<SO{v4>：用于显式图结构理解的单个大语言模型词元

Jingyao Wu[1], **Bin Lu**[1*], **Zijun Di**[1], **Xiaoying Gan**[1], **Meng Jin**[1], **Luoyi Fu**[1], **Xinbing Wang**[1], **周成虎**[2]

[1]上海交通大学，[2]地理科学与资源研究所，中国科学院{
wujingyao,robinlu1209,dzj75du,ganxiaoying,jinm,yiluofu,xwang8}@sjtu.edu.cn,
zhouch@lreis.ac.cn

arXiv:2602.01771v1 [cs.CL] 2 Feb 2026

摘要

大语言模型在理解非结构化数据方面展现出巨大潜力，但由于图结构幻觉，在处理图时仍面临重大挑战。现有方法主要有两种：将图转述为自然语言，这会导致词元消耗过多且注意力分散；或将图转换为可训练的连续嵌入（即软提示），但这会与原始文本词元严重不匹配。为解决这一问题，我们提出引入一个特殊词元 <SO$\mathcal{G}_k$>，在统一的词元空间中完整表示图的**结构之图**，从而支持显式拓扑输入和结构信息共享。具体而言，我们提出一种拓扑感知结构分词器，将每种图拓扑映射为一个高度选择性的单一词元。随后，我们构建一组混合结构问答语料，以使新的结构词元与现有文本词元对齐。通过这种方法，<SO$\mathcal{G}_k$> 使大语言模型能够以简洁且准确的方式进行理解、生成和推理。在五个图级基准数据集上的大量实验表明，我们的方法优于基线方法，性能提升了 9.9–41.4%，同时具备可解释性和一致性。此外，我们的方法还能灵活扩展到节点级任务，实现全局和局部结构理解。代码库已公开[1]。

1 引言

近年来，大语言模型（LLM）在非结构化文本的通用理解方面展现出了令人瞩目的能力（Guo 等人，2025年）。然而，大语言模型在理解文本属性图方面仍存在显著局限，例如知识图谱（Wang 等人，2025b；Lin 等人，2025年）、引文网络（Chen 等人，2024年；Tang 等人，2024年）和交易网络（Lei 等人，2025年）。与 LLM作为增强器方法（He 等人，2024a；Liu 等人，2024年；Li 等人，2024b）相比，LLM作为预测器方法（Ye等人，2024年；Tang 等人，2024年；Chen 等人，2024年；Huang等人，2024年）因其在不同层级任务和数据集／领域上的泛化能力而受到更广泛的关注。因此，对于大语言模型的序列化输入格式而言，理解图的核心挑战在于如何以完整、准确且高效的方式将结构输入模型。

现有工作主要通过两种方式转换图拓扑：一种是图到文本，另一种是图到嵌入。(1) 图到文本方法（Ye等人，2024年；Zhao 等人，2023b；Fatemi 等人，2024年）直接使用自然语言描述图拓扑，并在以往工作中得到广泛应用。然而，随着现实世界图规模不断增大，在文本中描绘完整拓扑会消耗大量词元，从而分散大语言模型的注意力，并损害整体结构理解。此外，现有工作（Cao 等人，2025年）揭示，大语言模型表现出明显的结构幻觉：即使节点/边顺序发生细微排列变化，也可能产生不同的输出，导致不稳定的结构理解。(2) 为了减少词元开销，图到嵌入方法（Tang 等人，2024年；Chen 等人，2024年；Huang等人，2024年；Liu 等人，2025a）将图

*联系作者：Bin Lu（robinlu1209@sjtu.edu.cn）[1]项目代码可在 https://github.com/Jingyao-Wu/SOG 获取。

拓扑压缩为连续嵌入，并将其投影到大语言模型中。然而，这不可避免地引入了模态不匹配。投影后的图嵌入与大语言模型固有的词元空间难以良好对齐，导致较弱的跨模态兼容性，并降低结构理解能力。因此，我们提出以下问题：如何简洁且准确地将图拓扑信息输入大语言模型，以避免结构幻觉？

为解决上述问题，我们提出为大语言模型引入**唯一一个特殊词元**，记作<SO$\mathcal{G}_k$>，以便在与文本词元相同的空间中完整表示**StructureOfGraph**。<SO$\mathcal{G}_k$>有助于显式输入拓扑并共享结构信息。具体而言，我们首先提出一种拓扑感知图结构分词器，用于忠实地编码每个图拓扑，并从离散词元词表中提取一个结构词元。这个紧凑而富有表达力的词元封装了关键的拓扑信息，同时具有高度选择性。为了使<SO$\mathcal{G}_k$>与大语言模型的文本输入对齐，我们随后构建了一组混合结构问答（QA），在任务无关的上下文中结合文本词元和结构词元。经过在这些新颖语料库上的微调后，该方法引导结构词元进入与文本词元相同的空间，从而促使大语言模型无缝地读取并理解拓扑。

总而言之，我们工作的主要贡献如下：

- 我们提出仅使用一个结构标记 <SO$\mathcal{G}_k$> ，该标记由具备拓扑感知能力的图词元化器生成，从而以高度精选的方式有效地向大语言模型提供结构信息。
- 为对齐两类词元的嵌入空间，我们设计了一组混合结构问答，指导大语言模型理解、生成并推理结构，从而协调文本与拓扑，实现统一理解。
- 在五个图级基准上的大量实验表明，我们的方法优于所有基线方法，性能提升幅度为 9.9% 至 41.4%，同时展现出良好的可解释性与一致性。此外，我们的方法还可有效扩展至节点级分类，表明 <SO$\mathcal{G}_k$> 兼具全局与局部结构感知能力。

2 相关工作

近年来，大语言模型在理解结构化图数据方面展现出良好前景，核心问题在于如何将非欧几里得图输入序列化的大语言模型。现有工作主要可分为两类：图到文本方法 和图到嵌入方法。

图到文本方法使用自然语言将图表述出来，把其结构转换为文本序列，这种做法在现有研究中得到广泛应用。最直接的方法（Ye等人，2024年；Wang 等人，2023年）通过枚举边来描述图连通性，并以“节点 A 与节点 B 相连”这种形式间接呈现边属性。此外，Li 等人（2024a）和 Wang 等人（2023年）直接将文本形式的邻接表输入大语言模型。为利用大语言模型的常识知识，Talk Like a Graph（Fatemi 等人，2024年）将节点和边刻画为各种现实世界关系，例如友谊、合著关系和人物关系。InstructGraph（Wang 等人，2024b）将图表示为一个 Python 类，并将节点和边设为属性，以利用大语言模型的代码推理能力。为进一步提高遍历效率，GraphText（Zhao 等人，2023b）通过邻居特征和标签描述中心节点的 k跳邻域结构，从而扩大大语言模型的感受野。LangTopo（Guan 等人，2024年）创新性地约束大语言模型的输出，并在量化层面上使大语言模型对文本格式图输入的理解与图神经网络的结构理解相一致。最近，Dr.E（Liu 等人，2025b）将每一跳的信息概括为大语言模型现有词汇表中的多个文本词元。然而，所有这些图到文本方法都需要消耗大量词元才能完整描述图结构，分散了模型对整体结构理解的注意力。此外，大语言模型对节点/边描述顺序的变化非常敏感，导致结构推理不稳定（Cao 等人，2025年）。

图到嵌入方法 将图压缩为嵌入，然后将其投影到大语言模型的词元空间中；在一些研究中，这也被称为“软提示”。GraphGPT（Tang 等人，2024年）分别对文本特征和结构特征进行编码，通过对比学习获得混合节点嵌入，并借助基于多层感知机的投影器将其映射到大语言模型的词元空间。为降低结构编码复杂度，LLaGA（Chen 等人，2024年）采用拉普拉斯嵌入和无参数节点聚合。超越节点级或跳数级表示后，G-Retriever（He 等人，2024b）对检索到的子图进行编码，并将其池化为图级嵌入，

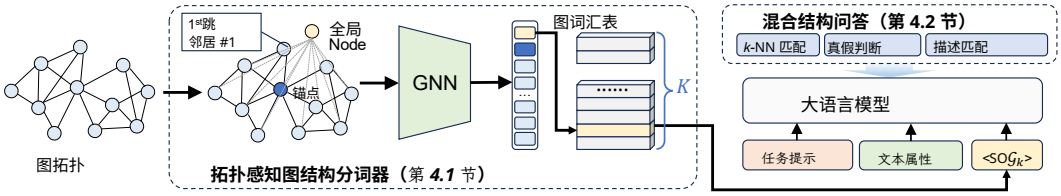

图1：使用结构标记 <SO$\mathcal{G}_k$> 的大语言模型理解总体架构。

随后通过多层感知机将其适配到大语言模型输入空间。还有一些研究进一步增强结构表达能力：TEA-GLM（Wang 等人，2024a）通过边删除和节点掩码构造对比视图，而 GraphLLM（Chai 等人，2023）则采用共享的多层感知机，将图嵌入投影到各大语言模型层的输入空间。GraphAdapter（Huang等人，2024）没有使用多层感知机进行投影，而是利用基于图神经网络的投影器，以提升其在文本属性图上的效率。除了直接投影，TEA-GLM（Wang 等人，2024a）和 GNP（Tian 等人，2024）还在空间对齐之前，分别采用基于主成分分析的映射或基于注意力的预对齐。尽管图到嵌入方法比图到文本方法消耗更少的词元，成功压缩了结构信息，但大语言模型固有的词元空间与投影后的图嵌入之间仍存在严重错位。

总之，我们的方法旨在仅使用一个结构标记，即 <SO$\mathcal{G}_k$>，它与大语言模型固有词元共享同一空间，从而充分刻画图结构。通过这种方式，我们遵循一种新的**图到词元** 范式，使大语言模型能够在图结构与文本属性之间实现和谐的词元映射与理解。

3 预备知识

给定一个文本属性图，它包含图结构 $G = (V, E)$、相应的文本信息T以及标签 y，其中 $V = \{v^1, v^2, \ldots, v^{|V|}\}$ 是节点集合，$E \subseteq V \times V$是边集合。G 的邻接矩阵为 $A \in \{0,1\}^{|V|\times|V|}$，其中 $A_{ij} = 1$ 表示边(v^i, v^j)的存在。我们采用大语言模型 $\mathcal{M}(\cdot; \mathcal{V}, \Theta)$进行图理解，其中 $\mathcal{V}$表示词汇表，Θ 表示参数。在本工作中，我们提出一组结构词元 {<SO$\mathcal{G}_1$>, $\cdots$, <SO$\mathcal{G}_K$>} 以明确反映图的整体结构信息，其中 K 是结构词元的数量。每个结构词元 <SO$\mathcal{G}_k$> 代表一种纯图结构的原型。这些结构词元与大语言模型原有的词汇表 $\mathcal{V}$共享同一个词元空间，由此形成一个新的词汇表 $\mathcal{V}' = \mathcal{V} \cup \{$<SO$\mathcal{G}_1$>, $\cdots$, <SO$\mathcal{G}_K$>$\}$。

随后，为了使用大语言模型作为预测器来获得答案 $\hat{y}$，输入到 $\mathcal{M}(\cdot; \mathcal{V}', \Theta)$ 的内容包括：(1) 任务提示 P，其中包含任务要求、期望的输出格式和候选标签；(2) 文本图属性 T；(3) 对应的结构词元 <SO$\mathcal{G}_k$>。因此，我们的目标是让大语言模型 $\mathcal{M}$生成输出 $\mathcal{O}$，其中包含与真实标签 y相匹配的预测标签 $\hat{y} \subset \mathcal{O}$。形式上，该过程可表示如下：

$$\mathcal{O} = \mathcal{M}(\{P, T, \text{<SO}\mathcal{G}_k\text{>}\} | G; \mathcal{V}', \Theta). \tag{1}$$

这种表述使结构信息能够无缝融入大语言模型的生成式推理过程，从而在图数据上实现更可靠的推断。

4 方法

本节将正式介绍如何生成和利用结构词元，即 <SO$\mathcal{G}_k$>，以如图1所示的两阶段方式增强大语言模型的图理解能力。具体而言，在第一阶段，我们提出了一种拓扑感知图结构分词器，用于提取图拓扑、编码全局信息，并进一步将其投影为一个结构词元。在第二阶段，我们构建多个混合结构问答（QA）对，并将其输入大语言模型，以确保文本词元和结构词元在统一空间中的融合。由此，大语言模型能够理解、生成并推理这些词元，从而在不同图任务中获得满意的结果。

4.1 拓扑感知图结构分词器

给定一个文本属性图，拓扑感知图结构分词器首先提取其图拓扑 $G(V, E)$，仅保留纯结构信息。为增强对拓扑的理解，我们为每个节点 $v^i \in V$ 赋予一个新的结构属性 t_s^i，采用分层

遍历策略。首先，根据节点重要性度量（例如节点度）识别一个锚节点。然后，根据每个节点相对于锚节点的位置来确定其在图中的位置，并形成其节点属性，例如“第一跳邻居 #1”或“第二跳邻居 #3”。同一跳内节点的编号（例如 #1、#3）由相同的节点重要性度量排序后确定。此外，我们引入一个虚拟全局节点来连接所有节点，作为池化机制以获取整体表示。最后，我们采用现成的文本编码器 $f_T(\cdot)$ 来嵌入新增属性：$x_s^i = f_T(t_s^i)$，$X_s = [x_s^1; x_s^2; \ldots\ldots; x_s^{|V|}; x_s^{\text{global}}] \in \mathbb{R}^{(|V|+1)\times d_s}$，其中 d_s 是文本编码器的维度。这种分层策略为每个节点提供了独特的空间坐标和丰富的特征，消除了节点不变置换所导致的模糊描述。

接着，为了聚合其多跳邻居中的节点信息，我们采用图神经网络（GNN）$f_G(\cdot)$作为编码模块（Kipf 和 Welling，2017年）。通过消息传递机制，GNN 捕获拓扑感知信息，将节点结构属性 X_s映射到连续空间中的潜在表示 H_s，其中 $H_s = f_G(X_s)$、$H_s \in \mathbb{R}^{(|V|+1)\times d}$和$d$ 是 GNN 编码器的维度。随后，将虚拟全局节点 $h_s^{\text{global}} \in \mathbb{R}^d$ 的表示作为图级结构表示。然而，这种连续表示无法与大语言模型的离散文本词元对齐，从而导致意外的结构幻觉。

因此，我们提出利用自监督拓扑重建（Yang 等人，2024a）将连续表示映射到 K 离散词元词表中，其中 K 是控制词表大小的超参数。具体而言，我们引入一个可学习的词汇表集合，记为 $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$，它作为结构模式的一组离散潜在嵌入。每个条目 $c_j \in \mathcal{C}$ 对应于在重建任务中学习到的一个结构原型。对于每个节点特征 h_s^i，我们根据欧几里得距离比较并检索最近的码本条目 $c_k = \mathbf{z}_e(h_s^i,\ \mathcal{C})$：

$$k = \arg\min_j \left\| h_s^i - c_j \right\|_2, \quad c_j \in \mathcal{C} = \{c_1, c_2, \ldots, c_K\}.$$

