Systemic Fingerprints for Predicting Liver Related Adverse Events

Palle S. Helmke[a], Barbara Füzi[a], Gerhard F. Ecker[a]

[a]*Department of Pharmaceutical Sciences, University of Vienna, Josef-Holaubek Platz 2, 1090 Vienna, Austria*

Background

Data science gains increasing importance in Next-Generation Risk Assessment (NGRA) due to the growing volume of data in life sciences. It enables researchers to identify patterns and predict toxicity, which refers to the harmful effects of a substance, such as a drug, on a living organism, leading to adverse events (AEs). Drug-induced liver injury is a commonly observed AE, resulting in liver damage caused by drug administration. Cholestasis, a subtype of DILI, is characterized by impaired bile flow in the liver. Understanding the underlying mechanisms of these conditions is crucial for identifying potential drug candidates and mitigating their hepatotoxic effects. This study aims to provide insights into cholestasis characterization, emphasizing the importance of compound target and pathway fingerprints for evaluating drug candidates. The integration of traditional machine learning algorithms offers promising opportunities for predictive modeling in this context.

Methods

- ✓ Liver specific compound target and pathway interactions of the cholestasis +/- dataset were collected using an open-source KNIME [1] workflow called "Path4drug" [2]
- ✓ Binary tables of three different descriptor sets (target, pathway, combined target/pathway) were created as input for machine learning models
- ✓ The cholestasis +/- dataset was partitioned 80/20 for training/test split (20% reserved as testing holdout sample after evaluation with 80% training set)
- ✓ Undersampling and Synthetic Minority Oversampling Technique (SMOTE) were applied on the 1:4 imbalanced cholestasis +/- dataset with the cholestasis positive (+) compounds as minority class
- ✓ 10-fold cross validation loop was built to evaluate model performances
- ✓ Tested machine learning algorithms in KNIME: Random Forest (RF), Gradient Boosted Tree (GBT), Extreme Gradient Boosting (XGBoost), Decision Tree (DC)
- ✓ Performance statistics were calculated to assess model performance and choose best performing model on cholestasis +/- dataset

Workflow

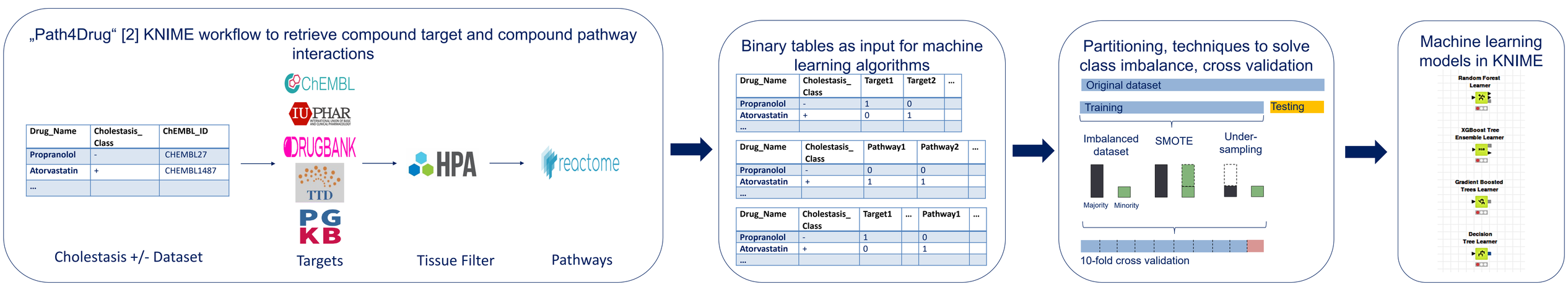

Figure 1. Workflow depicting the retrieval of descriptors in the open-source "Path4Drug" [2] KNIME workflow to create binary tables as input for four machine learning models in a 10-fold cross validation loop after application of techniques to solve class imbalance.

