Smajić *et al. Journal of Cheminformatics (2022) 14:54*
https://doi.org/10.1186/s13321-022-00635-2

Journal of Cheminformatics

EDUCATIONAL

Open Access

Using Jupyter Notebooks for re-training machine learning models

Aljoša Smajić, Melanie Grandits* and Gerhard F. Ecker

Abstract

Machine learning (ML) models require an extensive, user-driven selection of molecular descriptors in order to learn from chemical structures to predict actives and inactives with a high reliability. In addition, privacy concerns often restrict the access to sufficient data, leading to models with a narrow chemical space. Therefore, we propose a framework of re-trainable models that can be transferred from one local instance to another, and further allow a less extensive descriptor selection. The models are shared via a Jupyter Notebook, allowing the evaluation and implementation of a broader chemical space by keeping most of the tunable parameters pre-defined. This enables the models to be updated in a decentralized, facile, and fast manner. Herein, the method was evaluated with six transporter datasets (BCRP, BSEP, OATP1B1, OATP1B3, MRP3, P-gp), which revealed the general applicability of this approach.

Keywords: Classification models, Transporter proteins, Decentralization, Re-training, Jupyter Notebook

Introduction

The importance of machine learning (ML) approaches in drug discovery and in silico toxicity prediction has shown a significant increase in recent years. As available toxicity data has significantly increased [1–3], ML approaches became an essential part of the drug discovery pipeline. Public–private partnerships such as eTOX [4] and eTRANSAFE [5], as well as public databases (ChEMBL [6], PubChem [7]) enable trustful data supply for the establishment of predictive ML models. For training and improving the performances of ML models, a large amount of data is crucial [8]. However, when seeking to pool data from multiple sources, multiple restrictions occur. Companies quite often restrict access to in house data due to their business value. In addition, collecting, curating, and preserving data requires a lot of effort and time.

Furthermore, once a sufficient amount of qualitative data is established, additional challenges can occur on the path towards the creation of efficient ML models. The selection of chemical descriptors best suited to derive models of sufficient quality is one of them. The selection of a proper set of descriptors is an extensive, time-intensive, and still mostly manual process, especially when trying to understand relationships between chemical properties and their effect on biological targets [9]. Depending on the biological target, the descriptors best suited can considerably vary. Combined with the fact that additional hyper-parameters have to be tuned for each model, the creation of high accuracy ML models becomes an exhaustive process.

To overcome these issues and allow the user to establish predictive models in an easy and fast way, we created a framework that can be used in a semi-automated fashion for the creation and/or re-training of ML models for predicting inhibitory activity towards ABC and SLC transporters. Furthermore, in comparison to previous methods our approach does not require descriptor selection and hyperparameter search which enables fast and efficient model building.

A set of transporters, mainly used in this study, has caught the attention of regulatory agencies such as FDA, EMA, and the Japanese regulatory agency, as the inhibition of these proteins may play a role in

*Correspondence: melanie.grandits@univie.ac.at

Department of Pharmaceutical Sciences, University of Vienna, Vienna, Austria

drug-drug interactions and/or drug-induced liver injury. Therefore, the prediction of inhibitory profiles of small molecules towards these set of transporters can help to guide safety assessments of new drugs as often requested or recommended by regulatory agencies. Additionally, the knowledge can further help in terms of prioritization of compounds at the early Drug Discovery stage by medicinal chemists [10–17].

Combining Jupyter Notebooks (JNs) [18] as a framework for creating ML models and high-quality data regarding transport membrane proteins to train these models, shareable models can be built for the assessment of compounds for their interaction profile. In general, JN is a web-based interactive computing platform that enables the combination of computer code (e.g. python) and rich text elements (e.g. figures). A web browser is used to navigate in the JN app, and the established graphical user interface allows a better representation of files and so-called notebook documents. These notebook documents can be executed as well as read by users, as they contain code, rich text, images, plots, interactive figures and widgets. These notebooks can be easily shared since they are saved as structured text files (JSON format) and enable the transfer of the code of the model from one instance to another for re-training the model [19]. This allows the enrichment of the chemical space of the model. The notebook further provides a generalizable set of molecular descriptors for the ABC and SLC transporter families that has been shown to be applicable at least for the transporter proteins BCRP, BSEP, OATP1B1, OATP1B3, MRP3, and P-gp. The procedure was selected as it comprises the possibility of sharing the notebook in a facile manner and the creation of workflows for non-experienced users. By uploading data to the JN, the code can be executed which will allow the creation of models and the verification of the models within the JN. In addition, due to the ease of the integration of RDKit, JNs comprise a versatile tool for cheminformatics tasks.