为实现这一目标，我们采用了两种损失函数，即更新损失 $\ell_{\text{update}} = \left\|\text{sg}[h_s^i] - c_k\right\|_2^2$ 和承诺损失$\ell_{\text{commit}} = \left\|h_s^i - \text{sg}[c_k]\right\|_2^2$，以在 c_k 与 h_s^i之间建立桥梁。具体而言，与全局节点对应的词汇表索引被选作整个图的结构词汇表。为体现其拓扑感知能力，所选词汇表条目能够通过轻量级解码器实现拓扑重建。因此，我们引入一个共享的解码器模块 $f_q(\cdot)$，以获得重建的节点特征 $\hat{X} = f_q(\mathbf{z}_e(H_s))$，其中 $\mathbf{z}_e(H_s) \subset \mathcal{C}^{|V|+1}$ 表示与节点特征 H_s对应的所选词汇表条目。随后，利用 X 重建邻接矩阵 $\hat{A} = \hat{X}\hat{X}^T$，从而计算 $\hat{A}$ 与原始 A之间的重建损失。因此，总体损失函数由三个部分组成，进一步表示如下：

$$\mathcal{L} = \underbrace{\left\|A - \hat{A}\right\|_F^2}_{\text{Reconstruction loss}} + \underbrace{\|\text{sg}[H_s] - \mathbf{z}_e(H_s)\|_2^2}_{\text{Update loss}} + \beta \underbrace{\|H_s - \text{sg}[\mathbf{z}_e(H_s)]\|_2^2}_{\text{Commitment loss}},$$

其中 β 是加权超参数。

4.2 混合结构问答的词元对齐

获得图结构词元后，需要将其与大语言模型原有的文本词元对齐，以实现统一理解。为解决这一问题，我们构建混合结构问答，为新增的结构词元分配合理的嵌入，并调整大语言模型编码器，使其具备上下文感知能力。我们为结构问答设计了 3 类语料库，如下所示：

1. **k-最近词元邻居匹配**：问题为“以下是目标结构词元<SOG_i>，其五个最近的图结构词元是：”。随后，理想答案是根据词汇表中的向量余弦相似度得出的 Top-k 个最近词元。该问答语料库确保局部一致性：离散结构表示越接近，大语言模型词元嵌入也越接近。

2. **结构相似性判断（是/否）**：问题是”这里有两个词元 <SOG_i>和 <SOG_j>，判断它们是否表示相似的结构。” 如果余弦相似度

两个对应的全局节点嵌入在连续空间中的余弦相似度大于预定义的正阈值，则答案为”相似”；如果小于负阈值，则答案为”不相似”。该问答语料库借助对比学习的思想，明确令牌嵌入空间中结构相似性的边界。

3. **描述-令牌对匹配**：问题是 “给定目标图：节点 A 与节点 B 相连；节点 A 与节点 C 相连.....对应的图结构词元是：”。答案是其真实的结构词元。该问答语料库直接将文本令牌理解与结构标记理解联系起来，鼓励大语言模型在同一个“大脑”中明确统一这两种词元。

通过这些问答，我们对大语言模型进行监督微调。(1) 为保留预训练知识，我们冻结已有的文本词元，仅更新新添加的结构词元的令牌嵌入。(2) 为进一步提升拓扑感知能力，我们采用一种轻量级参数适配技术，即 LoRA（Hu 等，2022），以高效地将结构信息注入模型。值得注意的是，在这一微调过程中，所构建的问答仅依赖图结构，而忽略了原始文本信息或下游任务。因此，该方法在不同图任务和领域信息上具有很强的泛化能力和可扩展性。

4.3 D下游A应用与 <SO$\mathcal{G}_k$>

为了引导大语言模型适用于不同的下游应用，我们进行任务特定微调阶段，将任务转化为文本生成范式。具体而言，我们首先输入系统提示 S 作为通用指导：“根据目标图的图拓扑（即结构标记）及其文本属性，将其分类到正确的类别。每个结构标记代表一种独特的图模式（例如，相似分子结构的原型）。”随后 S，我们附加用户提示，以明确输入信息和应用场景。大语言模型的用户提示由以下三部分组成：(1) 任务提示 P：明确引导大语言模型理解任务目标，例如“[任务] 预测该分子是否具有毒性。”或“[任务] 将中心论文节点分类，从候选集合中为其分配正确的标签.....” (2) 文本图属性 T：提供图的原始语义特征和背景信息，例如“[分子]该分子的 SMILES 表达式为.....”或“[论文] 中心节点代表一篇标题为.....的论文.....” (3) 指定的结构标记 <SO$\mathcal{G}_k$>：使用“[结构标记]<SO$\mathcal{G}_k$>”表示图拓扑。在任务特定微调过程中，特定下游应用的真实标签 y 会添加在大语言模型输入（即系统提示和用户提示）之后，作为监督信号引导响应生成。损失函数表示如下：

$$\mathcal{L} = -\sum_{t} \log p_{\Theta}(y_t \mid y_{<t}, P, T, \texttt{<SO}\mathcal{G}_k\texttt{>}),$$

其中 t 索引输出序列中的位置。通过这种方式，下游应用可以进一步将文本/结构理解与不同的领域任务对齐。

5 实验

In t本节中，我们实证评估 <SO$\mathcal{G}_k$> vi的有效性与结构贡献 a
exte泛实验，通过回答以下研究问题（**RQ**s）：

- **RQ1**：与其他基线方法相比， <SO$\mathcal{G}_k$> 在图分类任务中的表现如何?
- **RQ2**：所提出的结构词元 <SO$\mathcal{G}_k$>效果如何?
- **研究问题3**：结构词元的可解释性如何？它们是否与文本词元保持对齐?
- **研究问题4**： <SO$\mathcal{G}_k$> 能否扩展，以提升大语言模型在节点级任务上的结构理解能力?
- **研究问题5**： <SO$\mathcal{G}_k$>对结构分词器的关键设计选择有多敏感?

5.1 实验设置

数据集。为评估性能，我们在 MoleculeNet（Wu 等人，2017年）的 5 个数据集上执行图级分类任务，包括 BBBP、Tox21、ClinTox、HIV 和 BACE，涵盖毒理学、临床研究和药物化学领域，共包含 17 个子任务。详细信息汇总于附录 A.1。

表 1：五个数据集上的性能比较，其中**粗体** 表示最佳性能，下划线表示次佳性能。

大语言模型骨干网络 / 方法		BBBP (↑)	Tox21 (↑)	ClinTox (↑)	HIV (↑)	BACE (↑)
GPT-4	零样本	61.5	55.2	51.6	65.9	62.5
GPT-4o	零样本	57.0	36.0	51.8	54.7	38.5
Deepseek-R1	零样本	63.6	60.1	48.2	50.5	52.7
LLaMA3-70B	零样本	60.1	44.9	48.8	58.3	50.9
Galactica-120B	零样本	66.1	68.9	82.6	74.5	61.7
非大语言模型分子方法	GPF-属性掩码	58.8	66.0	-	71.3	62.2
	GPF-GCL	56.5	63.6	-	49.3	52.1
	GIMLET	59.4	61.2	-	66.2	<u>69.6</u>
	MolCA-S	62.5	66.6	-	69.0	62.8
	MolCA-GS	63.6	68.5	-	72.7	63.9
LLaMA3-3B	零样本	50.2 ± 3.4	48.5 ± 2.7	50.2 ± 2.5	51.8 ± 0.8	52.7 ± 1.8
	少样本（随机）	46.5 ± 3.8	49.9 ± 10.2	57.8 ± 6.2	49.7 ± 16.7	68.2 ± 17.5
	少样本 (Morgan Sim)	49.7 ± 2.5	51.0 ± 8.7	53.3 ± 10.0	42.0 ± 14.7	49.7 ± 0.6
	LoRA 监督微调	60.2 ± 1.2	50.8 ± 0.8	63.8 ± 2.8	51.5 ± 0.7	50.5 ± 2.9
	软提示	27.9 ± 1.7	39.9 ± 1.9	35.0 ± 2.0	33.6 ± 0.5	28.5 ± 0.9
	<SO$\mathcal{G}_k$>（我们的方法）	**76.9 ± 3.1**	**83.4 ± 3.3**	<u>85.5 ± 3.7</u>	<u>75.7 ± 1.6</u>	63.3 ± 4.2
LLaMA2-7B	零样本	50.0 ± 0.0	48.4 ± 3.3	44.0 ± 2.6	50.0 ± 0.0	49.9 ± 0.4
	少样本（随机）	49.4 ± 0.6	49.3 ± 2.2	50.3 ± 1.7	46.9 ± 0.8	47.8 ± 1.8
	少样本（Morgan Sim）	51.3 ± 0.8	51.1 ± 2.5	56.8 ± 6.3	50.4 ± 0.7	53.2 ± 1.5
	LoRA 监督微调	55.0 ± 0.8	60.6 ± 3.4	59.3 ± 3.7	78.1 ± 2.2	61.7 ± 2.7
	软提示	53.7 ± 0.8	49.8 ± 1.5	49.7 ± 4.7	44.9 ± 0.2	49.4 ± 2.3
	<SO$\mathcal{G}_k$>（我们的方法）	<u>66.4 ± 2.7</u>	<u>72.4 ± 3.1</u>	**94.3 ± 0.1**	**83.2 ± 1.9**	**98.4 ± 0.8**

评估指标。我们采用生成式问答方法从大语言模型获取结果，并通过文本匹配进行评估，以确定回答中是否包含正确答案。最大输出标记长度限制为 100。大语言模型的采样参数设置为温度 0.6，以兼顾随机性和连贯性；同时将词元选择限制在累积概率质量为 90% 的范围内。为定量评估分类结果，我们使用 AUC-ROC 评估二元图分类任务的性能，因为大多数任务都存在显著的类别不平衡。每项实验运行三次，并报告平均性能。详细信息汇总于附录 A.2。

基线方法。我们将所提出的方法与 20 种基线方法进行比较，这些方法通常分为三类：（1）参数量庞大的大语言模型（≥ 70B），包括 GPT-4（OpenAI，2023年）、GPT-4o（Hurst 等人，2024年）、Deepseek-R1（Guo 等人，2025年）、LLaMA3-70B（Dubey 等人，2024年）、Galactica-120B（Taylor 等人，2022年）；（2）非大语言模型方法，包括 GPF-属性掩码（Fang 等人，2023a；Hu 等人，2020年）、GPF-GCL（Fang 等人，2023a；You 等人，2020年）、GIMLET（Zhao 等人，2023a）、MolCA-S（Liu 等人，2023b）、MolCA-GS（Liu 等人，2023b）；（3）参数量较少的大语言模型方法（即 LLaMA2-7B（Touvron 等人，2023年）、LLaMA3-3B（Dubey 等人，2024年）），包括零样本、少样本（Brown 等人，2020年）、基于 LoRA 的监督微调（LoRA 监督微调）（Hu 等，2022）和软提示（Perozzi 等人，2024年）。详细信息请参见附录 A.3。

实现细节。 我们在附录 B.1 中介绍了方法的实现细节。同时，各项下游任务的提示设计见附录B.4，其他训练配置和细节记录于附录 B.2。

5.2 总体性能（RQ1）

五个图级数据集上的实验结果如表 1 所示。我们得到以下观察结果：(1) <SO$\mathcal{G}_k$>始终优于所有基线方法，性能提升幅度令人满意，介于 9.9% 至 41.4% 之间。在 BBBP 和 Tox21 数据集上，我们的方法在 3B 和 7B 模型上的性能分别排名第一和第二。在其余三个数据集上，我们的方法在 7B 模型上取得了最佳性能。值得注意的是，在 BACE 数据集上，<SO$\mathcal{G}_k$>的 AUC-ROC 达到 98.4，显著优于次优模型。由于不同数据集的理解难度和训练样本数量各不相同，模型参数规模的增加并不会直接提升各个数据集上的性能。(2)与非大语言模型方法（GPF、GIMLET、MolCA）相比，

表2：在五个数据集上对不同结构标记 <SO$\mathcal{G}_k$> 进行消融实验。

模型	消融实验	BBBP	Tox21	ClinTox	HIV	BACE
LLaMA3-3B	(M1a) 无结构标记	61.2	72.3	81.8	53.0	53.2
	(M1b) 静态结构标记	60.7	67.8	52.5	54.6	54.3
	(M1c) 随机结构标记	57.3	62.2	52.8	55.3	55.8
	Ours	**76.9**	**83.4**	**85.5**	**75.7**	**63.3**
LLaMA2-7B	(M1a) 无结构标记	61.3	66.4	79.5	69.6	94.2
	(M1b) 静态结构标记	64.7	69.1	92.0	81.5	94.2
	(M1c) 随机结构标记	62.1	68.6	88.5	**83.4**	96.7
	Ours	**66.4**	**72.4**	**94.3**	83.2	**98.4**

参数规模较大的大语言模型（如 Galactica-120B）具备较强的文本理解能力，即使在零样本设置下，也能在部分数据集上取得更好的结果。然而，受结构理解能力的限制，其性能提升受到制约。我们提出的结构标记方法只需一个标记，就能使较小参数规模的大语言模型（3B 和 7B）超越大型模型的性能，彰显了该方法的有效性。同时，我们的方法在成本和资源利用方面也更具优势。(3)对于 3B/7B 大语言模型的基线方法，少样本方法的结果方差较大。例如，3B 模型在 HIV 和 BACE 上的标准差分别高达 16.7 和 17.5，导致结果不稳定、不可靠。LoRA 监督微调方法通过显式微调参数，使模型适应相应的数据集，从而提升性能。然而，面向大语言模型的图到文本输入方法存在局限，制约了模型性能的提升。软提示存在严重的错位问题，有时甚至会对 3B 等较小模型产生负面影响：相较于零样本性能，性能下降幅度为 8.6 至 24.2 个百分点。相比之下，我们的方法在不同模型规模下均能稳定提升性能，只需一个标记，便能有效且高效地弥合大语言模型文本推理与拓扑理解之间的差距。

5.3 消融研究（RQ2）

结构词元如何影响模型的有效性？ 首先，图2展示了一个简单示例，说明针对某个特定图实例选择不同结构标记 <SO$\mathcal{G}_k$> 时，性能会如何变化。该图的标签为“True”，我们绘制了输入不同结构标记时，大语言模型输出“True”的生成概率。当我们的模型选择预期的

<SO$\mathcal{G}_{157}$>，正确答案的输出概率达到最大。相比之下，其他结构词元的选择会导致结果出现显著差异，这表明我们生成的结构词元具有很强的选择性，选择合适的结构词元能够显著提升模型性能。

