 λ 无法对齐概念空间（可解释性较差），而过大的 λ则会迫使模型优先进行概念重构，而非完成主要分类任务。

5 结论

在本研究中，我们将概念瓶颈模型应用于分子性质预测，使模型能够在生成预测之前，将图神经网络和大语言模型嵌入表示投影到概念空间。通过利用 RDKit 进行自动化概念计算，并利用大语言模型进行任务感知概念选择，我们有效弥合了此前阻碍概念瓶颈模型应用于分子任务的标注差距与相关性差距。我们在十三个多样化化学数据集上的实验表明，GlassMol 总体上达到或超过了黑箱基线，这表明可解释性不必以牺牲性能为代价。案例研究进一步验证，所学得的概念归因与既有的结构重要性方法相一致。随着机器学习方法在质量与复杂度上不断发展，概念瓶颈模型架构为使人类专家在预测过程中保持知情并持续参与提供了一条有前景的路径，有望加速安全有效治疗药物的发现。

致谢。本研究部分得益于西北大学 Quest 高性能计算设施提供的计算资源和工作人员支持。该设施由教务长办公室、科研办公室和西北大学信息技术部门共同支持。本研究经费由西北大学 OURG 以 SURG 资助的形式提供。

利益披露。Kaize Ding 是 REAL 实验室的负责人，Oscar Rivera 和 ZiqingWang 是该实验室的成员。REAL 实验室正参与西北大学与艾伯维公司之间的一项合作研究。

参考文献

1. Belinkov, Y., Glass, J.：神经语言处理中的分析方法：综述。TACL。2019年 2. Blaschke, T., Arús-Pous, J., Chen, H., 等：Reinvent 2.0：一种用于从头药物设计的 AI 工具。JCIM。2020年 3. Cao, P.Y., He, Y., Cui, M.Y., 等：群组图：一种性能、效率和可解释性更强的分子图表示。Journal of Cheminformatics。2024年 4. Chen, X., Liu, Y., Ma, X., 等：通过多视图超图学习实现自解释的分子性质预测 5 2021年 6. Dai, E., Wang, S.：迈向自解释图神经网络。CIKM。 . Deng, J., Yang, Z., Wang, H., 等：对分子性质预测关键要素的系统研究。《自然通讯》。2023年 7. Ding, K., Xu, Z., Tong, H., 等：深度图学习的数据增强：综述。ACM SIGKDD Explorations Newsletter。2022年 8. Fan, Y., Wu, Y., French, R.H., 等：借助大语言模型增强的层次注意力图学习，实现分子溶解度预测。Digital Discovery。2026年 9. Fung, V., Zhang, J., Juarez, E., 等：材料化学领域图神经网络基准测试。《npj计算材料》。2021年 10. Gómez-Bombarelli, R., Aguilera-Iparraguirre, J., Hirzel, T.D., 等：通过高通量虚拟筛选和实验方法设计高效的分子有机发光二极管。《自然·材料》。2016年 11. Havasi, M., Parbhoo, S., Doshi-Velez, F.：解决概念瓶颈模型中的数据泄漏问题。NeurIPS。2022年 12. Hu, F., Chen, D., Liu, Q., 等：通过缺失标签填补改进用于分子性质预测的多任务图神经网络。《机器智能研究》。2025年 13. Hu, W., Liu, B., Gomes, J., 等：图神经网络预训练策略。arXiv 预印本。2019年 14. Huang, K., Fu, T., Gao, W., 等：Therapeutics Data Commons：用于药物发现与开发的机器学习数据集和任务。arXiv 预印本。2021年 15 4 . Hurst, A., Lerer, A., Goucher, A.P., 等：Gpt- o 系统卡。arXiv 预印本。2024年 16. Ismail, A.A., Oikarinen, T., Wang, A., 等：用于蛋白质设计的概念瓶颈语言模型。arXiv 预印本。2024年 17. Jiang, T., Yao, Q., Wang, Z., 等：用于分子性质预测的专家引导子结构信息瓶颈。JCIM。2025年 18. Jiménez-Luna, J., Grisoni, F., Schneider, G.：借助可解释人工智能开展药物发现。《自然·机器智能》。2020年 19 2 . Ju, J., Zheng, Y., Koh, H.Y., 等：M llm：利用大语言模型进行多视图分子表示学习。IJCAI。2025年 20. Kim, B., Wattenberg, M., Gilmer, J., 等：超越特征归因的可解释性：使用概念激活向量（TCAV）进行定量测试。国际机器学习会议。2018年 21. Kipf, T.N., Welling, M.：基于图卷积网络的半监督分类。2017年 22. Klucznik, T., Mikulak-Klucznik, B., McCormack, M.P., 等：由计算机规划并在实验室中完成的多种具有药用价值靶标的高效合成。《Chem》。2018年 23. Koh, P.W., Nguyen, T., Tang, Y.S., 等：概念瓶颈模型。国际机器学习会议。2020年 24. Kumar, N., Berg, A.C., Belhumeur, P.N., 等：用于人脸验证的属性分类器和相似性分类器。ICCV。2009年 25. Labadie-Tamayo, R., Slijepˇcevi´c, D., Chen, X., 等：从大语言模型中蒸馏知识：一种用于仇恨言论与反仇恨言论识别的概念瓶颈模型。IP&M。2026年