Subsequently, JNs are great tools for educational purposes. The TeachOpenCADD platform by AG Volkamer has demonstrated this by creating JNs with step-by-step tutorials that can be used as a teaching platform for classroom lessons and self-studying. Open-source data and Python packages are used as tools for establishing both ligand- and structure-based approaches. The usage of these JNs provides knowledge in the field of cheminformatics and structural bioinformatics for students and users interested in these topics [20]. Therefore, our JN not only offers the possibility of improving the ML models, model building and predictions for the six endpoints but also offers students, universities and interested users to learn more about model building, data handling, datasets, standardization procedure, descriptor calculation and model evaluation in cheminformatics.

Methods

Dataset preparation

In this study, datasets of six different transmembrane transport proteins (BCRP, BSEP, OATP1B1, OATP1B3, MRP3, P-gp) were used as a case study [21–46]. Firstly, datasets from the Vienna LiverTox Workspace (LiverTox) [47] were chosen, as these datasets were already published and used for the development of predictive models. The corresponding web service allows the prediction of substrates and inhibitors for a set of ABC and SLC transporters.

Secondly, an in-house KNIME workflow was used for the retrieval of additional new data from public platforms such as ChEMBL and PubChem (ChEMBL26 [48], CheEMBL27 [49], ChEMBL28 [50], PubChem [7]). The data from ChEMBL 26 and 27 were used as additional training sets (see below), while data from ChEMBL 28 and additional data from Pubchem served as test sets. Activity values were taken from the original publication and class labeling for binary classification was applied based on a threshold of an IC50 value of 10 µM. All data sets were provided in sdf-format together with a binary classification (0/inactive or 1/active) for each of the six endpoints. For each compound the InChIs (IUPAC International Chemical Identifiers), InChI Keys and SMILES (Simplified Molecular Input Entry Specification) were calculated. All datasets are available on GitHub at https://github.com/PharminfoVienna/Retraining_Notebook/tree/main/data.

Before following the standardization protocol, stereochemistry information was removed from the InChIs and duplicated InChIs were identified. In case duplicates show the same class label, one of the compounds was kept. Otherwise, both compounds were removed. Data cleaning and standardization was performed using a modified version of the Standardizer provided by Atkinson (available at https://github.com/flatkinson/standardiser). This tool was applied to remove salts, neutralize, and discard non-organic compounds. Tables 1 and 2 show the number of data points available per transporter for LiverTox and for the newly collected datasets. The datasets were further used to generate classification models which allow the prediction of inhibitors for a number of liver transporters involved in severe side effects.

Descriptor selection

For the characterization of the chemical space related to ABC and SLC transporter inhibition, a variety of molecular descriptors from the RDKit library (version 2020.09.1) were used [51]. These molecular descriptors

Table 1 Overview of the six transporter datasets which are provided on the LiverTox workspace

Endpoint	LiverTox training		LiverTox test	
	Actives	Inactives	Actives	Inactives
Breast cancer resistance protein (BCRP)	432	542	109	86
Bile salt export pump (BSEP)	114	410	43	116
Organic anion transporting polypeptide 1B1 (OATP1B1)	178	1472	64	137
Organic anion transporting polypeptide 1B3 (OATP1B3)	116	1547	40	169
Multidrug resistance associated protein (MRP3)	32	52	–	–
P-glycoprotein (Pgp)	612	549	86	48

Table 2 Overview of the six transporter datasets that were used for the training of the models

