[BIOINFO 2025]

Hierarchical molecular representation learning with multi-level vector quantization for toxicity prediction

Yewon Shin[1], and Hojung Nam[1,2,3*]

[1]AI Graduate School, Gwangju Institute of Science and Technology (GIST)
[2]Department of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST)
[3]Center for AI-Applied High Efficiency Drug Discovery (AHEDD), Gwangju Institute of Science and Technology (GIST)
*Corresponding author: hjnam@gist.ac.kr

Toxicity identification plays a critical role in ensuring drug safety and minimizing the risk of drug withdrawals during preclinical and clinical stages of drug development. Recently, many computational approaches have been developed to efficiently predict toxicity by leveraging molecular structural information to address the limitations of time-consuming and costly experimental methods. Substructure patterns, including functional groups, not only determine molecular properties but are also closely related to specific toxicities. However, many approaches do not fully capture the influence of structural context at the atom or fragment level, even though identical local structural patterns can lead to different molecular properties depending on the surrounding chemical environment. In this study, we propose a hierarchical molecule pretraining framework that can effectively incorporate atom and fragment-level context for toxicity prediction. First, the proposed model is trained to learn discrete vector codebooks at both levels, encoding the chemical properties and local structural patterns of molecules through multi-level self-supervised tasks and vector quantization. Afterward, the model learns the overall hierarchical structure of molecules using a heterogeneous graph with a chemical knowledge-guided pretraining strategy. Our model outperforms other existing pretrained methods across various toxicities. Furthermore, we confirm that the learned multi-level codebooks effectively capture local structural context and fragment characteristics, enabling the representation to encode the semantic similarity of molecules.