此外，我们通过三种不同变体进行了消融实验，如表2所示：**(1)无结构词元**：从大语言模型输入中移除结构词元 <SO$\mathcal{G}_k$>，仅保留任务提示和文本图属性。**(2)随机结构词元**：将特定的 <SO$\mathcal{G}_k$> 替换为随机词元。**(3)静态结构标记**：输入一个不加区分的静态结构词元。我们观察到，所有变体都会导致性能下降，这证实了结构词元在引导大语言模型理解图拓扑方面发挥着至关重要的作用。统计指标表明，结构词元具有很强的选择性：只有我们的拓扑感知图结构分词器生成的正确词元，才能忠实地传达相应图的正确拓扑信息，从而持续提升下游任务表现；而随机或静态替换则无法提供有意义的引导。这一观察有力地证明，我们的大语言模型已经学会将特定的拓扑感知词元与相应的

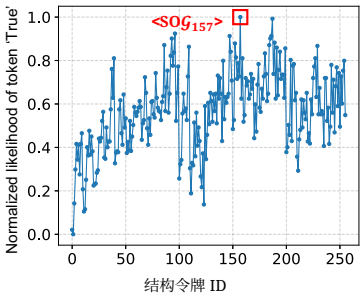

图2：不同结构词元选择下性能变化的一个简单示例。更多示例见附录C中的图7、8、9。

ing 图结构。因此，只有正确映射的词元才能可靠地传递图 h
top论并生成准确的预测。

使用不同的混合问答方法进行指令微调，会如何影响模型性能？ 图3展示了使用三种方法进行指令微调时的模型性能

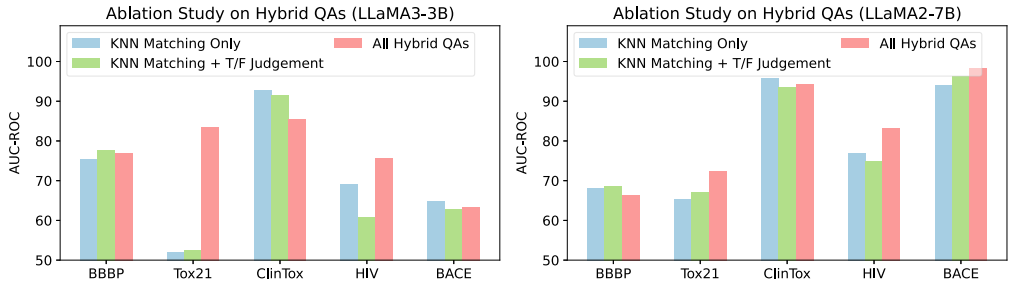

图3：在五个数据集上使用不同混合问答语料库对大语言模型进行调优的消融实验。

不同的问答语料库。我们发现，使用全部混合问答时，模型整体表现更好，尤其是 Tox21 在 3B 模型上取得了显著成果。同时，我们观察到，采用描述词元匹配进行微调有时会导致性能略有下降。这是因为它要求大语言模型进行细粒度的结构表征，从而在某些场景下增加了模型的学习难度。

5.4 案例研究与模型分析（RQ3）

显式结构标记是否具有可解释性？ 为验证这一点，我们选取前 50 个结构标记，并如图 4 所示，展示它们的相关矩阵热图。我们观察到，对角线上的值均为 1.0；除平凡的自相似性外，矩阵中的非对角线元素整体较弱（约为 0.6–0.7）；此外，某些标记（例如标记 <SO$\mathcal{G}_{11}$>）与其他所有标记的相关性都普遍较低。这种模式表明，<SO$\mathcal{G}_k$>之间的冗余度较低，并说明图拓扑的词元化方式紧凑且近乎正交。

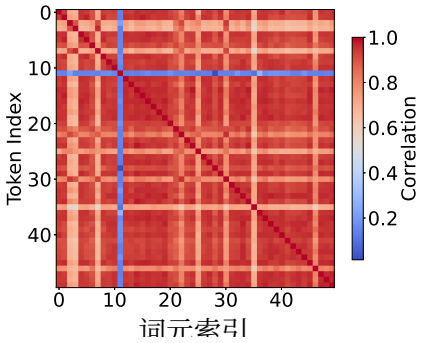

图 4：前 50 个结构标记的相关性热图。

不同数据集中的结构标记选择是否具有相似的结构信息？ 我们考察了不同分子数据集中结构标记的选择及其对应结构，如图所示

在图5中。每一行代表一组示例，其中相应的结构标记选自不同的数据集。我们发现，这些分子结构具有相同的 Bemis–Murcko 骨架（第一列），该骨架通过 RDKit 开源工具包（https://www.rdkit.org/）提取。结果表明，具有相同骨架的分子始终被映射到相同的结构标记。这种骨架级一致性证实，<SO$\mathcal{G}_k$> 能够捕获核心拓扑基元，从而对图结构进行简洁且无冗余的压缩。

结构标记与文本词元是否共享同一嵌入空间？ 在此，我们比较LLaMA2-7B的结构词元、文本词元与图软提示之间的关系。我们随机采样300个图，并通过图卷积网络生成其软提示，然后从大语言模型的词汇表中提取结构词元与高频文本词元的嵌入。我们先使用主成分分析，再使用t-SNE，将这三者投影到二维空间。图6的可视化结果表明，软提示与文本词元占据明显可分离的区域，揭示了图派生嵌入与语言派生嵌入之间的严重错位。相比之下，结构词元嵌入于文本词元的邻域之中，有效桥接了图空间与语言空间。这一模式表明，结构词元将拓扑知识均匀注入大语言模型的词元空间，从而使大语言模型能够读取并推理图结构信息。

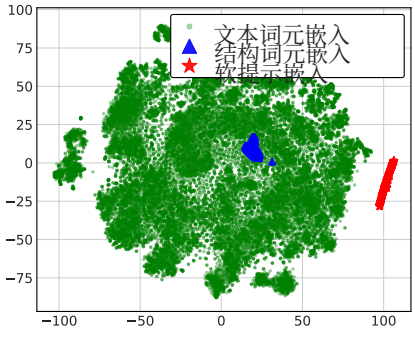

图6：结构标记、文本词元和图派生软提示嵌入的 t-SNE 可视化

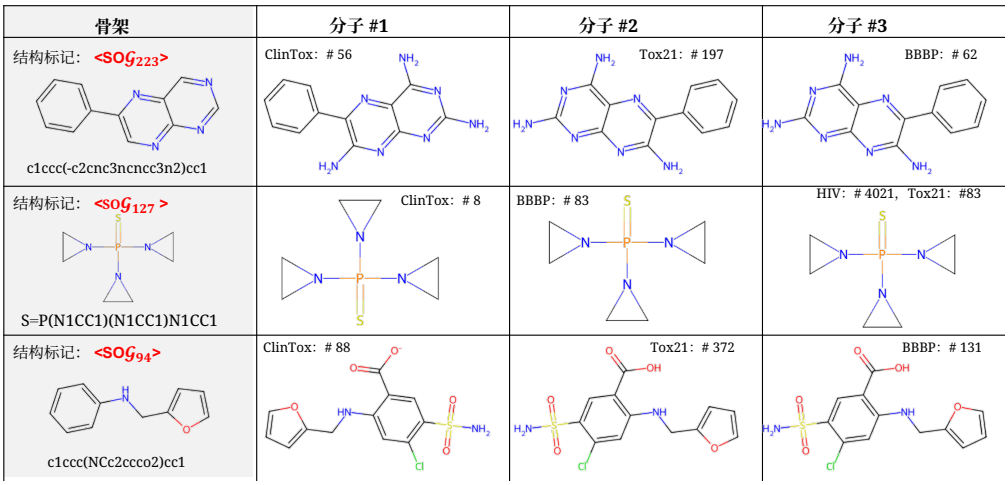

图5：共享相同骨架的分子始终会选择相同的结构标记。附录C中的图10和图11展示了更多示例。

5.5 扩展至节点级分类（研究问题4）

除图级任务外，我们的方法也易于扩展到节点级分类。除全局虚拟节点外，图中的每个节点也会被分配一个对应的结构标记。因此，我们将节点级结构标记注入大语言模型，并在两个广泛采用的基准数据集 Cora 和 Pubmed 上比较其性能（Hu 等人，2020a；Li 等人，2024c）。具体操作请参阅附录 B.3。为进行全面比较，我们还与一系列基线方法进行了比较，包括非大语言模型方法（DGI（Velickovic 等人，2019）、GraphMAE（Hou 等人，2022））、参数量较大的通用大语言模型（LLaMA3-70B（Dubey 等人，2024）、GPT-3.5-turbo（Ouyang 等人，2022）、GPT-4o（Hurst 等人，2024）、DeepSeek-chat（Guo 等人，2025））以及图专用的大语言模型方法（Emb w/ NA（Li 等人，2024c）、OFA（Liu 等人，2024）、ZEROG（Li 等人，2024b）、GraphGPT（Tang 等人，2024））。

表3：节点分类性能比较，其中**粗体**表示最佳性能，下划线表示次佳性能。

模型	Cora		Pubmed	
	准确率 (↑)	F1 分数 (↑)	准确率 (↑)	F1 分数 (↑)
DGI	17.50	12.44	44.88	38.72
GraphMAE	27.08	23.66	22.03	15.65
LLaMA3 (70B)	67.99	68.05	77.00	64.18
GPT-3.5-turbo	65.67	63.22	75.99	69.90
GPT-4o	68.62	68.49	77.96	71.79
DeepSeek-chat	65.62	65.77	79.23	74.30
Emb w/ NA	63.59	58.23	74.66	73.15
OFA	23.11	23.30	46.60	35.04
ZEROG	62.52	57.53	79.08	77.94
GraphGPT	24.90	7.98	39.85	20.07
我们的方法（LLaMA3-3B）	**91.58**	**78.62**	**97.46**	85.50
我们的方法（LLaMA2-7B）	88.80	71.28	96.27	**89.64**

如表3所示，我们的方法在所有数据集上始终取得最先进的性能。值得注意的是，所提出的 <SO$\mathcal{G}_k$> 仅有 3B 个参数，却超越了 GPT-4o 和 DeepSeek-chat 等规模大得多的模型，在准确率上分别高出 GPT-4o 22.96 个百分点，高出 DeepSeek-chat 25.96 个百分点。与在 GraphLLM 基线方法中表现出色的 ZEROG 相比，我们的方法准确率提升了 23%，F1 分数提升了 15%。节点级实验表明，得益于我们的任务无关对齐策略，结构词元能够体现全局和局部结构视角。

5.6 超参数与设计敏感性分析（RQ5）

我们进行结构超参数研究，以考察关键设计选择如何影响 <SO$\mathcal{G}_k$>的有效性。首先，我们评估不同的词汇表大小 K （例如，64、128、256、512），结果如表 4 所示。我们发现，K=256 的整体表现最佳。较小的取值无法充分区分多样的结构模式，而过大的取值则容易导致过拟合并降低泛化能力；因此， K=256 在表达能力与鲁棒性之间提供了良好的平衡，

表5：锚节点选择策略的比较。

Task	随机	PageRank	介数中心性	度*
BBBP p_np	74.4	**82.2**	80.9	76.9
HIV_HIV_活性	72.0	75.4	62.3	**79.7**
BACE_分类	63.5	64.3	65.5	**98.4**
ClinTox FDA 批准	92.4	**99.6**	98.0	77.8
Tox21 NR-AR	69.6	59.6	69.7	**76.4**
Tox21 NR-AR-LBD	66.6	58.7	64.0	**76.0**
Tox21 SR-p53	**66.6**	54.6	56.4	63.3
avg.	72.2	70.6	71.0	78.4

能力和鲁棒性。我们选择 K=256，为建模基准数据集中呈现的结构多样性提供了最合适的容量。

此外，我们对锚点节点选择进行了敏感性分析。默认采用节点度（*），并将其与其他策略进行比较，即 PageRank（Yang 等，2024b）、介数中心性（Xiang 等，2023）和随机选择，以评估它们对结构编码和下游性能的影响，如表5所示。基于度的策略在大多数基准数据集中仍最为稳定有效。由于能够更好地捕捉稀疏分子图中的全局中心性，PageRank 在 BBBP 和 ClinTox 上带来了小幅提升；而随机选择往往会不可预测地改变结构中心，导致编码不稳定、性能下降。

表 4：结构词汇表大小 K 对模型性能的影响。

数据集	结构词汇表大小			
	64	128	256*	512
BBBP	84.6	71.8	76.9	96.2
ClinTox	91.8	90.1	76.9	90
HIV	52.7	51.9	79.7	55.9
BACE	69.8	62.3	98.4	85.7
Tox21	52.6	53.4	81.6	59.8
avg.	**60.1**	**59.3**	**81.7**	**66.8**

6 结论

本文提出使用一个结构词元 <SO$\mathcal{G}_k$>，简洁且准确地帮助大语言模型显式理解拓扑结构。我们的方法解决了大语言模型结构理解中两种主流范式存在的词元效率低下和跨模态错位等关键问题。我们通过实验验证了 <SO$\mathcal{G}_k$> 在图级和节点级基准数据集上的有效性，证明这些结构词元不仅与大语言模型原生的词元空间高度对齐，而且能够有选择地捕捉图特有的结构模式。未来，我们将致力于进一步探索其在更多样化图数据上的能力，例如动态图和异构图。

致谢

本文作者获得了国家自然科学基金委员会（项目编号：T2421002、92579211、62272301、623B2071、62525209）、中国博士后科学基金会博士后资助计划（资助编号：GZB20250806）以及上海交通大学人工智能赋能科学种子计划（项目编号：2025年AI4S-QY01）的资助。

参考文献

Tom B. Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhari- wal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、Christopher Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。语言模型是少样本学习者。载于神经信息处理系统进展第*33*卷：神经信息处理系统*2020*年会，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上举行，2020年。

Yukun Cao、Shuo Han、Zengyi Gao、Zezhong Ding、Xike Xie 和 S. Kevin Zhou。Graphinsight：解锁大语言模型在图结构理解方面的洞见。载于第*63*届计算语言学协会年会论文集（第*1*卷：长篇论文），*ACL 2025*，奥地利维也纳，*2025*年*7*月*27*日至*8*月*1*日，第12096–12134页。计算语言学协会，2025年。

Ziwei Chai、Tianjie Zhang、Liang Wu、Kaiqiao Han、Xiaohai Hu、Xuanwen Huang 和 Yang Yang。Graphllm：提升大语言模型的图推理能力。*CoRR*，abs/2310.05845，2023 年。

Runjin Chen、Tong Zhao、Ajay Kumar Jaiswal、Neil Shah 和 Zhangyang Wang。Llaga：大型语言与图助手。发表于 第*41*届国际机器学习大会，*ICML 2024*，维也纳，奥地利，*2024*年*7*月*21*日至*27*日。OpenReview.net，2024年。

