Development of Robust and Predictive Machine Learning QSAR Models for Hepatic Stability

Guilherme M. Silva[a], Holli-Joi Sullivan[b], Marielle K. Rath[b], Vinicius M. Alves[b], Eugene N. Muratov[b], Carlos H. T. P. Silva[a], Alexander Tropsha[b]

[a] *Computational Laboratory of Pharmaceutical Chemistry, School of Pharmaceutical Sciences, University of São Paulo, Ribeirão Preto, SP, 14040-903, Brazil.*
[b] *Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, 27599, USA.*

University of São Paulo

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

ELRIG UK
TRANSLATING IDEAS INTO THERAPIES 2021
CO-HOSTED WITH THE BRITISH PHARMACOLOGICAL SOCIETY
JUNE 30 - JULY 1 - ONLINE

INTRODUCTION

Assessment of pharmacokinetic properties of compounds is a critical step in drug discovery. Measuring hepatic stability is essential in establishing the drug accumulation and clearance in the body. Usually, this endpoint is evaluated *in vivo*, using rats, or *in vitro*, using human liver microsomes. Recently, *in silico* approaches been recognized as alternative approaches to evaluating the pharmacokinetic properties of bioactive compounds.[1] Herein, we describe (i) the collection, curation, and integration of the largest publicly available dataset of human hepatic stability measured *in vitro* with liver microsomes and (ii) the development and statistical validation of robust and predictive QSAR models for this endpoint.

MATERIALS AND METHODS

Datasets

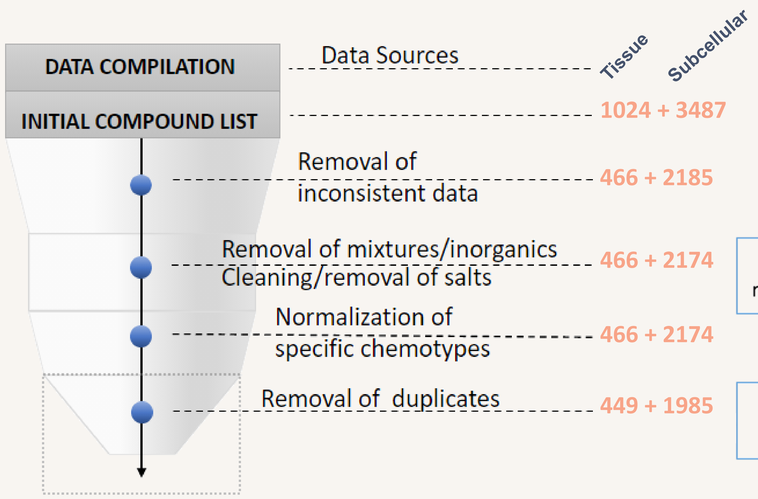
Data was compiled from the European Bioinformatics Institute's ChEMBL database, considering two types of assays for human liver microsome data (expressed in terms of half-life T1/2 values): tissue (ChEMBL ID: CHEMBL2367379) and subcellular (CHEMBL613373). These two datasets totaled 1,024 and 3,487 compounds, respectively.

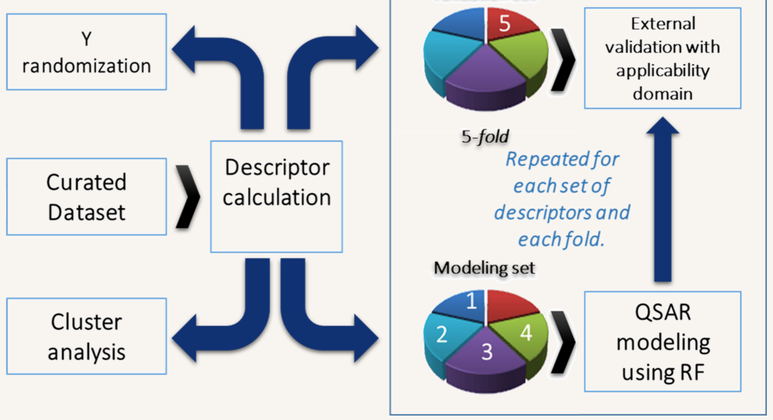
The threshold of 30 minutes was considered to categorize compounds as stable (T1/2 > 30 min) and unstable (T1/2 ≤ 30 min). Thus, the tissue dataset was balanced and consisted of 155 stable and 155 unstable compounds, whereas the subcellular dataset consisted of 1,248 stable and 737 unstable compounds.