Results

Descriptor	Cross-Validation	Techniques to solve imbalance	Best Model	Precision (Mean)	Sensitivity (Mean)	Specificity (Mean)	Accuracy (Mean)	MCC (Mean)
Target	10-fold	Original imbalanced classes	XGBoost	0.53 ± 0.10	0.33 ± 0.09	0.92 ± 0.02	0.80 ± 0.03	0.31 ± 0.10
Target	10-fold	SMOTE	XGBoost	0.40 ± 0.07	0.51 ± 0.09	0.81 ± 0.03	0.74 ± 0.04	0.29 ± 0.10
Target	10-fold	Undersampling	XGBoost	0.69 ± 0.07	0.65 ± 0.08	0.70 ± 0.12	0.67 ± 0.05	0.35 ± 0.10
Pathway	10-fold	Original imbalanced classes	RF	0.55 ± 0.10	0.20 ± 0.07	0.96 ± 0.02	0.80 ± 0.02	0.24 ± 0.08
Pathway	10-fold	SMOTE	XGBoost	0.39 ± 0.10	0.41 ± 0.10	0.83 ± 0.04	0.74 ± 0.04	0.23 ± 0.12
Pathway	10-fold	Undersampling	RF	0.66 ± 0.08	0.60 ± 0.11	0.68 ± 0.09	0.64 ± 0.07	0.29 ± 0.14
Target/ Pathway	10-fold	Original imbalanced classes	XGBoost	0.51 ± 0.11	0.31 ± 0.10	0.92 ± 0.04	0.79 ± 0.03	0.28 ± 0.10
Target/ Pathway	10-fold	SMOTE	XGBoost	0.37 ± 0.08	0.41 ± 0.11	0.82 ± 0.04	0.73 ± 0.04	0.22 ± 0.11
Target/ Pathway	10-fold	Undersampling	XGBoost	0.65 ± 0.06	0.68 ± 0.07	0.62 ± 0.10	0.65 ± 0.06	0.30 ± 0.11

Table 1. Performance statistics of descriptor sets in combination with 10-fold cross validation, techniques to solve class imbalance and four different machine learning algorithms. For the target as well as the combined target/pathway descriptor set, the highest Matthews Correlation Coefficient and Sensitivity was reached with undersampling and the XGBoost model. The pathway descriptor set reached the highest Matthews Correlation Coefficient and Sensitivity in combination with undersampling and a Random Forest model. Performance statistics are defined as mean ± standard deviation (SD).

Descriptor	Feature
Target	Bile salt export pump
Target	Solute carrier family 15 member 1
Pathway	Metabolism
Target	Carbonic anhydrase 2
Pathway	Miscellaneous substrates
Pathway	Initiation of Nuclear Envelope (NE) Reformation
Target	Albumin
Target	Cytochrome P450 1A1
Pathway	Paracetamol ADME
Pathway	Negative regulation of the PI3K/AKT network

Table 2. Feature importance analysis of a Gradient Boosted Tree model with the combined descriptor set in the H2O framework in KNIME with the Bile Salt Export Pump (BSEP) as top 1 ranked feature.

References and tools:

1. Berthold, M.R. et al. *SIGKDD Explor. Newsl. 11 (1),* **2009**, 26–31
2. Füzi, B; Gurinova, J; Hermjakob, H; Ecker, GF; Sheriff, R. *Front. Pharmacol. Sec. Predictive Toxicology, 12*, **2021**
3. Viennois, E; Pujada, A; Zen, J; Merlin, D. *Compr Physiol. 8(2),* **2018**, 731-760

Discussion & Conclusion

- ✓ Inhibition of the top 1 ranked transporter Bile Salt Export Pump (BSEP) causes toxic accumulation of bile acids in the liver and is considered one of the primary reasons for cholestasis.
- ✓ Recently, sites beyond the small intestine, where functional expression of the top 2 ranked transporter SLC15A1 (PEPT1) has been discovered, include the bile duct epithelium [3].
- ✓ The primary mode of transportation for bile salts and acids in the bloodstream is through their binding to serum albumin (ALB), which is the top 7 ranked feature.
- ✓ The target descriptor set in combination with undersampling utilizing the XG Boosted Tree algorithm overall performed best regarding the Matthews Correlation Coefficient.