[BIOINFO 2023]

Evaluating BERT-based large language models in target druggability prediction

Sera Park[1], and Hyunjin Shin[1,*]

[1]*MOGAM Institute for Biomedical Research, Seoul 06730, Republic of Korea*

**Corresponding author: hyunjin.shin@mogam.re.kr*

It is extremely important to discover innovative and efficacious drug targets for successful drug development. Among many factors that may affect a target's efficacy, druggability must be properly considered because the target should be modulated by the given modality, such as small molecules, for inducing sufficient therapeutic effects. Recently, there have been many attempts to predict target druggability using machine learning and more advanced artificial intelligence (AI) techniques. In particular, large language models (LLMs) are expected to be useful in terms that it can detect long-distance dependencies in the amino acid sequences of a target protein without the need for structural information. To test this idea, we compared BERT-based LLMs over traditional machine learning methods such as random forest (RF), extreme tree (ET), support vector machine (SVM), and stacking ensemble-learning model. In this comparative study, we estimated the predictive performance of these models on 3,003 protein sequences as benchmark data. In conclusion, this study could provide insights into the potential opportunities and challenges of LLM in druggable target discovery.