[BIOINFO 2024]

TransGOM: Transformer for Phenotype-based Molecular Generation using Gene Set Enrichment Scores

Jitae Kim[1], Woojong Sim[1], and Minji Jeon[1,2,3*]

[1]Department of Medicine, Korea University College of Medicine
[2]Department of Biomedical Informatics, Korea University College of Medicine
[3]Biomedical Research Center, Korea University Anam Hospital
**Corresponding author: mjjeon@korea.ac.kr*

Many generative models for target-specific drug discovery have been proposed. However, they rely on the 3D structures of targets and this limits their ability to generate new bioactive molecules when the accurate 3D structures of targets re unavailable. Recently, phenotype-based molecule generation AI models have been proposed to address this limitation by using differentially expressed genes (DEG) profiles to generate molecular structures that may induce the given DEG profiles. However, these models have not consistently produced robust results, largely due to the inherent instability of DEG profiles. DEG profiles, generated through microarray or RNA sequencing, are highly affected by environmental factors, causing batch effects and noise in the data. As a result, achieving consistent and reliable outcomes from the DEG profile-based molecule generation models is challenging.

We propose TransGOM (Transformer-based model using Gene Ontology for Molecule Generation), which is a phenotype-based de novo molecule structure generation model. To mitigate the noise in DEG profiles from the ConnectivityMap, we convert them into normalized enrichment scores (NES) of Gene Ontology (GO) terms using Gene Set Enrichment Analysis (GSEA) analysis. These GO terms, ranked by NES, represent biological phenotypes of compounds, serving as the input for the model. The transformer-based model generates molecules that are likely to induce the biological phenotypes encoded by the GO terms. Due to the long length of the input data, rotary position embedding is utilized to account for both absolute and relative positional information.

TransGOM performance was validated through evaluation metrics such as uniqueness, novelty, validity, drug-likeness and synthetic accessibility. We further assessed TransGOM by inputting GO terms induced by vorinostat. The model successfully generated a set of molecules structurally similar to vorinostat, demonstrating its ability to capture relevant molecular features. TransGOM shows potential to advance drug discovery, particularly in phenotype-based approaches, offering a novel

paradigm for de novo drug design.