26. Lampert, C.H., Nickisch, H., Harmeling, S.：通过类间属性迁移学习检测未见过的对象类别。计算机视觉与模式识别会议。2009年 27. Landrum, G. 等：RDKit：开源化学信息学软件。2016年 28. Leon, M., Perezhohin, Y., Peres, F. 等：比较 SMILES 和 SELFIES 分词，以增强化学语言建模。《科学报告》。2024年 29. Li, H., Li, D., Zhang, J. 等：3D-MolGL：将 3D 分子图整合到语言模型中的多模态框架。《专家系统与应用》。2026年 30. Li, L., Zhang, Y., Wang, G. 等：用于分子性质预测的Kolmogorov–Arnold图神经网络。《自然·机器智能》。2025年 31. Liu, Y., Ding, K., Lu, Q. 等：迈向自我可解释的图级异常检测。NeurIPS。2023年 32. Madsen, A., Reddy, S., Chandar, S.：神经自然语言处理的事后可解释性：一项综述。《ACM计算调查》。2022年 33. Niyongabo Rubungo, A., Arnold, C., Rand, B.P. 等：LLM-Prop：利用大语言模型预测晶体材料的性质。《npj计算材料》。2025年 34. Oikarinen, T., Das, S., Nguyen, L.M. 等：无标签概念瓶颈模型。arXiv 预印本。2023年 35. Qadri, Y.A., Shaikh, S., Ahmad, K. 等：可解释人工智能：药物发现的一个视角。《药剂学》。2025年 36. Radhakrishnan, A., Nguyen, K., Chen, A. 等：问题分解提升模型生成推理的忠实度。arXiv 预印本。2023年 37. Ross, J., Belgodere, B., Chenthamarakshan, V. 等：大规模化学语言表征捕捉分子结构与性质。《自然·机器智能》。2022年 38. Schneider, G., Fechner, U.：类药分子的计算机辅助从头设计。《自然综述：药物发现》。2005年 39. Tan, Z., Cheng, L., Wang, S. 等：通过概念瓶颈解释预训练语言模型。亚太知识发现与数据挖掘会议。2024年 40. Thieme, M., Hassan, M., Rupakheti, C. 等：TopoPool：用于提取分子和蛋白质子结构的自适应图池化层。NeurIPS。2024年 41. Wang, Z., Wen, Y., Pattie, W. 等：POLO：面向先导化合物优化的偏好引导式多轮强化学习。arXiv 预印本。2025年 42. Wang, Z., Zhang, K., Zhao, Z. 等：文本引导的分子发现大语言模型综述：从分子生成到优化。arXiv 预印本。2025年 43. Wen, Y., Xu, C., Hu, J.Y.C. 等：AlignAb：用于设计类天然抗体的帕累托最优能量对齐。2024年 44. Ye, X., Durrett, G.：少样本文本推理提示中解释的不可靠性，2022年。arXiv 预印本。2022年 45. Yuan, H., Yu, H., Gui, S. 等：图神经网络的可解释性：一项分类综述。《IEEE模式分析与机器智能汇刊》。2022年 46. Yuan, H., Yu, H., Wang, J. 等：通过子图探索实现图神经网络的可解释性。国际机器学习会议。2021年 47. Zhao, H., Chen, H., Yang, F. 等：大语言模型的可解释性：一项综述。《ACM智能系统与技术汇刊》。2024年 48. Zheng, Y., Koh, H.Y., Ju, J. 等：用于分子性质预测的科学发现大语言模型。《自然·机器智能》。2025年