ABSTRACT

Purpose: Using ligand-based approach and Quantitative Structure Activity Relationship (QSAR), this project aims to apply machine learning (ML) to compounds screening stage, to speed up drug discovery processes and reduce failure rate.

Study design: In this project, QSAR model for exploring potential inhibitors of lung cancer target, C-Met, has been designed by using regression. 4771 molecules are collected from the ChEMBL database and calculated to chemical descriptors for ML model building. pIC50 is the focus as the higher the value, the better the potency of a drug.

Findings: Molecular weight, solubility, hydrogen bond donors and acceptors are four metrics that could measure drugs' absorption and toxicity. Random forest is chosen for predicting the potency of hit compounds for C-Met as its R2 (0.58) was decent. Also, it's a highly interpretable model.

Originality/value: Generating Shapley feature importance graph from random forest was the last step of the ML processes, but it's a starting point for lead-optimization because it could serve as a guideline for scientists. This project built on the foundation of QSAR, which is valuable for future research with the proteochemometric (PCM) model as it could generate interactions between multi protein targets instead of one target.

Practical Implications: Companies could expedite drug development using ML. They could filter out toxic drugs by following Lipinski's druglikeness rules. Instead of trial and error, drug developers would have a clear direction by using the Shapley graph when they alter compounds.

Keywords: AI, Drug Discovery, QSAR, Compound Screening, Random Forest