Knowledge Transfer Office
香港城市大學
City University of Hong Kong

用于分子图的图神经网络预训练方法及系统

信息和通信

计算机/人工智能/数据处理和信息技术

机会

自监督预训练，特别是使用掩码语言建模（MLM）策略，已成为自然语言处理（NLP）等领域的基石，并已适用于图神经网络（GNN）从分子图数据中学习。然而，现有的掩码策略，如属性掩码（AttrMask），通常采用随机掩码，这导致了一个关键问题：分子数据集中原子的不平衡分布。在典型的分子数据集中，常见原子（如碳、氧、氮）占绝大多数（例如约96%），而痕量元素（如氯、氟）出现频率较低。随机掩码不成比例地选择高频原子进行掩码，导致预训练模型过度关注学习这些常见原子的表示，而忽略了稀有但可能具有重要化学意义的元素。这种不平衡限制了模型捕捉全面化学知识的能力，因为痕量元素可能在分子性质和功能中起关键作用。尽管一些以模型为中心的方法，如MOLE-BERT和GraphMAE，试图通过添加复杂模块（例如标记器或重构层）来解决这个问题，但它们增加了计算负担和参数数量。因此，显然需要一种更高效、以数据为中心的解决方案，直接解决原子不平衡问题，而不增加模型复杂性，从而改进GNN在下游分子预测任务（如药物发现和性质预测）中的预训练。

备注
IDF:1618

IP状态
已申请专利

技术成熟度等级 (TRL) 4

发明人

馮志聰教授

林煒先生

查询： kto@cityu.edu.hk

技术

本发明引入了一种以数据为中心的加权随机掩码策略（WMM），以解决预训练期间的原子不平衡问题。该方法不是随机掩码原子，而是根据原子类型在特定分子中的频率为每个原子分配权重。权重使用定义的公式计算：$w_a(i) = \frac{\ln(k(n_a(i) + 1))}{n_a(i)}$，其中 $n_a(i)$ 是分子中原子类型 $a(i)$ 的数量，k 是一个超参数（通常≥0.8）。这种加权确保数量较多的原子类型（如碳）获得较低的权重，降低其被掩码的概率，而稀有原子类型获得较高的权重，增加其掩码可能性。该方法通过鼓励模型预测较少出现的原子来平衡学习信号，从而捕捉更多样化的化学信息。该策略可与现有的自监督学习框架（如AttrMask、GraphMAE或掩码原子建模MAM）无缝集成，无需额外模型参数。在预训练期间，GNN处理掩码后的分子样本，预测被掩码的原子，并通过迭代训练将预测与原始原子进行比较。该策略最大化从每个分子独特原子组成中学习的能力，增强了模型在不同分子结构间的泛化能力，并提高了下游任务的性能。

优势

- 通过以数据为中心的加权策略直接解决原子不平衡问题，改进对稀有但关键元素的学习。
- 在不增加模型参数或计算复杂性的情况下提升预训练性能，保持高效性。
- 兼容多种现有的自监督学习模型（如AttrMask、GraphMAE、MAM），提供灵活性和易于集成性。
- 提高下游分子性质预测任务的整体准确性，实验结果显示性能提升（例如，AttrMask提升1.38%，GraphMAE提升0.89%）。

- 提供可调超参数（k）来控制掩码概率，允许针对不同数据集和任务进行优化。
- 通过使预训练任务更具挑战性，减少对高频原子的过拟合，从而获得更好的泛化能力。

应用

- 药物发现与设计：预训练的GNN可预测分子性质，帮助识别潜在候选药物。
- 分子性质预测：应用包括化合物毒性（如ClinTox）、溶解度和生物活性预测。
- ADMET（吸收、分布、代谢、排泄、毒性）分析：增强药物开发中的药代动力学和安全性评估模型。
- 化学信息学与材料科学：通过改进分子表示加速具有所需性质的新材料发现。
- 蛋白质序列预训练：该方法可扩展至生物序列（如蛋白质），以处理不平衡的氨基酸分布。
- 教育与研究工具：为化学和生物学领域的学术和工业研究提供稳健的预训练模型。