Endpoint	Training		Test	
	Actives	Inactives	Actives	Inactives
Breast cancer resistance protein (BCRP)	904	786	149	38
Bile salt export pump (BSEP)	221	1100	3	7
Organic anion transporting polypeptide 1B1 (OATP1B1)	292	1675	18	3
Organic anion transporting polypeptide 1B3 (OATP1B3)	168	1818	13	4
Multidrug resistance associated protein (MRP3)	74	569	0	3
P-glycoprotein (Pgp)	1281	953	136	236

The training set comprises data from LiverTox plus those extracted from ChEMBL 26 and 27, the test set contains data extracted from ChEMBL 28 and PubChem

enable the translation of chemical structures into numerical representations of atomic or molecular properties of compounds. In total, 197 two-dimensional (2D) descriptors were chosen as a starting point for the selection of features applicable for ABC and SLC transporters. Herein, three different feature selection methods from the scikit-learn Python library (version 0.24.2) were applied: VarianceThreshold, Univariate feature selection, and Recursive feature elimination. By applying VarianceThreshold all calculated molecular descriptors with zero variance were removed. As a next step, best descriptors were selected based on a univariate statistical approach. ANOVA-f was chosen over mutual information due to the nature of the six transporter datasets. This method estimates the degree of linear dependency by using the F-test approach. In parallel, recursive feature elimination (RFE) was performed to select features by recursively considering subsets of molecular descriptors. A random forest wrapper was used for assigning the weights. As a last step, the results of both univariate feature selection and recursive feature elimination methods were compared from each dataset. Molecular descriptor results were then manually aligned with each other. 70 features were found to match within all transporters using the top 50 scored ANOVA-f method and 170 using the RFE method. The resulting 70 descriptors were used for the creation of the final models (see Fig. 1). A graphical representation of the workflow can be seen in Fig. 2.

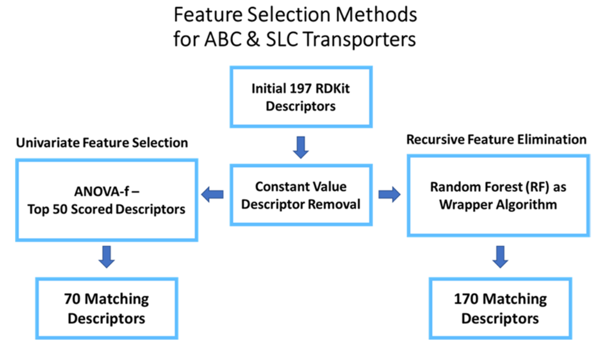
Feature Selection Methods
for ABC & SLC Transporters
Initial 197 RDKit Descriptors
Univariate Feature Selection
Recursive Feature Elimination
ANOVA-f – Top 50 Scored Descriptors
Constant Value Descriptor Removal
Random Forest (RF) as Wrapper Algorithm
70 Matching Descriptors
170 Matching Descriptors

Fig. 1 Schematic overview of the descriptor analysis carried out for both ABC and SLC transporters

Model generation

Four different classifiers, namely logistic regression, support vector machine, random forest, and k-nearest neighbor were used for model generation. The scikit-learn Python library (version 0.24.2) implementations were used to train binary classification models for the six above mentioned datasets.

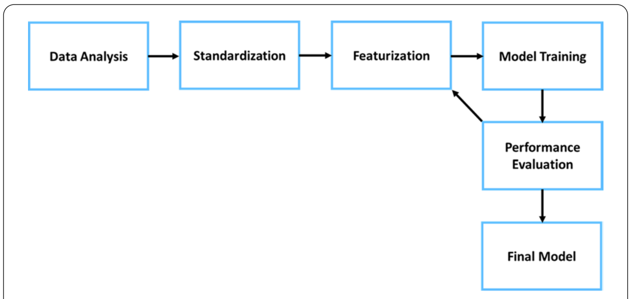

Fig. 2 Graphical Illustration of the workflow for model generation

Hyperparameter grid search

To find the optimal parameters for each classifier, a grid search of the hyperparameters was performed. The following parameters were used:

Logistic regression:
C: 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 1, 10, 50, 100, 1000
max_iter 1,10,100,100,1000,10000.