QSAR modeling

QSAR models were developed in KNIME employing three types of molecular descriptors (Morgan and MACCS fingerprints, and RDKit properties) using Random Forest (RF) as machine learning algorithm. Models were developed and validated following the best practices proposed by OECD[2] and by employing 5-fold external cross-validation procedure to estimate the robustness of developed models.

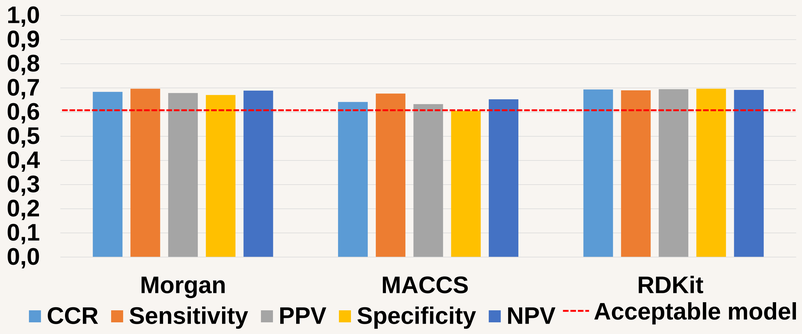
Moreover, we estimated the applicability domain[3] and performed 20 rounds of Y-randomization to ensure the absence of chance correlations.



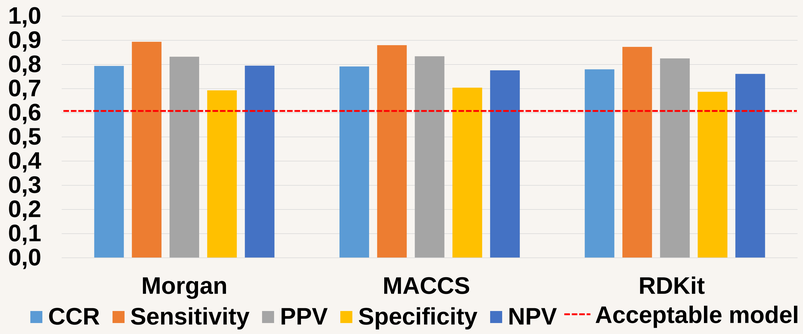


RESULTS AND DISCUSSION

Statistical characteristics of developed QSAR models with tissue T1/2 values



Statistical characteristics of developed QSAR models with subcellular T1/2 values



In general, all models had high predictive power. For tissue models, RDKit showed the highest correct classification rate (CCR) of 0.69, sensitivity of 0.69, positive predictive values (PPV) of 0.70, specificity of 0.70, negative predictive value (NPV) of 0.69, and coverage of 1.0. Best subcellular model was built using Morgan fingerprints with respective values of 0.79, 0.89, 0.83, 0.69, 0.80, and 1.00.

Only one compound was present in both datasets. Therefore, we performed a cross-screening between the models developed for each endpoint. These predictions showed low cross-assay predictivity, indicating that subcellular and tissue assays are not similar to estimate hepatic stability

CONCLUSIONS

We collected, curated, and integrated the largest publicly available dataset for human hepatic stability and developed robust and predictive QSAR models. These can be employed to predict the half-life of compounds in human liver microsomes with high accuracy.

The models will be implemented as part of a comprehensive platform for the prediction of pharmacokinetic parameters, which will be freely available for the scientific community.

REFERENCES

[1] Liu, R. et al. *J. Chem. Inf. Model.* **2015**, 55, 1566-1575.
[2] Tropsha, A. *Mol. Inform.* **2010**, *29*, 476–488.
[3] Tropsha, A. et al. *Curr. Pham. Des.* **2007**, 13, 3494-3504.

ACKNOWLEDGEMENTS

CAPES CNPq FAPESP