Julian Coda-Forno、Marcel Binz、Zeynep Akata、Matt M. Botvinick、Jane X. Wang 和 Eric Schulz。大语言模型中的元上下文学习。发表于 神经信息处理系统进展第*36*卷：*2023*年神经信息处理系统年会（*NeurIPS 2023*），美国路易斯安那州新奥尔良，*2023*年*12*月*10*日至*16*日，2023年。

Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Let man、Akhil Mathur、Alan Schelten、Amy Yang、Angela Fan、Anirudh Goyal、Anthony Hartshorn、Aobo Yang、Archi Mitra、Archie Sravankumar、Artem Korenev、Arthur Hinsvark、Arun Rao、Aston Zhang、Aur´elien Rodriguez、Austen Gregerson、Ava Spataru、Baptiste Rozi`ere、Bethany Biron、Binh Tang、Bobbie Chern、Charlotte Caucheteux、Chaya Nayak、Chloe Bi、Chris Marra、Chris McConnell、Christian Keller、Christophe Touret、Chunyang Wu、Corinne Wong、Cristian Canton Ferrer、Cyrus Nikolaidis、Damien Allonsius、Daniel Song、Danielle Pintz、Danny Livshits、David Esiobu、Dhruv Choudhary、Dhruv Mahajan、Diego Garcia-Olano、Diego Perino、Dieuwke Hupkes、Egor Lakomkin、Ehab AlBadawy、Elina Lobanova、Emily Dinan、Eric Michael Smith、FilipRadenovic、FrankZhang、Gabriel Synnaeve、Gabrielle Lee、Georgia Lewis Anderson、Graeme Nail、Gr´egoire Mialon、Guan Pang、Guillem Cucurell、Hailey Nguyen、Hannah Korevaar、Hu Xu、Hugo Touvron、Iliyan Zarov、Imanol Arrieta Ibarra、Isabel M. Kloumann、Ishan Misra、Ivan Evtimov、Jade Copet、Jaewon Lee、Jan Geffert、Jana Vranes、Jason Park、Jay Mahadeokar、Jeet Shah、Jelmer van der Linde、Jennifer Billock、Jenny Hong、Jenya Lee、Jeremy Fu、Jianfeng Chi、Jianyu Huang、Jiawen Liu、Jie Wang、Jiecao Yu、Joanna Bitton、Joe Spisak、Jongsoo Park、Joseph Rocca、Joshua Johnstun、Joshua Saxe、Junteng Jia、Kalyan Vasuden Alwala、Kartikeya Upasani、Kate Plawiak、Ke Li、Kenneth Heafield、Kevin Stone等。Llama 3模型群。*CoRR*，abs/2407.21783，2024年。

Taoran Fang、Yunchao Zhang、Yang Yang、Chunping Wang 和 Lei Chen。图神经网络的通用提示调优。载于神经信息处理系统进展第*36*卷：*2023*年神经信息处理系统年会，*NeurIPS 2023*，美国路易斯安那州新奥尔良，*2023*年*12*月*10*日至*16*日，2023年。

Bahare Fatemi、Jonathan Halcrow 和 Bryan Perozzi。像图一样说话：为大语言模型编码图。载于第十二届国际学习表征会议，*ICLR 2024*，奥地利维也纳，*2024*年*5*月*7*日至*11*日。OpenReview.net，2024年。

Zhong Guan、Hongke Zhao、Likang Wu、Ming He 和 Jianpin Fan。LangTopo：通过词元化拓扑建模对齐图的语言描述。*CoRR*，abs/2406.13250，2024年。

Daya Guo、Dejian Yang、Haowei Zhang、Junxiao Song、Peiyi Wang、Qihao Zhu、Runxin Xu、Ruoyu Zhang、Shirong Ma、Xiao Bi、Xiaokang Zhang、Xingkai Yu、Yu Wu、Z. F. Wu、Zhibin Gou、Zhi- hong Shao、Zhuoshu Li、Ziyi Gao、Aixin Liu、Bing Xue、Bingxuan Wang、Bochao Wu、Bei Feng、Chengda Lu、Chenggang Zhao、Chengqi Deng、Chong Ruan、Damai Dai、Deli Chen、Dongjie Ji、Erhang Li、Fangyun Lin、Fucong Dai、Fuli Luo、Guangbo Hao、Guanting Chen、Guowei Li、H. Zhang、Hanwei Xu、Honghui Ding、Huazuo Gao、Hui Qu、Hui Li、Jianzhong Guo、Jiashi Li、Jingchang Chen、Jingyang Yuan、JinhaoTu、J unjie Qiu、Junlong Li、J.L. Cai、JiaqiNi、Jian Liang、Jin Chen、Kai Dong、Kai Hu、Kaichao You、Kaige Gao、Kang Guan、Kexin Huang、Kuai Yu、Lean Wang、Lecong Zhang、Liang Zhao、Litong Wang、Liyue Zhang、Lei Xu、Leyi Xia、Mingchuan Zhang、Minghua Zhang、Minghui Tang、Mingxu Zhou、Meng Li、Miaojun Wang、Mingming Li、Ning Tian、Panpan Huang、Peng Zhang、Qiancheng Wang、Qinyu Chen、Qiushi Du、Ruiqi

Ge、Ruisong Zhang、Ruizhe Pan、Runji Wang、R. J. Chen、R. L. Jin、Ruyi Chen、Shanghao Lu、Shangyan Zhou、Shanhuang Chen、Shengfeng Ye、Shiyu Wang、Shuiping Yu、Shunfeng Zhou、Shuting Pan、S. S. Li、Shuang Zhou、Shaoqing Wu、Tao Yun、Tian Pei、Tianyu Sun、T. Wang、WangdingZeng、Wen Liu、Wenfeng Liang、Wenjun Gao、Wenqin Yu、Wentao Zhang、W.L. Xiao、Wei An、Xiaodong Liu、Xiaohan Wang、Xiaokang Chen、Xiaotao Nie、Xin Cheng、Xin Liu、Xin Xie、Xingchao Liu、Xinyu Yang、Xinyuan Li、Xuecheng Su、Xuheng Lin、X. Q. Li、Xiangyue Jin、Xiaojin Shen、Xiaosha Chen、Xiaowen Sun、Xiaoxiang Wang、Xinnan Song、Xinyi Zhou、Xianzu Wang、Xinxia Shan、Y. K. Li、Y. Q. Wang、Y. X. Wei、Yang Zhang、Yanhong Xu、Yao Li、Yao Zhao、Yaofeng Sun、Yaohui Wang、Yi Yu、Yichao Zhang、Yifan Shi、Yiliang Xiong、Ying He、Yishi Piao、Yisong Wang、Yixuan Tan、Yiyang Ma、Yiyuan Liu、Yongqiang Guo、Yuan Ou、Yuduan Wang、Yue Gong、Yuheng Zou、Yujia He、Yunfan Xiong、Yuxiang Luo、Yuxiang You、Yuxuan Liu、Yuyang Zhou、Y. X. Zhu、Yanping Huang、Yaohui Li、Yi Z heng、Yuchen Zhu、Yunx- ian Ma、Ying Tang、Yukun Zha、Yuting Yan、Z. Z. Ren、Zehui Ren、Zhangli Sha、Zhe Fu、Zhean Xu、Zhenda Xie、Zhengyan Zhang、Zhewen Hao、Zhicheng Ma、Zhigang Yan、Zhiyu Wu、Zihui Gu、Zijia Zhu、Zijun Liu、Zilin Li、Ziwei Xie、Ziyang Song、Zizheng Pan、Zhen Huang、Zhipeng Xu、Zhongyu Zhang 和 Zhen Zhang。DeepSeek-R1 通过强化学习激励大语言模型进行推理。《自然》，645(8081):633–638，2025年9月。

Taicheng Guo、Kehan Guo、Bozhao Nan、Zhenwen Liang、Zhichun Guo、Nitesh V. Chawla、Olaf Wiest 和 Xiangliang Zhang。大语言模型能在化学领域做什么？涵盖八项任务的综合基准。在 神经信息处理系统进展第*36*卷：*2023*年神经信息处理系统年会，*NeurIPS 2023*，美国路易斯安那州新奥尔良，*2023*年*12*月*10*日至*16*日，2023年。

Xiaoxin He、Xavier Bresson、Thomas Laurent、Adam Perold、杨立昆和 Bryan Hooi。善用解释：用于增强文本属性图表示学习的 LLM到LM解释器。在 第十二届国际学习表征会议，*ICLR 2024*，奥地利维也纳，*2024*年*5*月*7*日至*11*日。OpenReview.net，2024a。

Xiaoxin He、Yijun Tian、Yifei Sun、Nitesh V. Chawla、Thomas Laurent、杨立昆、Xavier Bresson 和 Bryan Hooi。G-retriever：面向文本图理解与问答的检索增强生成。发表于神经信息处理系统进展 *38*：*2024*年神经信息处理系统年会，*NeurIPS 2024*，温哥华，不列颠哥伦比亚省，加拿大，*2024*年*12*月*10*日至*15*日，2024b。

Zhenyu Hou、Xiao Liu、Yukuo Cen、Yuxiao Dong、Hongxia Yang、Chunjie Wang 和 唐杰。Graphmae：自监督掩码图自编码器。发表于*KDD '22*：第*28*届 *ACM SIGKDD*知识发现与数据挖掘会议，华盛顿特区，美国，*2022*年*8*月*14*日至*18*日，第594–604页。ACM，2022年。

Edward J. Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang 与 Weizhu Chen。LoRA：大语言模型的低秩适配。载于第十届国际学习表征会议，*ICLR 2022*，线上活动，*2022*年*4*月*25*日至*29*日。OpenReview.net，2022年。

Linmei Hu、Zeyi Liu、Ziwang Zhao、Lei Hou、Liqiang Nie 与李涓子。知识增强预训练语言模型综述。*IEEE*知识与数据工程汇刊，36(4):1413–1430，2024年。

Weihua Hu、Matthias Fey、Marinka Zitnik、Yuxiao Dong、Hongyu Ren、Bowen Liu、Michele Catasta 与 Jure Leskovec。开放图基准：面向图机器学习的数据集。载于神经信息处理系统进展第*33*卷：*2020*年神经信息处理系统年会，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上，2020a年。

Weihua Hu、Bowen Liu、Joseph Gomes、Marinka Zitnik、Percy Liang、Vijay S. Pande 和 Jure Leskovec。图神经网络预训练策略。载于第 *8* 届国际学习表征会议，*ICLR 2020*年，埃塞俄比亚亚的斯亚贝巴，*2020*年 *4* 月 *26* 日至 *30* 日。OpenRe- view.net，2020年。

Xuanwen Huang、Kaiqiao Han、Yang Yang、Dezheng Bao、Quanjin Tao、Ziwei Chai 和 Qi Zhu。图神经网络能成为大语言模型的优秀适配器吗？载于*ACM* 网络会议论文集，*WWW 2024*，新加坡，*2024*年*5*月*13*日至*17*日，第 893–904 页。ACM，2024年。

Aaron Hurst、Adam Lerer、Adam P. Goucher、Adam Perelman、Aditya Ramesh、Aidan Clark、AJ Ostrow、Akila Welihinda、Alan Hayes、Alec Radford、Aleksander Madry、Alex Baker-Whitcomb、Alex Beutel、Alex Borzunov、Alex Carney、Alex Chow、Alex Kirillov、Alex Nichol、Alex Paino、Alex Renzin、Alex Tachard Passos、Alexander Kirillov、Alexi Christakis、Alexis Conneau、Ali Kamali、Allan Jabri、Allison Moyer、Allison Tam、Amadou Crookes、Amin Tootoonchian、Ananya Kumar、Andrea Vallone、Andrej Karpathy、Andrew Braunstein、Andrew Cann、Andrew Codispoti、Andrew Galu、Andrew Kondrich、Andrew Tulloch、Andrey Mishchenko、Angela Baek、Angela Jiang、Antoine Pelisse、Antonia Woodford、Anuj Gosalia、Arka Dhar、Ashley Pantuliano、Avi Nayak、Avital Oliver、Barret Zoph、Behrooz Ghorbani、Ben Leimberger、Ben Rossen、Ben Sokolowsky、Ben Wang、Benjamin Zweig、Beth Hoover、Blake Samic、Bob McGrew、Bobby Spero、Bogo Giertler、Bowen Cheng、Brad Lightcap、Brandon Walkin、Brendan Quinn、Brian Guarraci、Brian Hsu、Bright Kellogg、Brydon Eastman、Camillo Lugaresi、Carroll L. Wainwright、Cary Bassin、Cary Hudson、Casey Chu、Chad Nelson、Chak Li、Chan Jun Shern、Channing Conger、Charlotte Barette、Chelsea Voss、Chen Ding、Cheng Lu、Chong Zhang、Chris Beaumont、ChrisHallacy、Chris Koch、Christian Gibson、Christina Kim、ChristineChoi、Christine McLeavey、Christopher Hesse、Claudia Fischer、Clemens Winter、Coley Czarnecki、Colin Jarvis、Colin Wei、Constantin Koumouzelis 和 Dane Sherburn。GPT-4o 系统卡。*CoRR*，abs/2410.21276，2024年。

Diederik P. Kingma 与 Jimmy Ba。Adam：一种随机优化方法。载于第*3*届国际学习表征会议，*ICLR 2015*，美国加利福尼亚州圣迭戈，*2015*年*5*月*7*日至*9*日，会议论文集，2015年。

Thomas N. Kipf 与 Max Welling。基于图卷积网络的半监督分类。载于第*5*届国际学习表征会议，*ICLR 2017*，法国土伦，*2017*年*4*月*24*日至*26*日，会议论文集。OpenReview.net，2017年。

Yuchen Lei、Yuexin Xiang、Qin Wang、Rafael Dowsley、Tsz Hon Yuen 与 Jiangshan Yu。面向加密货币交易分析的大语言模型：一项比特币案例研究。*CoRR*，abs/2501.18158，2025年。

Xin Li、Weize Chen、Qizhi Chu、Haopeng Li、Zhaojun Sun、Ran Li、Chen Qian、Yiwei Wei、Chuan Shi、Zhiyuan Liu、Maosong Sun 与 Cheng Yang。大语言模型能否像专业人士一样分析图？基准、数据集与模型。载于《神经信息处理系统进展》第*38*卷：*2024*年神经信息处理系统年度会议，*NeurIPS 2024*，加拿大不列颠哥伦比亚省温哥华，*2024*年*12*月*10*日至*15*日，2024a。

Yuhan Li、Peisong Wang、Zhixun Li、Jeffrey Xu Yu 和 Jia Li。ZEROG：研究图中的跨数据集零样本迁移能力。载于第*30*届*ACM SIGKDD*知识发现与数据挖掘会议论文集，*KDD 2024*，巴塞罗那，西班牙，*2024*年*8*月*25*日至*29*日，第1725–1735页。ACM，2024b。