Support vector machine:
C: 0.01, 0.1, 1.0, kernel: linear.
C: 0.01, 0.1, 0.5, 1.0, 10.0, 50, 100, 1000, kernel: rbf,
C: 0.01, 0.1, 0.5, 1.0, 10, 50, 100, 1000, gamma: 0.0001, 0.001, 0.01, 0.1, 0.5, 1.0, 10.0, 50.0, 100.0, C: 0.0001, 0.001, 0.01, 0.1, 0.5, 1, 10, 50, 100, kernel: rbf

Random forest:
n_estimators: 10, 25, 50, 75, 100, 250, 500.
max_depth: 2, 3, 4, 6, 10, 15, 20.

k-Nearest neighbor:
n_neighbors: 3, 5, 9, 11, 13, 17, 19,
weights: uniform.
distance metric: Euclidean.

Training procedure, cross-validation, and evaluation

In a first step, prediction models were generated simply based on the LiverTox dataset and the settings mentioned above. The performance of these models was compared with the ones obtained from the LiverTox models [47] to validate our approach.

In a next step, the newly collected datasets of the six transporters were used for the training of the actual models. The performance of the models was evaluated using a tenfold cross-validation, and the statistical metrics, such as accuracy, sensitivity, specificity, and balanced accuracy were calculated (see Table 3). For that, the scikit-learn Python library was used. Additionally, an external test set was used to test the new generated models. This test set was collected from ChEMBL28 and PubChem and only data which was novel to the training set was kept.

Applicability domain

Local outlier factor (LOF) as described by Breunig and coworkers was used for the calculation of the applicability domain [52] and as implemented in the scikit-learn Python library (version 0.24.2). In this approach the local densities of the nearest neighbors of a compound are compared to its local densities, and a factor from 0 to 1 is assigned. In brief, if the local density is greater or equal to its surrounding, a compound is considered inside the domain, otherwise it is considered outside the domain.

The following parameters were used:

- 5 nearest neighbors
- novelty = True
- Contamination = 0.1
- Euclidean metric
- Minmax scaled descriptors
- First two principal components were chosen as input

Results

Descriptor analysis

Three different feature selection methods were applied. Variance threshold setting to zero, univariate feature selection using ANOVA-f, and RFE with a random forest wrapper (default settings) were used for the retrieval of the most relevant molecular descriptors from the RDKit module for each dataset. However, once molecular descriptors with constant values were removed, the ANOVA-f and RFE method were applied. For the RFE method, different sets of descriptors were obtained for each transporter. The obtained descriptors were then aligned with each other for the identification of the most frequent descriptors occurring in each dataset. However, 170 descriptors were obtained, which is still considered as a high number considering the basic principle of parsimony in QSAR. Therefore, we conducted in parallel the ANOVA-f approach. Instead of using all scored molecular descriptors, we decided to use only the best 50 scored molecular descriptors for the alignment procedure. As our idea was to keep the number of descriptors as low as possible, we set the threshold to 50 for the alignment, as the performance of the models decreased in individual cases when a lower number was applied. The alignment of each set of resulted descriptors from the six transporter proteins was then conducted. This resulted in a final set of 70 descriptors. The impact of 197, 170 and 70 descriptors on all four models were then examined by calculating the balanced accuracy for each dataset.

