Hierarchical Graph Tokenization for Molecule-Language Alignment

Yongqiang Chen [*1] **Quanming Yao** [2] **Juzheng Zhang** [2] **James Cheng** [1] **Yatao Bian** [3]
[*]Work done during an internship at Tencent AI Lab. [1]*The Chinese University of Hong Kong* [2]*Tsinghua University* [3]*Tencent AI Lab*. Correspondence to: Yatao Bian yatao.bian@gmail.com.

Abstract

Recent years have witnessed growing interest in extending the capabilities of large language models (LLMs) to molecular science. As LLMs are predominantly trained with text data, most existing approaches adopt a graph neural network to represent a molecule as a series of node tokens for molecule-language alignment, which, however, have overlooked the inherent hierarchical structures in molecules. Notably, high-order molecular structures contain rich semantics of functional groups, which encode crucial biochemical functionalities of the molecules. We show that neglecting the hierarchical information in tokenization will lead to subpar molecule-language alignment and severe hallucination. To address this limitation, we propose a novel strategy called **HI**erarchical **G**rap**H** **T**okenization (HIGHT). HIGHT employs a hierarchical graph tokenizer that encodes the hierarchy of atom, motif, and molecular levels of informative tokens to improve the molecular perception of LLMs. HIGHT also adopts an augmented instruction tuning dataset, enriched with the hierarchical graph information, to further enhance the molecule-language alignment. Extensive experiments on 14 real-world benchmarks verify the effectiveness of HIGHT in reducing hallucination by 40%, and significant improvements in various molecule-language downstream tasks.