Yuhan Li、Peisong Wang、Xiao Zhu、Aochuan Chen、Haiyun Jiang、Deng Cai、Wai Kin (Victor) Chan 和 Jia Li。Glbench：面向图与大型语言模型的综合基准。载于神经信息处理系统进展第*38*卷：*2024*年神经信息处理系统年会，*NeurIPS 2024*，温哥华，不列颠哥伦比亚省，加拿大，*2024*年*12*月*10*日至*15*日，2024c。

Zhixun Li、Xin Sun、Yifan Luo、Yanqiao Zhu、Dingshuo Chen、Yingtao Luo、Xiangxin Zhou、Qiang Liu、Shu Wu、Liang Wang 和 Jeffrey Xu Yu。GSLB：图结构学习基准。载于神经信息处理系统进展第*36*卷：*2023*年神经信息处理系统年会，*NeurIPS 2023*，新奥尔良，路易斯安那州，美国，*2023*年*12*月*10*日至*16*日，2023年。

Zhucong Li、Jin Xiao、Bowei Zhang、Zhijian Zhou、Qianyu He、Fenglei Cao、Jiaqing Liang 和 Yuan Qi。Chemhts：用于增强化学智能体的分层工具堆叠。*CoRR*，abs/2502.14327，2025年。

Qika Lin、Tianzhe Zhao、Kai He、Zhen Peng、Fangzhi Xu、Ling Huang、Jingying Ma 和 Mengling Feng。《用于无缝整合知识图谱与大语言模型的自监督量化表示》。载于计算语言学协会第*63*届年会论文集（第*1*卷：长篇论文），*ACL 2025*，维也纳，奥地利，*2025*年*7*月*27*日至*8*月*1*日，第13587–13602页。计算语言学协会，2025年。

Hao Liu、Jiarui Feng、Lecheng Kong、Ningyue Liang、Dacheng Tao、Yixin Chen 和 Muhan Zhang。《一以当全：迈向适用于所有分类任务的单一图模型训练》。载于第十二届国际学习表征会议，*ICLR 2024*，维也纳，奥地利，*2024*年*5*月*7*日至*11*日。OpenReview.net，2024年。

Jiawei Liu、Cheng Yang、Zhiyuan Lu、Junze Chen、Yibo Li、Mengmei Zhang、Ting Bai、Yuan Fang、Lichao Sun、Philip S. Yu 和 Chuan Shi。图基础模型：概念、机遇与挑战。*IEEE* 模式分析与机器智能汇刊，47(6):5023–5044，2025a。

Zequn Liu、Wei Zhang、Yingce Xia、Lijun Wu、Shufang Xie、Tao Qin、Ming Zhang 和 刘铁岩。Molxpt：以文本封装分子，实现生成式预训练。载于 第 *61* 届计算语言学协会年会论文集（第 *2* 卷：短文），*ACL 2023*，多伦多，加拿大，*2023* 年 *7* 月 *9* 日至 *14* 日，第 1606–1616 页。计算语言学协会，2023a。

Zhiyuan Liu、Sihang Li、Yanchen Luo、Hao Fei、Yixin Cao、Kenji Kawaguchi、Xiang Wang 和蔡达成。Molca：借助跨模态投影器和单模态适配器进行分子图语言建模。载于*2023*年自然语言处理实证方法会议论文集（*EMNLP 2023*），*2023*年*12*月*6*日至*10*日，新加坡，第15623–15638页。计算语言学协会，2023b。

Zipeng Liu、Likang Wu、Ming He、Zhong Guan、Hongke Zhao 和 Nan Feng。面向语言模型的多视图增强结构图词化。载于*AAAI-25*，由人工智能促进协会赞助，*2025*年*2*月*25*日至*3*月*4*日，费城，宾夕法尼亚州，美国，第24714–24722页。AAAI出版社，2025b。

Ilya Loshchilov 和 Frank Hutter。解耦权重衰减正则化。载于第七届国际学习表征会议，*ICLR 2019*，新奥尔良，路易斯安那州，美国，*2019*年*5*月*6*日至*9*日。OpenReview.net，2019年。

Harry L Morgan。化学结构唯一机器描述的生成——化学文摘社开发的一种技术。化学文献杂志，5(2):107–113，1965年。

OpenAI。GPT-4 技术报告。*CoRR*，abs/2303.08774，2023年。

Long Ouyang、Jeffrey Wu、Xu Jiang、Diogo Almeida、Carroll L. Wainwright、Pamela Mishkin、Chong Zhang、SandhiniAgarwal、Katarina Slama、Alex Ray、John Schulman、Jacob Hilton、Fraser Kelton、Luke Miller、Maddie Simens、Amanda Askell、Peter Welinder、Paul F. Christiano、Jan Leike 和 Ryan Lowe。通过人类反馈训练语言模型遵循指令。载于神经信息处理系统进展第*35*卷：神经信息处理系统年会 *2022*，*NeurIPS 2022*，新奥尔良，路易斯安那州，美国，*2022*年*11*月*28*日至*12*月*9*日，2022年。

Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、JamesBradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga、Alban Desmaison、Andreas K¨ opf、Ed- ward Z. Yang、Zachary DeVito、Martin Raison、Alykhan Tejani、Sasank Chilamkurthy、Benoit Steiner、Lu Fang、Junjie Bai 和 Soumith Chintala。《PyTorch：一种命令式风格的高性能深度学习库》。载于 神经信息处理系统进展第 *32* 卷：神经信息处理系统 *2019* 年年会，*NeurIPS 2019*，*2019* 年 *12* 月 *8—14* 日，加拿大不列颠哥伦比亚省温哥华，第 8024–8035 页，2019年。

Bryan Perozzi、Bahare Fatemi、Dustin Zelle、Anton Tsitsulin、Seyed Mehran Kazemi、Rami Al- Rfou 和 Jonathan Halcrow。《让你的图来表达：为大语言模型编码结构化数据》。*CoRR*，abs/2402.05862，2024年。

Nils Reimers 和 Iryna Gurevych。Sentence-BERT：使用孪生 BERT 网络进行句子嵌入。在*2019*年自然语言处理实证方法会议和第九届国际自然语言处理联合会议（*EMNLP-IJCNLP 2019*）论文集，*2019*年*11*月*3*日至*7*日，中国香港，第 3980–3990 页。计算语言学协会，2019年。

Jiabin Tang、Yuhao Yang、Wei Wei、Lei Shi、Lixin Su、Suqi Cheng、Dawei Yin 和 Chao Huang。GraphGPT：面向大语言模型的图指令调优。载于 第*47*届信息检索研究与发展国际 *ACM SIGIR*会议论文集，*SIGIR 2024*，华盛顿特区，美国，*2024*年*7*月*14*日至*18*日，第 491–500 页。ACM，2024年。

Ross Taylor、Marcin Kardas、Guillem Cucurull、Thomas Scialom、Anthony Hartshorn、Elvis Saravia、Andrew Poulton、Viktor Kerkez 和 Robert Stojnic。《Galactica：面向科学的大语言模型》。*CoRR*，abs/2211.09085，2022年。

Yijun Tian、Huan Song、Zichen Wang、Haozhu Wang、Ziqing Hu、Fang Wang、Nitesh V. Chawla 和 Panpan Xu。利用大语言模型进行图神经提示。载于第三十八届*AAAI*人工智能会议，*AAAI 2024*，第三十六届人工智能创新应用会议，*IAAI 2024*，第十四届人工智能教育进展研讨会，*EAAI 2014*，*2024*年*2*月*20*日至*27*日，温哥华，加拿大，第19080–19088页。AAAIPress，2024年。

Hugo Touvron、Louis Martin、Kevin Stone、Peter Albert、Amjad Almahairi、Yasmine Babaei、Niko- lay Bashlykov、Soumya Batra、Prajjwal Bhargava、Shruti Bhosale、Dan Bikel、Lukas Blecher、Cristian Canton-Ferrer、Moya Chen、Guillem Cucurull、David Esiobu、Jude Fernandes、Jeremy Fu、Wenyin Fu、Brian Fuller、Cynthia Gao、Vedanuj Goswami、Naman Goyal、Anthony Hartshorn、Saghar Hosseini、Rui Hou、Hakan Inan、Marcin Kardas、Viktor Kerkez、Madian Khabsa、Isabel Kloumann、Artem Korenev、Punit Singh Koura、Marie-Anne Lachaux、Thibaut Lavril、Jenya Lee、Diana Liskovich、Yinghai Lu、Yuning Mao、Xavier Martinet、Todor Mihaylov、Pushkar Mishra、Igor Molybog、Yixin Nie、Andrew Poulton、Jeremy Reizenstein、Rashi Rungta、Kalyan Saladi、Alan Schelten、Ruan Silva、Eric Michael S mith、Ranjan Subramanian、Xiaoqing Ellen Tan、Binh Tang、Ross Taylor、Adina Williams、Jian Xiang Kuan、Puxin Xu、Zheng Yan、Iliyan Zarov、Yuchen Zhang、Angela Fan、Melanie Kambadur、Sharan Narang、Aur´elien Rodriguez、Robert Stojnic、Sergey Edunov 和 Thomas Scialom。《Llama 2：开放式基础模型与微调聊天模型》。*CoRR*，abs/2307.09288，2023年。

Petar Velickovic、William Fedus、William L. Hamilton、Pietro Li`o、Yoshua Bengio 和 R. Devon Hjelm。《深度图互信息最大化》。载于第七届国际学习表征会议，*ICLR 2019*，美国路易斯安那州新奥尔良，*2019*年*5*月*6*日至*9*日。OpenReview.net，2019年。

Duo Wang、Yuan Zuo、Fengzhi Li 和 Junjie Wu。《大语言模型作为零样本图学习器：图神经网络表征与大语言模型词元嵌入的对齐》。载于《神经信息处理系统进展》第*38*卷：*2024*年神经信息处理系统年会，*NeurIPS 2024*，加拿大不列颠哥伦比亚省温哥华，*2024*年*12*月*10*日至*15*日，2024a。

Heng Wang、Shangbin Feng、Tianxing He、Zhaoxuan Tan、Xiaochuang Han 和 Yulia Tsvetkov。《语言模型能否用自然语言解决图问题？》。载于《神经信息处理系统进展》第*36*卷：*2023*年神经信息处理系统年会，*NeurIPS 2023*，美国路易斯安那州新奥尔良，*2023*年*12*月*10*日至*16*日，2023年。

Jianing Wang、Junda Wu、Yupeng Hou、Yao Liu、Ming Gao 和 Julian J. McAuley。《Instructgraph：通过以图为中心的指令微调和偏好对齐提升大语言模型》。载于计算语言学协会论文集，*ACL 2024*，泰国曼谷及线上会议，*2024*年*8*月*11*日至*16*日，第 13492–13510 页。计算语言学协会，2024b。

Runze Wang、Mingqi Yang 和 Yanming Shen。《连接分子图与大语言模型》。载于*AAAI-25*，由人工智能促进协会主办，*2025*年*2*月*25*日至*3*月*4*日，美国宾夕法尼亚州费城，第 21234–21242 页。AAAI出版社，2025a。

Zhuang Wang、Linna Zhou、Xuekai Chen、Zhili Zhou 和 Zhongliang Yang。《STLC-KG：融合大规模语言模型与常识知识图谱的社交文本隐写分析方法》。载于*AAAI-25*，由人工智能促进协会主办，*2025*年*2*月*25*日至*3*月*4*日，美国宾夕法尼亚州费城，第 25461–25469 页。AAAI出版社，2025b。

Zhenqin Wu、Bharath Ramsundar、Evan N. Feinberg、Joseph Gomes、Caleb Geniesse、Aneesh S. Pappu、Karl Leswing 和 Vijay S. Pande。《MoleculeNet：分子机器学习的基准》。*CoRR*，abs/1703.00564，2017年。

Nan Xiang、Qilin Wang 和 Mingwei You。使用渐进算法和最短路径近似估计并更新介数中心性。《科学报告》，13(1):17110，2023年。

Ling Yang、Ye Tian、Minkai Xu、Zhongyi Liu、Shenda Hong、Wei Qu、Wentao Zhang、Bin Cui、Muhan Zhang 和 Jure Leskovec。Vqgraph：重新思考图表示空间，以衔接图神经网络与多层感知机。载于第十二届国际学习表征会议，*ICLR 2024*，维也纳，奥地利，*2024*年*5*月*7*日至*11*日。OpenReview.net，2024a。

Mingji Yang、Hanzhi Wang、Zhewei Wei、Sibo Wang 和 Ji-Rong Wen。个性化PageRank计算的高效算法：综述。*IEEE*知识与数据工程汇刊，36(9):4582–4602，2024b。

Ruosong Ye、Caiqi Zhang、Runhui Wang、Shuyuan Xu 和 Yongfeng Zhang。语言是图所需的一切。载于计算语言学协会论文集：*EACL 2024*，圣朱利安斯，马耳他，*2024*年*3*月*17*日至*22*日，第1955–1973页。计算语言学协会，2024年。

Yuning You、Tianlong Chen、Yongduo Sui、Ting Chen、Zhangyang Wang 和 Yang Shen。通过数据增强进行图对比学习。载于神经信息处理系统进展第*33*卷：*2020*年神经信息处理系统年会，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上，2020年。

Haiteng Zhao、Shengchao Liu、Chang Ma、Hannan Xu、Jie Fu、Zhihong Deng、Lingpeng Kong 和 Qi Liu。GIMLET：用于基于指令的分子零样本学习的统一图文本模型。载于神经信息处理系统进展第*36*卷：*2023*年神经信息处理系统年会，*NeurIPS 2023*，美国路易斯安那州新奥尔良，*2023*年*12*月*10*日至*16*日，2023a。

Jianan Zhao、Le Zhuo、Yikang Shen、Meng Qu、Kai Liu、Michael M. Bronstein、Zhaocheng Zhu 和 Jian Tang。GraphText：文本空间中的图推理。*CoRR*，abs/2310.01089，2023b。

Zihan Zhao、DaMa、Lu Chen、Liangtai Sun、Zihao Li、Yi Xia、Bo Chen、Hongshen Xu、Zichen Zhu、Su Zhu 等人。开发 chemdfm，将其打造为面向化学的大型语言基础模型。《细胞报告：物理科学》，6(4)，2025年。

A 实验设置的补充细节

A.1 数据集

MoleculeNet 是一个基准套件，旨在为数据集制定标准，并为计算化学中的机器学习模型建立一致的评估框架（Wu 等人，2017年）。以下五个数据集（BBBP、BACE、HIV、CinTox、Tox21）均来自该基准。我们遵循 GIMLET（Zhao 等人，2023a）中规定的数据集选择和划分标准。对于下游任务，我们按 8:1:1 的比例划分数据集，并报告测试集上的结果。实验设置包括单任务和多任务，同一数据集中的多个任务仅在任务标签上有所不同，而使用的分子完全相同。数据统计信息见表 6。

