[BIOINFO 2024]

Predicting Drug-likeness through knowledge alignment and EM-like one-class boundary optimization

Dongmin Bang[1,2], Inyoung Sung[3], Yinhua Piao[4], Sangseon Lee[5] and Sun Kim[1,2,4,6,*]

[1] *Interdiciplinary Program in Bioinformatics, Seoul National University*
[2] *AIGENDRUG Co., Ltd.*
[3] *BK21 FOUR Intelligence Computing, Seoul National University*
[4] *Department of Computer Science and Engineering, Seoul National University*
[5] *Department of Artificial Intelligence, Inha University*
[6] *Interdiciplinary Program in Artificial Intelligence, Seoul National University*
** Corresponding author: sunkim.bioinfo@snu.ac.kr*

The advent of generative AI models is revolutionizing drug discovery, generating de novo molecules at unprecedented speed. However, accurately identifying drug candidates among generated molecules remains an open problem. The essence of this drug-likeness prediction task lies in constructing a compact subspace that encompasses majority of approved drugs with only a small number of unknown compounds (drug candidates) inside. Computational challenges arises in constructing a decision boundary on an unbound chemical space that lacks definite negatives, i.e, non drug-likeness. Approved drugs exist highly dispersed across structural space, making it more harsh to effectively separate drugs from non-drugs through existing classifiers. Addressing such challenges, we introduce a novel approach for learning a compact boundary of drug-likeness through an Expectation-Maximization (EM)-like iterative optimization process. Specifically, we refine both the boundary and the distribution of the embedding space via metric learning, allowing the model to iteratively tighten the drug-like boundary while pushing non-drug-like compounds outside. Augmented by integration of biomedical context within knowledge graphs via multi-modal alignment, our model demonstrates up to five orders of magnitude improvement in drug-likeness prediction performance metrics. Our model further showcases its utility in large-scale screening of AI-generated compounds through zero-shot toxic compound filtering.