SHAP-based tools for ADMET compounds profiling on the example of metabolic stability optimization

Agnieszka Wojtuch[1], Rafał Jankowski[1], Sabina Podlewska[2]*

[1]Faculty of Mathematics and Computer Science, Jagiellonian University, 6 S. Łojasiewicza Street, 30-348 Kraków, Poland.

[2]Maj Institute of Pharmacology, Polish Academy of Sciences, 12 Smętna Street, 31-343, Kraków, Poland

e-mail: smusz@if-pan.krakow.pl

INTRODUCTION

The importance of optimization of ADMET properties within the drug design process is undeniable. The earlier the ADMET profiling is introduced, the better, as even the most active compounds can be rejected from further research due to the unfavorable physicochemical or pharmacokinetic properties or due to the induction of toxic effects.

In the study, we focused on metabolic stability, which is important not only due to the fact that the candidate drug should be stable enough to have sufficient time to produce therapeutic effect, but also due to the possibility of the formation of toxic products after transformations. We developed a tool, which enables not only *in silico* evaluation of metabolic stability (expressed as compound half-lifetime), but also provides pieces of information, which can guide the process of structure optimization with the reference to the evaluated parameter.

MATERIALS AND METHODS

CHEMBL-derived datasets describing human and rat metabolic stability (expressed as half-lifetime) were prepared. Two compounds representations were used: MACCSFP and Klekota and Roth Fingerprint (KlekFP). Both classification and regression models were developed. For classification, the following cut-offs were used:

- ≤0.6 – low stability,
- (0.6–2.32> – medium stability,
- >2.32 – high stability.

Naïve Bayes classifier, Support Vector Machines (SVMs), and several models based on trees were applied.

Shapley Additive exPlanations (SHAP) were used to examine the influence of particular chemical substructures on the model's outcome.

The scheme of the whole protocol is presented in Figure 1.

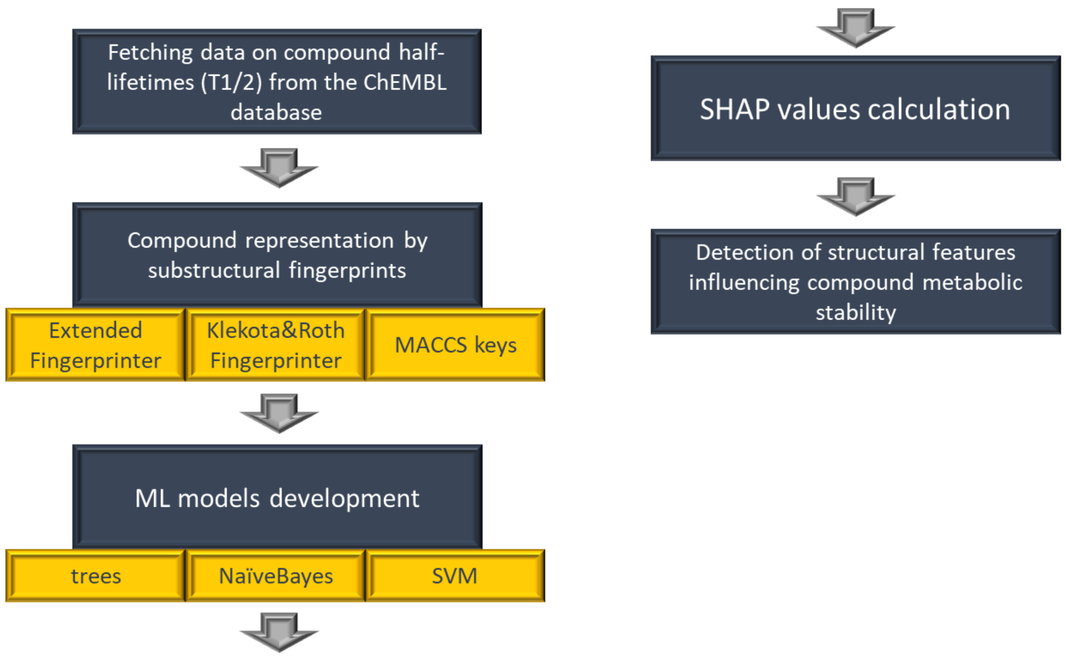

Figure 1. Scheme of the developed protocol for metabolic stability evaluation and optimization.z

RESULTS

The prediction power of ML models is evaluated with the Area Under the Receiver Operating Characteristic Curve (AUC). In the case of regression the prediction correctness was assessed with the use of the Root Mean Square Error (RMSE). The results are presented in Figure 2. In general, the predictions of compound half-lifetimes are satisfactory with AUC values over 0.8 and RMSE below 0.4-0.45. All class assignments performed on human data are more effective for KRFP with the improvement over MACCSFP ranging from ~0.02 for SVM and trees up to 0.09 for Naïve Bayes. Classification efficiency performed on rat data is more consistent for different compound representations with AUC variation of around 1 percentage point. When regression experiments are considered, the KRFP provides better half-lifetime predictions than MACCSFP for 3 out of 4 experimental setups – only for studies on rat data with the use of trees, the RMSE is higher by 0.01 for KRFP than for MACCSFP.