BBBP BBBP（血脑屏障通透性）数据集是 MoleculeNet 数据集的一部分。具体而言，该数据集用于预测给定化合物能否穿过血脑屏障（BBB）。BBBP 包含化合物的二分类标签，用于标示其能否穿过血脑屏障：能穿过 BBB 的化合物标记为 1，不能穿过的标记为 0。这些化合物以 SMILES 表达式表示。数据集包含 2,039 种化合物，具有多种化学特征，包括原子类型、键类型、分子量等分子性质。不过，在此任务中，这些特征不会作为辅助信息用于模型输入。该数据集广泛用于药物开发研究。这对于设计治疗脑部疾病的药物至关重要，因为只有能够穿过 BBB 的化合物才能对中枢神经系统产生作用。

表 6：图级分类数据集统计。

	# 总样本数			正样本比例(%)			负样本比例(%)		
数据集	训练集	验证集	test	训练集	验证集	test	训练集	验证集	test
BACE_分类	1210	151	152	42.6	55.6	60.5	57.4	44.4	39.5
BBBP_阳性_阴性	1631	204	204	82.2	54.9	52.5	17.8	45.1	47.6
ClinTox_临床毒性	1184	148	148	8.0	4.7	6.8	92.0	95.3	93.2
ClinTox FDA 批准	1184	148	148	93.3	96.0	93.9	6.7	4.1	6.1
HIV_HIV_活性	32896	4112	4112	3.8	2.0	3.2	96.3	98.0	96.8
Tox21_NR-AR	6258	782	783	4.0	4.0	3.5	96.0	96.0	96.6
Tox21_NR-AR-LBD	6258	782	783	3.1	3.2	2.4	96.9	96.8	97.6
Tox21_核受体芳香烃受体	6258	782	783	9.4	11.1	11.8	90.6	88.9	88.3
Tox21_核受体芳香化酶	6258	782	783	3.3	5.8	6.0	96.7	94.3	94.0
Tox21_核受体雌激素受体	6258	782	783	10.3	9.6	8.9	89.7	90.4	91.1
Tox21_核受体雌激素受体配体结合域	6258	782	783	4.8	3.7	2.7	95.2	96.3	97.3
Tox21 NR-PPAR-γ	6258	782	783	2.1	4.1	2.8	97.9	95.9	97.2
Tox21_SR-ARE	6258	782	783	11.5	13.6	15.1	88.5	86.5	84.9
Tox21_SR-ATAD5	6258	782	783	3.1	4.5	4.2	96.9	95.5	95.8
Tox21_SR-HSE	6258	782	783	4.5	5.6	6.0	95.5	94.4	94.0
Tox21_SR-MMP	6258	782	783	11.4	14.2	12.3	88.6	85.8	87.7
Tox21_SR-p53	6258	782	783	4.4	9.6	9.2	95.6	90.4	90.8

BACE MoleculeNet 中的 BACE数据集旨在进行二分类，目标是预测某种分子是否会与人类 β-分泌酶结合。该酶是一种天冬氨酸蛋白酶，在周围神经细胞髓鞘的形成过程中发挥重要作用。在 BACE 二分类任务中，标签 1 表示该分子是 BACE抑制剂，而标签 0 表示该分子不是 BACE抑制剂。该数据集包含 1513 个分子，每个分子都具有分子特征，即其 SMILES 表达式。它广泛用于评估药物发现中的机器学习模型。

HIV MoleculeNet 中的 HIV 数据集旨在预测药物能否抑制 HIV 复制。该数据集包含 41,120 个分子及其特征，例如 SMILES 表达式，以及相应的抗 HIV 病毒活性。此任务属于二分类：标签为 1 表示给定药物分子能够抑制 HIV 复制，标签为 0 则表示不能。该数据集广泛用于评估药物发现和分子生物信息学领域的机器学习模型，尤其是 HIV 研究和药物开发方面的模型。

ClinTox ClinTox数据集源自药物发现研究，并收录于MoleculeNet。它聚焦于临床试验结局与药物安全性评价，包含1480种化合物，并标注了两项二分类任务。其中一项任务FDA批准用于预测药物分子是否获得美国食品药品监督管理局（FDA）批准，标签1表示FDA批准药物，0表示未获批准药物。另一项任务CT-TOX用于判断药物分子是否因毒性导致临床试验失败，1表示失败，0表示成功并继续。该数据集广泛用于毒性预测和药物安全性评估。

Tox21 Tox21 数据集源自“21世纪毒理学”（Tox21）计划，并收录于 MoleculeNet。该数据集旨在评估化合物对 12 个生物学靶点的人体毒性。它包含 7,823 个分子，每个分子均标注了 12 个二分类标签，用于指示该化合物是否会激活或抑制特定靶点。具体而言，核受体（NR）任务包括 NR-AR、NR-AR-LBD、NR-AhR、NR-芳香化酶、NR-ER、NR-ER-LBD 和 NR-PPAR-gamma，每项任务都需要预测分子是否会激活或抑制相应受体。应激反应（SR）任务包括 SR-ARE、SR-ATAD5、SR-HSE、SR-MMP 和 SR-p53，用于评估化合物是否会诱发抗氧化信号传导、DNA损伤、热休克、线粒体功能障碍或 p53介导的应激通路等反应。所有任务均构建为二分类问题，其中 1 表示激活，0 表示未激活。该数据集已成为计算毒理学和药物安全性评价中广泛采用的基准。

A.2 评估指标

为定量评估图分类结果，我们采用 AUC-ROC（受试者工作特征曲线下面积）评估二元图分类任务的性能，因为大多数任务都存在显著的类别不平衡。AUC-ROC 通过在不同阈值下绘制真正例率（TPR）与假正例率（FPR），衡量模型区分不同类别的能力。该值越高，模型的判别能力越强。

在实际计算 AUC-ROC 指标时，最大输出标记长度限制为 100。为生成输出，将大语言模型的采样参数设为温度 0.6，以兼顾随机性与连贯性，同时将词元选择限制在累积概率质量为 90% 的范围内。随后，我们解析生成的文本以提取答案。具体而言，我们通过定义正例集（例如[“是”“正确”“活跃”“已批准”]）和负例集（例如[“否”“错误”“不活跃”“已拒绝”“未批准”]）来进行匹配。如果生成文本中出现正例集中的任意词语或短语，则将预测赋值为 1；否则，如果出现负例集中的任意词语或短语，则赋值为 0。为避免子串冲突（例如，肯定短语嵌入否定表达中），我们规定匹配优先级：先检查负例，只有在未匹配到负例时才考虑正例。最后，将所得预测与真实标签（0/1）配对，以计算 AUC-ROC。

对于多任务评估，报告的均值是对各个任务的平均表现取平均得到的。假设各任务之间相互独立，则总体标准差为各任务标准差平方的均值的平方根。

为定量评估节点分类结果，我们采用准确率和 F1分数作为标准化指标。准确率是正确分类的节点数与节点总数之比，可简明地衡量整体分类表现，其计算公式如下：

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} I(y_i = \hat{y}_i)}{N},$$

其中， $I(y_i = \hat{}y_i)$ 是一个指示函数：当真实标签 y_i 与预测标签 $\hat{}y_i$ 匹配时取值为 1，否则取值为 0； N 是节点总数。然而，在处理分类任务时，准确率可能不足以全面评估表现。因此，我们还使用 F1分数，即精确率和召回率的调和平均。精确率衡量所有正类预测中真正例所占的比例，而召回率则衡量所有实际正例中被识别出的真正例所占的比例。F1分数综合权衡这两项指标，提供更可靠的性能度量。给定混淆矩阵计数（TP、FP、TN、FN），其中 TP 为真正例的数量，FP 为假正例的数量，TN 为真负例的数量，FN 为假负例的数量。每个二元任务的 F1分数计算如下：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

在多分类设置中，计算精确率和召回率时，微平均F1分数会汇总所有类别的计数。令 TP_k、FP_k 和 FN_k 分别表示类别 k 的真正例、假正例和假负例。定义

$$\text{Precision}_{\text{micro}} = \frac{\sum_k \text{TP}_k}{\sum_k (\text{TP}_k + \text{FP}_k)}, \qquad \text{Recall}_{\text{micro}} = \frac{\sum_k \text{TP}_k}{\sum_k (\text{TP}_k + \text{FN}_k)},$$

and

$$\text{F1-score}_{\text{micro}} = \frac{2\,\text{Precision}_{\text{micro}}\,\text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}}.$$

微平均F1反映了整体上按样本频率加权的性能。本文所有 F1分数评估均使用微平均F1。

在实际计算准确率时，我们还将最大输出标记长度限制为 100。若真实标签出现在生成输出中，则将该预测计为正确，并报告正确预测数量占样本总数的比例。对于 F1 分数，我们通过扫描输出提取首个匹配项作为预测标签（若未找到，则使用未知），然后根据预测标签和真实标签计算多个类别上的微平均F1。我们报告了本方法三次运行的平均准确率和 F1分数。

A.3 基线方法

图级分类的基线方法 我们将非大语言模型方法和基于大语言模型的方法都纳入图分类任务的基线方法。非大语言模型方法包括图提示学习方法 GPF（Fang 等人，2023年），其预训练图神经网络模型为 AttrMasking（Hu 等人，2020年）和 GCL（You 等人，2020年）；还包括图语言调优方法 MolCA（Liu 等人，2023b）及其变体 MolCA-GS。对于基于大语言模型的方法，我们将这些方法归纳如下。

- 零样本推理：将文本图属性和指令提示直接输入大语言模型，利用其预训练知识和推理能力（Hu 等，2024年）。
- 少样本学习：将文本图属性、指令提示以及包含完整文本属性和真实标签的示例直接输入大语言模型，利用其上下文学习能力（Coda-Forno 等人，2023年），即少样本示例可以引导大语言模型理解任务。少样本示例的采样策略包括随机采样，以及基于 Morgan 相似度选择相似样本（Morgan，1965年）。在我们的设置中，我们采用 8样本学习（Guo 等人，2023年），每个类别选取 4 个样本。
- 监督微调：通过低秩适配（Hu 等，2022）进行监督微调，使模型适应下游任务。在我们的设置中，我们将所有任务的数据合并为训练集，并使用联合训练的大语言模型评估每项任务的测试集划分。

- 软提示：采用一个投影器，即一个带有 ReLU 激活的两层多层感知机（MLP），将图嵌入映射到大语言模型空间中，仅训练 MLP，同时冻结大语言模型。我们将大语言模型对 token ”True” 预测的 logit 与真实标签（0/1）进行比较，并使用交叉熵损失，该过程应用于生成的第一个 token。

节点级分类的基线方法我们针对三类基线方法进行节点分类评估：(i) 图自监督学习（SSL）方法，这类方法学习可迁移的结构表示（Li 等人，2023年），代表方法包括 DGI（Velickovic 等人，2019年）和 GraphMAE（Hou 等人，2022年）；(ii) 纯大语言模型基线，仅依赖语义信息，例如 LLaMA3-70B（Dubey 等人，2024年）、GPT-3.5-turbo（Ouyang 等人，2022年）、GPT-4o（Hurst 等人，2024年）和 DeepSeek-chat（Guo 等人，2025年）；(iii) 大语言模型–图集成框架，在基于大语言模型的流程中利用结构和语义信息进行节点分类。Emb w/ NA 是一种简单的免训练基线，结合结构和语义线索进行零样本推理（Li 等人，2024c）。我们还纳入了两个先进的集成框架：OFA（Liu 等人，2024年），该方法将所有节点和边转换为人类可读的文本形式，将异构领域编码到共享语言空间中，并通过插入特定于任务的提示子结构来适应下游任务；以及 ZeroG（Li 等人，2024b），该方法使用语言模型编码节点属性和类别描述，并采用基于提示的子图采样与轻量微调，以应对跨数据集零样本迁移。总体而言，这些基线方法涵盖了仅结构、仅语义以及结构–语义联合范式。

部分基线方法的结果取自其他研究。表 1 中，GPT-4 的 AUC-ROC 结果来自 Zhao 等人（2025年）；GPT-4o、Deepseek-R1 和 LLaMA3-70B 的结果来自 Li 等人（2025年）；Galactica-120B 的结果来自 Liu 等人（2023a）；GIMLET 的结果来自 Zhao 等人（2023a）；GPF-属性掩码、GPF-GCL、MolCA-S 和 MolCA-GS 的结果来自 Wang 等人（2025a）。表3 中所有基线方法的准确率和 F1分数结果均取自 Li 等人（2024c）。

B 实现的补充细节

B.1 模型细节与训练超参数

我们使用 SentenceTransformers all-MiniLM-L6-v2（Reimers 和 Gurevych，2019年）对新分配的结构属性进行编码，以获得初始节点特征。我们在结构分词器中采用双层图卷积网络作为编码器，并将结构词汇表的大小设为 256。我们选用两种广泛使用的大语言模型 LLaMA2-7B-chat（Touvron 等人，2023年）和 Llama-3.2-3B-Instruct（Dubey 等人，2024年）作为大语言模型骨干网络。所有模型均使用 PyTorch（Paszke 等人，2019年）实现。在结构分词器训练阶段，我们首先以学习率 1×10^{-2} 在重建任务上对图卷积网络进行预训练，作为预热。随后，我们使用两个参数组优化整个结构分词器：学习率为 5×10^{-2} 的图卷积网络，以及学习率为 0.5 的结构词元离散嵌入，使用

Adam优化器（Kingma与Ba，2015年）。在与混合问答进行词元对齐的阶段以及任务特定微调阶段，我们以学习率 5×10^{-4} 和 LoRA 秩 16 对新添加的词元嵌入和大语言模型骨干网络进行优化，使用AdamW优化器（Loshchilov与Hutter，2019年）。

