Leveraging chemical structural knowledge and in vitro toxicity information for enhancing multi-task in vivo toxicity prediction

Minsu Park[1], Yewon Shin[2], Hyunho Kim[1,3†*], Hojung Nam[1, 2, 4*]

[1]Department of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Buk-gu, Gwangju, 61005, Republic of Korea

[2]AI Graduate School, Gwangju Institute of Science and Technology (GIST), Buk-gu, Gwangju, 61005, Republic of Korea

[3]Division of Advanced Predictive Research, Korea Institute of Toxicology, 141 Gajeong-ro, Yuseong-gu, Daejeon, 34114, Republic of Korea

[4]Center for AI-Applied High Efficiency Drug Discovery (AHEDD), Gwangju Institute of Science and Technology (GIST), Buk-gu, Gwangju, 61005, Republic of Korea

* Corresponding authors

† Current affiliation

Email addresses:

MP : 15pms@gm.gist.ac.kr

YS : yewon.shin@gm.gist.ac.kr

HK : hyunho.kim@kitox.re.kr

HN : hjnam@gist.ac.kr

Abstract

The evaluation of potential drug toxicity is a crucial step in early drug development. As part of the New Approach Methodologies (NAMs) aimed to overcome the limitations of animal-based in vivo assessments, computational approaches for in vivo toxicity prediction have been developed. However, these models often suffer from low performance due to limited data availability. To address this issue, we introduce a novel knowledge transfer model that systemically leverages both chemical structure and in vitro toxicity information through transfer learning and multi-task learning.

Our model utilizes three types of knowledge transfer strategies. First, the GNN layers are pre-trained on the ChEMBL dataset to learn the representation about molecular structure. Second, this pre-trained model is further trained on Tox21 dataset with multi-task learning to incorporate the auxiliary in vitro toxicity patterns. Lastly, for the in vivo toxicity prediction, we employed a cross-attention mechanism between molecular structural embeddings and in vitro toxicity embeddings. With this hierarchical knowledge transfer strategy, the model can selectively leverage the information from both chemical structure and in vitro toxicity context for each in vivo toxicity.

As a result, our proposed model outperformed existing baseline models including transfer learning and multi-task learning based methods across three in vivo toxicities: Carcinogenicity, Drug-Induced Liver Injury (DILI), and Genotoxicity. Furthermore, the model's applicability as a prediction tool for early-stage drug development was validated through the extensive screening of molecules in the DrugBank database.