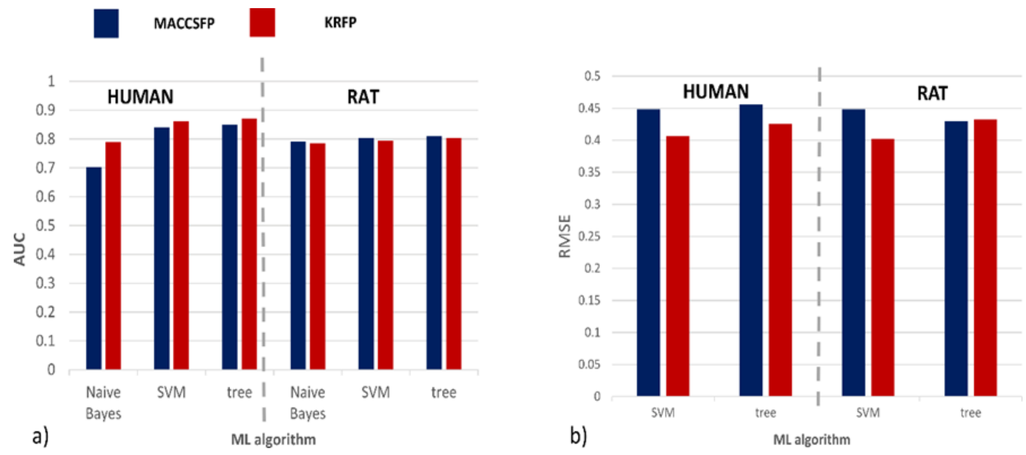

Figure 2. Global prediction power of the ML algorithms in a) classification and b) regression studies.

SHAP-BASED ANALYSIS

Analysis of the methodology output for the example compund, CHEMBL2207577 (Figure 3), shows that the highest contribution to the compound stability is indicated to be the aromatic ring with the chlorine atom attached (feature 3545) and tiophene (feature 1915), whereas compound instability was mostly contributed by the secondary amine (feature 677).

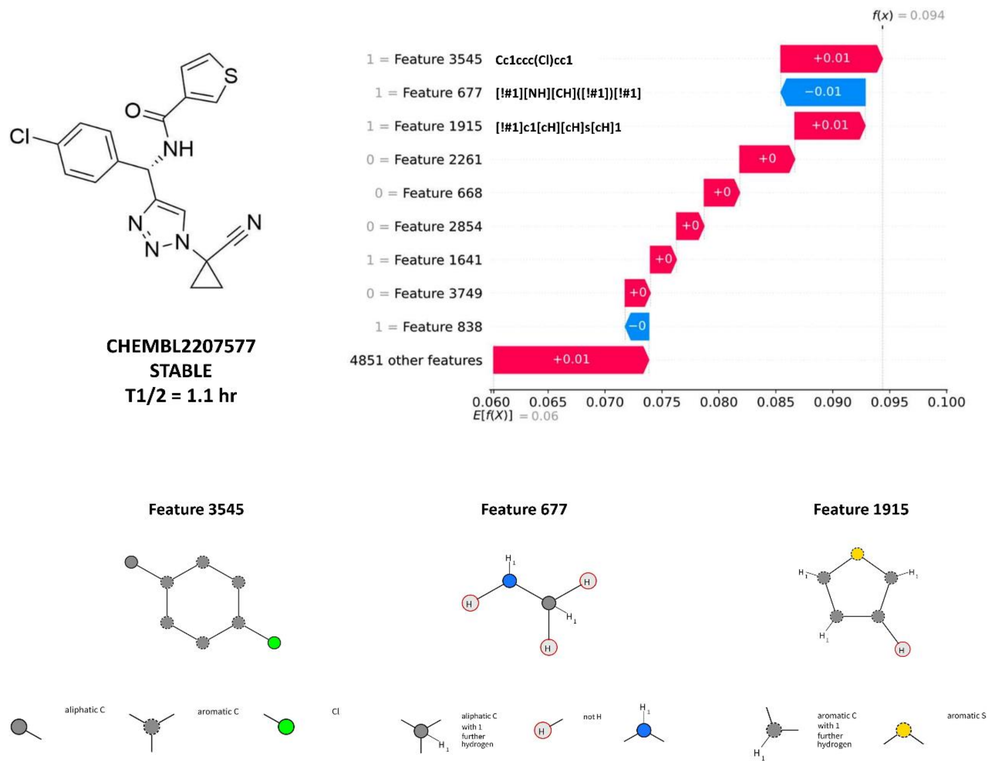

Figure 3. Analysis of the metabolic stability prediction for CHEMBL2207577 for human/KRFP/trees predictive model.

WEB SERVICE

The results of all experiments can be analyzed in detail with the use of the web service, which can be found at https://metstab-shap.matinf.uj.edu.pl/. In addition, the user can submit their own compound and its metabolic stability will be evaluated with the use of the constructed models and the contribution of particular structural features will be evaluated with the use of the SHAP values.

Moreover, in order to enable manual comparisons, the most similar compound from the ChEMBL set (in terms of the Tanimoto coefficient calculated on Morgan fingerprints) is provided for each submitted compound (if the similarity is above the 0.3 threshold).

Obtaining such information enables optimization of metabolic stability as the substructures influencing this parameter are detected. Moreover, the comparison of several ML models and compound representations allows to provide a comprehensive overview of the problem.

References

Wojtuch, A.; Jankowski, R.; Podlewska, S. *How can SHAP values help to shape metabolic stability of chemical compounds?* J. Cheminf. 2021, 13, 74

Acknowledgments

The study was supported by the grant OPUS 2018/31/B/NZ2/00165 financed by the National Science Centre, Poland (www.ncn.gov.pl)