B.2 下游应用详述

在通过混合问答完成词元对齐后，我们对与新增结构词元相对应的嵌入以及大语言模型骨干网络的参数进行微调，以用于下游任务。由于大多数任务的训练集中存在严重的类别不平衡现象（见表 6），我们在微调期间采用了各种数据平衡技术。对于某些任务，我们还与其他任务联合训练大语言模型。具体而言，对于 LLaMA3-3B 模型，在 BBBP p np、ClinTox FDA 批准、ClinTox 临床毒性和 HIV 活性任务中，我们通过复制少数类样本，使其数量与多数类样本相匹配，从而按 1:1 的比例进行类别平衡；对于 BACE 分类，则保留原始训练数据分布。每项任务均单独微调，并使用第三个 epoch 的检查点进行评估。对于 Tox21 基准，我们使用全部 17 项任务的原始训练数据进行联合微调，其中包括来自其他四个基准的数据，并使用第二个 epoch 的检查点评估模型在 Tox21 的 12 个子任务上的表现。对于 LLaMA2-7B 模型，Tox21 NR-ER 和 Tox21 SR-HSE 使用未经平衡的原始训练数据；Tox21 NR-AR-LBD、Tox21 NR-ER-LBD、Tox21 SR-ARE、ClinTox FDA 批准、ClinTox 临床毒性和 HIV 活性任务则采用上述 1:1 的平衡比例。对于 BBBP p np、Tox21 NR-AR、Tox21 NR-AhR 和 Tox21 NR-PPAR-gamma，我们对数据进行重采样，使少数类样本数调整为多数类样本数的五分之一；对于BACE 分类，则保留原始分布。所有任务均单独微调，并基于第二个 epoch 的检查点进行评估，但 Tox21 NR-ER-LBD 使用第三个 epoch 的检查点。对于其余 Tox21 任务（NR-芳香化酶、SR-ATAD5、SR-MMP、SR-p53），我们使用全部 17 项任务的未经平衡的原始训练数据进行联合微调，并使用第三个 epoch 的检查点评估这四项任务。值得注意的是，我们没有对验证集和测试集的分布做任何调整。

B.3 扩展至节点级分类

为将我们的方法扩展到节点级任务，我们通过为每个目标节点采样邻域结构来构建自我图，从而能够在标准图范式下进行处理。对于每个节点，我们采样其二跳自我图，并使用在图任务上训练的拓扑感知结构分词器，将其映射为节点级结构令牌。我们不再需要在构建的自我图中添加全局节点。选择与目标节点对应的特定 <SO$\mathcal{G}_k$> 作为结构令牌，以表示目标节点的局部邻域拓扑，并将其插入文本形式的大语言模型输入中，用于下游节点的分类。

B.4 提示设计 <SO$\mathcal{G}_k$>

我们将结构词元融入任务提示和文本图属性中，以构建完整的大语言模型输入提示，从而通过生成式推理获得类别标签。每项任务的详细大语言模型输入提示见相应方框（第 21 至 26 页）。

C 补充案例研究

我们提供了更多示例，说明不同的结构词元如何影响第5.3节所讨论的大语言模型输出概率，如图7、图8和图9所示。此外，我们还考察了数据集训练集划分（而非第5.4节中的测试集划分）中结构词元的选择及其对应结构的更多示例，如图10和图11所示。

D 大语言模型的使用

本文的核心研究将大语言模型直接作为所提出流程中不可或缺的组成部分，因为我们的目标是提升大语言模型的图推理能力并减少结构幻觉。具体而言，我们在大语言模型的输入空间中添加特殊结构词元，以便更准确、简洁地理解图的拓扑结构。在图基准评估中，大语言模型充当主要推理模型来生成答案，评估任务包括图分类和节点分类。

必须澄清的是，大语言模型仅用于这一操作角色，并未参与本文的构思、概念化或撰写。研究思路与写作完全出自我们自己。

用于 BBBP p np 的大语言模型输入提示

< |文本开始| >< |开始标题 id| > 系统< |结束标题 id| >你是一个乐于助人且可靠的助手。< |eot id| >< |文本开始| >< |开始标题 id| >用户< |结束标题 id| >你是一名化学专家。请根据给定分子的分子结构（标记）和 SMILES 表达式，将其归入正确类别。每个结构标记代表一种独特的图模式（例如，一类相似的分子图）。[分子] *smiles expression of target molecule*[结构标记] <SO$\mathcal{G}_k$>[任务] 该分子是否具有血脑屏障通透性（BBB 穿透）？ BBB 可通透为 True，不可通透为 False。请从以下两个选项中输出完整的正确答案：1. True 2. False[答案]< |eot id| >< |开始标题 id| >大语言模型助手< |结束标题 id| >

用于 Tox21 NR-AR 的大语言模型输入提示

< |文本开始| >< |起始标头 ID| > system< |结束标头 ID| >你是一位乐于助人且可靠的助手。< |消息结束 ID| >< |文本开始| >< |起始标头 ID| >user< |结束标头 ID| >你是一位化学专家。请根据给定分子的分子结构（标记）和 SMILES 表达式，将其归类到正确的类别。每个结构标记代表一种独特的图模式（例如，一类相似的分子图）。[分子] *smiles expression of target molecule*[结构标记] <SO$\mathcal{G}_k$>[任务] 该分子是否会激活雄激素受体（NR-AR）？ 活性分子为 True，非活性分子为 False。请从以下两个选项中输出完整且正确的答案：1. True 2. False[答案]< |消息结束 ID| >< |起始标头 ID| >大语言模型助手< |结束标头 ID| >

Tox21 NR-AR-LBD 的大语言模型输入提示

< |begin of text| > < |start header id| > system< |end header id| >你是一位乐于助人且可靠的助手。< |eot id| >< |begin of text| >< |start header id| >user< |end header id| >你是一位化学专家。请根据给定分子的分子结构（标记）和 SMILES 表达式，将其归入正确类别。每个结构标记都代表一种独特的图模式（例如，一类相似的分子图）。[分子] *smiles expression of target molecule*[结构标记] <SO$\mathcal{G}_k$>[任务] 该分子是否会激活雄激素受体（NR- AR-LBD）的配体结合域？ 活性为 True，无活性为 False。请从以下两个选项中输出完整且正确的答案：1. True 2. False[答案]< |eot id| >< |start header id| >大语言模型助手< |end header id| >

Tox21 核受体芳香烃受体的大语言模型输入提示

< |begin of text | > < |start header id | > system< |end header id | >你是一位乐于助人且可靠的助手。< |eot id | >< |begin of text | >< |start header id | >user< |end header id | >你是一位化学专家。请根据给定分子的分子结构（标记）和 SMILES 表达式，将其归入正确类别。每个结构标记都代表一种独特的图模式（例如，一类相似的分子图）。[分子] *smiles expression of target molecule*[结构标记] <SO$\mathcal{G}_k$>[任务] 该分子是否会激活芳香烃受体（NR-AhR）？ 活性为 True，无活性为 False。请从以下两个选项中输出完整且正确的答案：1. True 2. False[答案]< |eot id | >< |start header id | >大语言模型助手< |end header id | >

Tox21 核受体芳香化酶的大语言模型输入提示

< |文本开始 | >< |起始标题 id | > 系统< |结束标题 id | >你是一个乐于助人且可靠的助手。< |eot id | >< |文本开始 | >< |起始标题 id | >用户< |结束标题 id | >你是一名化学专家。请根据给定分子的分子结构（标记）和 SMILES 表达式，将其归入正确类别。每个结构标记代表一种独特的图模式（例如，一类相似的分子图）。[分子] *smiles expression of target molecule*[结构标记] <SO$\mathcal{G}_k$>[任务] 该分子是否抑制芳香化酶（NR-芳香化酶）？ 活性对应 True，非活性对应 False。请从以下两个选项中输出完整的正确答案：1. True 2. False[答案]< |eot id | >< |起始标题 id | >大语言模型助手< |结束标题 id | >

Tox21 核受体雌激素受体的大语言模型输入提示

< |begin of text | > < |start header id | > system< |end header id | >你是一位乐于助人且可靠的助手。< |eot id | >< |begin of text | >< |start header id | >user< |end header id | >你是一位化学专家。请根据给定分子的分子结构（标记）和 SMILES 表达式，将其归入正确的类别。每个结构标记都代表一种独特的图模式（例如，一类相似的分子图）。[分子] *smiles expression of target molecule*[结构标记] <SO$\mathcal{G}_k$>[任务] 该分子是否会激活雌激素受体（NR-ER）？ 激活为 True，未激活为 False。请从以下两个选项中选出完整且正确的答案：1. True 2. False[答案]< |eot id | >< |start header id | >大语言模型助手< |end header id | >

Tox21 NR-ER-LBD 的大语言模型输入提示

<|begin of text|><|start header id|>system<|end header id|>你是一位乐于助人且可靠的助手。<|eot id|><|begin of text|><|start header id|>user<|end header id|>你是一位化学专家。请根据给定分子的分子结构（标记）和 SMILES 表达式，将其归入正确的类别。每个结构标记都代表一种独特的图模式（例如，一类相似的分子图）。[分子] *smiles expression of target molecule*[结构标记] <SO$\mathcal{G}_k$>[任务] 该分子是否会激活雌激素受体（NR-ER- LBD）的配体结合域？激活为 True，未激活为 False。请从以下两个选项中选出完整且正确的答案：1. True 2. False[答案]<|eot id|><|start header id|>大语言模型助手<|end header id|>

Tox21 NR-PPAR-γ 的大语言模型输入提示

<|文本开始|><|头部开始 id|>系统<|头部结束 id|>你是一位乐于助人且可靠的助手。<|消息结束 id|><|文本开始|><|头部开始 id|>用户<|头部结束 id|>你是一位化学专家。请根据给定分子的分子结构（标记）和 SMILES 表达式，将其归类到正确的类别中。每个结构标记都表示一种独特的图模式（例如，一类相似的分子图）。[分子] *smiles expression of target molecule*[结构标记] <SO$\mathcal{G}_k$>[任务] 该分子是否会激活过氧化物酶体增殖物激活受体γ（NR- PPAR-gamma）？激活时为 True，不激活时为 False。请从以下两个选项中输出完整的正确答案：1. True 2. False[答案]<|消息结束 id|><|头部开始 id|>大语言模型助手<|头部结束 id|>

Tox21 SR-ARE 的大语言模型输入提示

<|文本开始|><|起始标题id|>系统<|结束标题id|>你是一个乐于助人且可靠的助手。<|回合结束id|><|文本开始|><|起始标题id|>用户<|结束标题id|>你是一名化学专家。请根据给定分子的分子结构（标记）和SMILES 表达式，将其归入正确类别。每个结构标记代表一种独特的图模式（例如，一类相似的分子图）。[分子] *smiles expression of target molecule*[结构标记] <SO$\mathcal{G}_k$>[任务] 该分子是否会激活抗氧化反应元件（SR-ARE）？激活为True，未激活为False。请从以下两个选项中输出完整的正确答案：1. True 2. False[答案]<|回合结束id|><|起始标题id|>大语言模型助手<|结束标题id|>

Tox21 SR-ATAD5 的大语言模型输入提示

<|文本开始|><|起始标题id|>系统<|结束标题id|>你是一个乐于助人且可靠的助手。<|回合结束id|><|文本开始|><|起始标题id|>用户<|结束标题id|>你是一名化学专家。请根据给定分子的分子结构（标记）和SMILES 表达式，将其归入正确类别。每个结构标记代表一种独特的图模式（例如，一类相似的分子图）。[分子] *smiles expression of target molecule*[结构标记] <SO$\mathcal{G}_k$>[任务] 该分子是否会激活ATAD5信号（SR-ATAD5）？激活为True，未激活为False。请从以下两个选项中输出完整的正确答案：1. True 2. False[答案]<|回合结束id|><|起始标题id|>大语言模型助手<|结束标题id|>

Tox21 SR-HSE 的大语言模型输入提示

<|文本开始|><|头部标识开始|>系统<|头部标识结束|>你是一位乐于助人且值得信赖的助手。<|消息结束|><|文本开始|><|头部标识开始|>用户<|头部标识结束|>你是一位化学专家。请根据给定分子的分子结构（标记）和 SMILES 表达式，将其分类到正确的类别中。每个结构标记都代表一种独特的图模式（例如，一类相似的分子图）。[分子] *smiles expression of target molecule*[结构标记] <SO$\mathcal{G}_k$>[任务] 该分子是否会激活热休克元件（SR-HSE）？激活时为 True，未激活时为 False。请从以下两个选项中输出完整的正确答案：1. True 2. False[答案]<|消息结束|><|头部标识开始|>大语言模型助手<|头部标识结束|>

Tox21 SR-MMP 的大语言模型输入提示

<|begin of text|><|start header id|>system<|end header id|>你是一位乐于助人且可靠的助手。<|eot id|><|begin of text|><|start header id|>user<|end header id|>你是一位化学专家。请根据给定分子的分子结构（标记）和 SMILES 表达式，将其归类到正确的类别。每个结构标记都代表一个独特的图模式（例如，一类相似的分子图）。[分子] *smiles expression of target molecule*[结构标记] <SO$\mathcal{G}_k$>[任务]该分子是否会激活线粒体膜电位应激反应（SR-MMP）？激活为 True，未激活为 False。请从以下两个选项中输出完整且正确的答案：1. True 2. False[答案]<|eot id|><|start header id|>大语言模型助手<|end header id|>

Tox21 SR-p53 的大语言模型输入提示

< |begin of text | > < |start header id | > system< |end header id | >你是一位乐于助人且可靠的助手。< |eot id | >< |begin of text | >< |start header id | >user< |end header id | >你是一位化学专家。请根据给定分子的分子结构（标记）和 SMILES 表达式，将其归类到正确的类别。每个结构标记都代表一个独特的图模式（例如，一类相似的分子图）。[分子] *smiles expression of target molecule*[结构标记] <SO$\mathcal{G}_k$>[任务] 该分子是否会激活 p53 应激反应通路（SR-p53）？ 激活为 True，未激活为 False。请从以下两个选项中输出完整且正确的答案：1. True 2. False[答案]< |eot id | >< |start header id | >大语言模型助手< |end header id | >

ClinTox FDA 批准的大语言模型输入提示

< |文本开始 | >< |开始标头 ID | > 系统< |结束标头 ID | >你是一位乐于助人且可靠的助手。< |消息结束 ID | >< |文本开始 | >< |开始标头 ID | >用户< |结束标头 ID | >你是一位化学专家。请根据给定分子的分子结构（标记）和 SMILES 表达式，将其分类到正确的类别。每个结构标记都代表一个独特的图模式（例如，一类相似的分子图）。[分子] *smiles expression of target molecule*[结构标记] <SO$\mathcal{G}_k$>[任务] 该分子是否已获 FDA 批准？已获 FDA 批准则为 True，未获批准则为 False。请从以下两个选项中输出完整且正确的答案：1. True 2. False[答案]< |消息结束 ID | >< |开始标头 ID | >大语言模型助手< |结束标头 ID | >

ClinTox 临床毒性的大语言模型输入提示

< |文本开始 | >< |标头 ID 开始 | > 系统< |标头 ID 结束 | >你是一名有帮助且可靠的助手。< |文本结束 ID | >< |文本开始 | >< |标头 ID 开始 | >用户< |标头 ID 结束 | >你是一名化学专家。请根据给定分子的分子结构（标记）和 SMILES 表达式，将其分类到正确的类别中。每个结构标记代表一种独特的图模式（例如，一类相似的分子图）。[分子] *smiles expression of target molecule*[结构标记] <SO$\mathcal{G}_k$>[任务] 该分子是否与临床毒性相关？对于具有临床毒性的分子，回答 True；对于无毒的分子，回答 False。请从以下两个选项中输出完整且正确的答案：1. True 2. False[答案]< |文本结束 ID | >< |标头 ID 开始 | >大语言模型助手< |标头 ID 结束 | >

