Title

A bootstrap aggregating machine learning model for balanced prediction of androgen receptor agonistic toxicity with uncertainty quantification.

Jidon Jang, Byung Ho Lee, Jeong Hyun Lee, Woo Dae Jang, Ho Won Seo, Mi Young Lee, Ye Sol Han, Su Ah Kim, Su Jin Park, Yumi Noh, and Kwang-Seok Oh

Abstract

This study aims to develop and validate a machine learning-based model for accurately predicting androgen receptor (AR) agonistic toxicity, addressing challenges posed by data imbalance and low sensitivity in existing models. AR agonistic activity is a key indicator of reproductive toxicity, potentially leading to prostate cancer. While machine learning models have been created to predict AR agonists, they often suffer from biased learning due to imbalanced datasets. To overcome this, we developed a bootstrap aggregated k-nearest neighbor model which significantly improved the balance of prediction. By quantifying uncertainty and using it as an additional screening metric, the model further enhanced the reliability of its predictions, offering a robust approach for large-scale screening of AR-based toxicity. This study demonstrates that our model effectively addresses data imbalance and delivers balanced classification metrics compared to previous approaches, offering a promising tool for drug screening applications.