[BIOINFO 2024]

Enhancing Molecular Property Prediction through Two-Stage Pre-training, Multi-Modal Fusion, and Scaffold Ensemble

Minyoung Kim[1], Heesang Moon[2], and Mina Rho[1,2,3,*]

[1] *Department of Artificial Intelligence, Hanyang University*
[2] *Department of Computer Science, Hanyang University*
[3] *Department of Biomedical Informatics, Hanyang University*
** Corresponding author: minarho@hanyang.ac.kr*

Accurate molecular property prediction is crucial for preclinical research, drug screening, repositioning and design. Traditional experiments are labor-intensive, time-consuming and expensive, driving the need for in silico techniques such as deep learning models. However, these models often suffer from poor generalization and overfitting due to limited labeled data. To address these challenges and incorporate chemical knowledge into the embedding space, we propose a novel two-stage pretraining method that effectively captures informative features at both local- and global-level. The pre-trained embeddings are then fine-tuned for molecular property prediction tasks, utilizing a multi-modal fusion learning strategy to leverage the strengths and compensate the drawbacks of diverse molecular representations derived from sequence and graph models. We applied scaffold splitting which results in significant performance fluctuations depending on the distribution of scaffolds across train, validation, and test sets. To overcome this issue, a scaffold ensemble strategy is adopted. We evaluated this model on eight benchmark datasets: BBBP, BACE, ClinTox, Tox21, ToxCast, LIPO, ESOL, and FreeSolv, achieving superior or comparable performance against existing models. Specifically, our method achieves a 2.2% improvement in classification tasks and 1.8% improvement in regression tasks. These results demonstrate the model's ability to learn chemically meaningful features and exhibit robust generalization capabilities.