HIV 活性的输入提示

<|文本开始|>≥<|标头 ID 开始|>系统<|标头 ID 结束|>你是一名有帮助且可靠的助手。<|文本结束 ID|><|文本开始|><|标头 ID 开始|>用户<|标头 ID 结束|>你是一名化学专家。请根据给定分子的分子结构（标记）和 SMILES 表达式，将其分类到正确的类别中。每个结构标记代表一种独特的图模式（例如，一类相似的分子图）。[分子] *smiles expression of target molecule*[结构标记] <SO$\mathcal{G}_k$>[任务] 该分子是否能抑制 HIV复制？对于能够抑制 HIV 的活性分子，回答 True；对于无活性的分子，回答 False。请从以下两个选项中输出完整且正确的答案：1. True 2. False[答案]<|文本结束 ID|><|标头 ID 开始|>大语言模型助手<|标头 ID 结束|>

BACE 分类的输入提示

<|文本开始|>≥<|标题开始 id|>系统<|标题结束 id|>你是一名有帮助且可靠的助手。<|文本结束 id|><|文本开始|><|标题开始 id|>用户<|标题结束 id|>你是一名化学专家。请根据给定分子的分子结构（标记）和 SMILES 表达式，将其分类到正确的类别中。每个结构标记代表一种独特的图模式（例如，一类相似的分子图）。[分子] *smiles expression of target molecule*[结构标记] <SO$\mathcal{G}_k$>[任务] 该分子与 β-分泌酶1 的结合结果是真是假？请从以下两个选项中输出完整且正确的答案：1. 真 2. 假[答案]<|文本结束 id|><|标题开始 id|>大语言模型助手<|标题结束 id|>

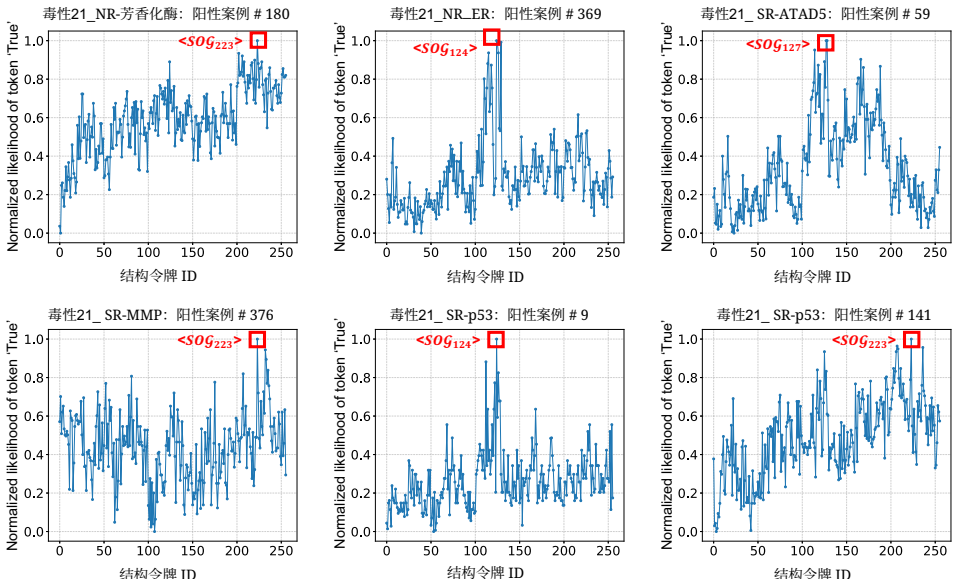

图7：在Tox21数据集中选择不同结构词元时性能变化的示例。

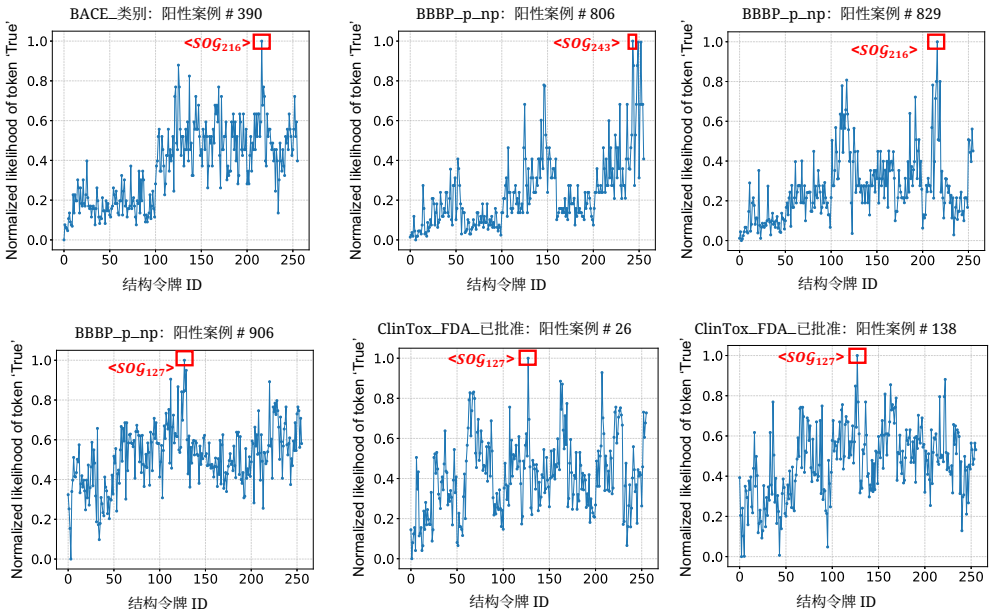

图8：在BBBP、BACE和ClinTox数据集中选择不同结构词元时的性能变化示例。

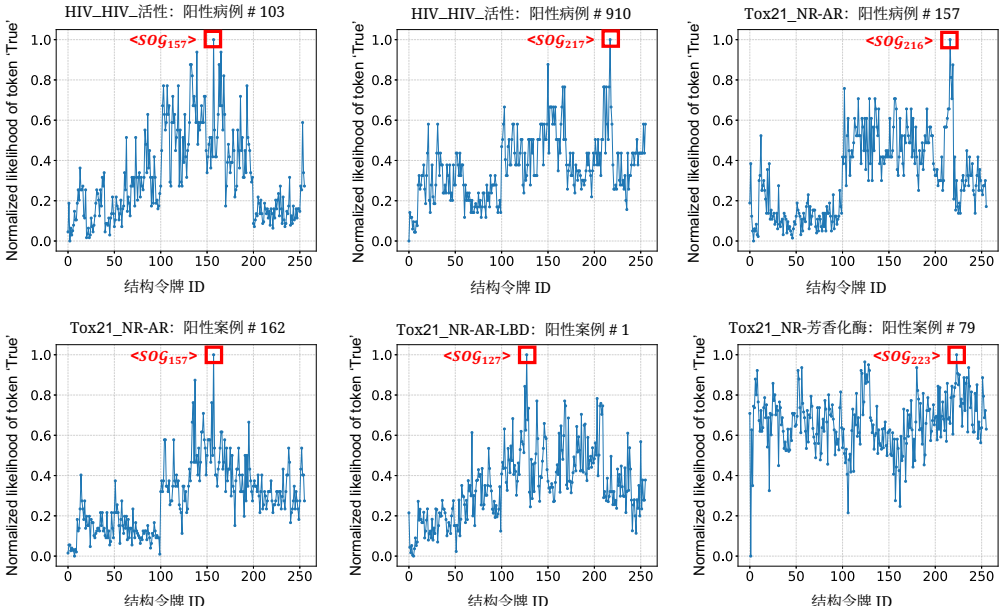

Figure 9：选择不同结构词元时的性能变化示例
在Tox21
1和HIV数据集。

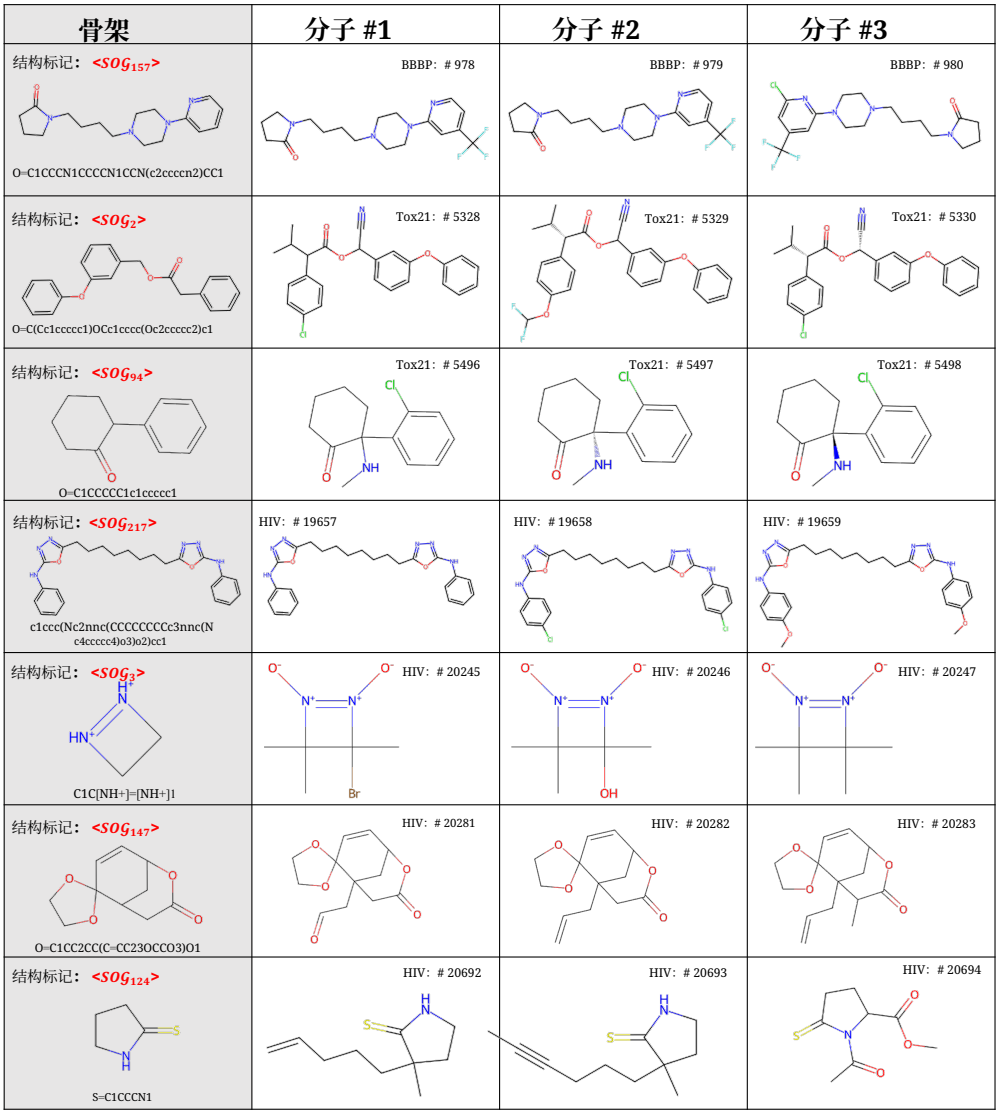

图10：训练集中共享相同骨架且使用相同结构标记的分子。

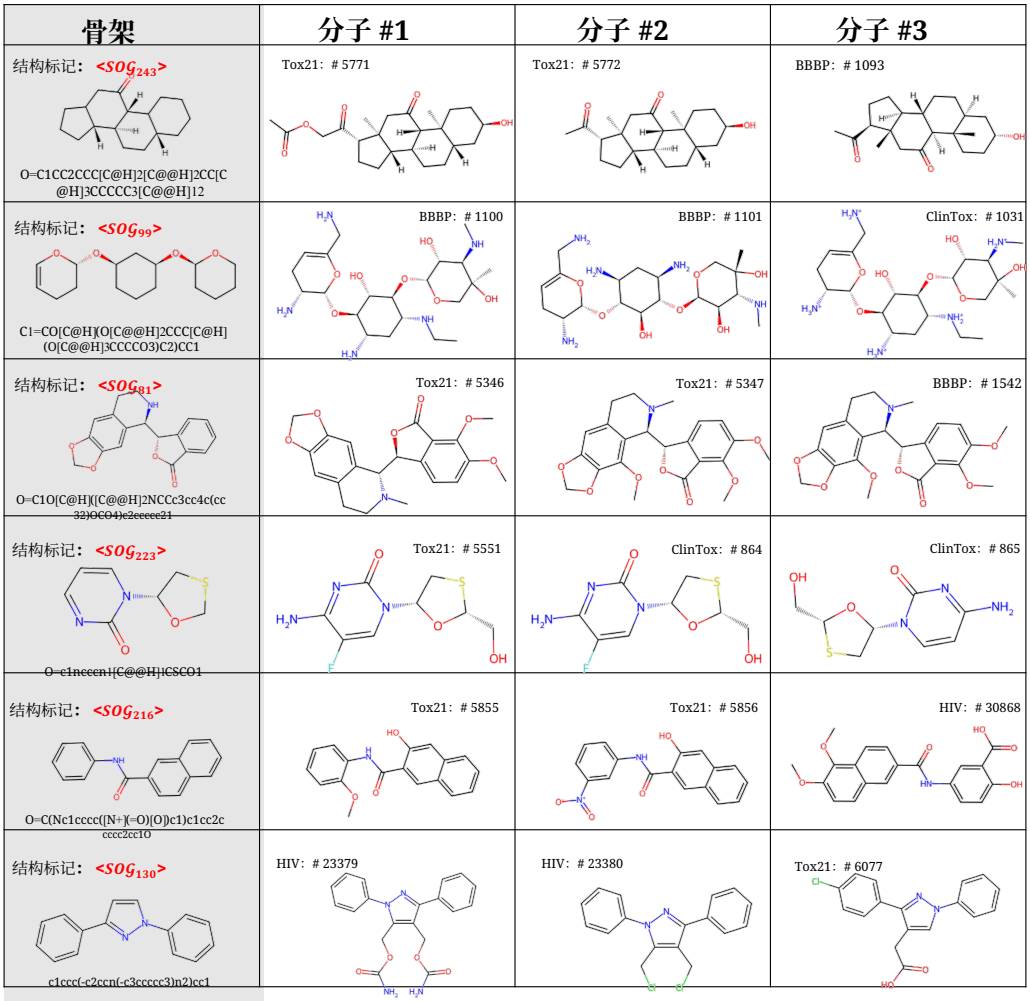

图11：不同训练集中的分子具有相同的骨架和结构标记。