Table 3 Statistical metrics for all four models of each dataset

Models	**LR**		**SVM**		**RF**		**k-NN**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
BCRP								
Accuracy	0.73	0.74	0.76	0.67	0.80	0.70	0.76	0.70
Sensitivity	0.75	0.81	0.75	0.69	0.79	0.71	0.79	0.74
Specificity	0.69	0.50	0.77	0.61	0.83	0.63	0.73	0.53
Balanced accuracy	0.72	0.65	0.76	0.65	0.80	0.67	0.76	0.63
F1-score	0.74	0.83	0.77	0.77	0.80	0.79	0.78	0.79
AUC	0.80	0.65	0.83	0.65	0.88	0.67	0.80	0.63
Precision	0.74	0.86	0.79	0.87	0.85	0.88	0.77	0.86
MCC	0.46	0.28	0.53	0.25	0.61	0.28	0.52	0.23
BSEP								
Accuracy	0.84		0.72	–	0.78	–	0.83	-
Sensitivity	0.22	–	0.84	–	0.79	–	0.52	-
Specificity	0.96	–	0.69	–	0.77	–	0.89	-
Balanced accuracy	0.59	–	0.77	–	0.77	–	0.71	-
F1-score	0.30	–	0.54	–	0.57	–	0.53	-
AUC	0.73	–	0.85	–	0.87	–	0.79	-
Precision	0.59	–	0.42	–	0.50	–	0.60	-
MCC	0.28	–	0.44	–	0.49	–	0.45	-
OATP1B1								
Accuracy	0.86	0.38	0.80	0.76	0.85	0.71	0.87	0.71
Sensitivity	0.20	0.33	0.74	0.83	0.63	0.72	0.35	0.67
Specificity	0.97	0.67	0.81	0.33	0.89	0.67	0.96	1
Balanced accuracy	0.59	0.50	0.77	0.58	0.74	0.69	0.65	0.83
F1-score	0.27	0.48	0.52	0.86	0.55	0.81	0.43	0.80
AUC	0.77	0.50	0.83	0.58	0.84	0.69	0.81	0.83
Precision	0.47	0.86	0.40	0.88	0.49	0.93	0.58	1
MCC	0.24	-	0.44	0.15	0.47	0.29	0.34	0.47
OATP1B3								
Accuracy	0.91	0.35	0.84	0.71	0.86	0.59	0.92	0.65
Sensitivity	0.14	0.23	0.81	0.77	0.77	0.69	0.36	0.62
Specificity	0.98	0.75	0.84	0.50	0.87	0.25	0.97	0.75
Balanced accuracy	0.56	0.49	0.83	0.64	0.82	0.47	0.67	0.68
F1-score	0.20	0.35	0.46	0.80	0.48	0.72	0.41	0.73
AUC	0.79	0.49	0.88	0.64	0.89	0.47	0.80	0.68
Precision	0.45	0.75	0.32	0.83	0.35	0.75	0.50	0.89
MCC	0.20	− 0.02	0.44	0.25	0.46	− 0.05	0.38	0.31
MRP3								
Accuracy	0.88	–	0.60	–	0.59	–	0.78	–
Sensitivity	0	–	0.77	–	0.68	–	0.20	–
Specificity	0.99	–	0.58	–	0.59	–	0.86	–
Balanced accuracy	0.5	–	0.67	–	0.62	–	0.53	–
F1-score	0	–	0.43	–	0.37	–	0.21	–
AUC	0.44	–	0.67	–	0.63	–	0.57	–
Precision	0.1	–	0.35	–	0.32	–	0.34	–
MCC	0	–	0.30	–	0.20	–	0.12	–
P-gp								
Accuracy	0.74	0.65	0.72	0.68	0.76	0.68	0.71	0.64
Sensitivity	0.81	0.92	0.72	0.81	0.81	0.92	0.76	0.88

Table 3 (continued)

Models	LR		SVM		RF		k-NN	
	Train	Test	Train	Test	Train	Test	Train	Test
Specificity	0.64	0.28	0.71	0.50	0.70	0.35	0.64	0.32
Balanced accuracy	0.73	0.60	0.71	0.65	0.76	0.64	0.70	0.60
F1-score	0.78	0.75	0.73	0.74	0.79	0.77	0.75	0.74
AUC	0.80	0.60	0.77	0.65	0.80	0.64	0.76	0.60
Precision	0.77	0.64	0.78	0.69	0.80	0.66	0.75	0.64
MCC	0.46	0.27	0.44	0.33	0.53	0.34	0.43	0.25

Test: External Dataset

*Train: tenfold cross-validation

Interestingly, we obtained similar results using only 70 descriptors from the ANOVA-f approach (see Fig. 3) compared to 197 and 170 descriptors obtained from the RFE method. Therefore, we decided to implement the resulted 70 molecular descriptors retrieved from the ANOVA-f approach in the Jupyter Notebook.

Performance of the ML models

For the development of predictive models that can be shared in an easy manner and used for all six transporter datasets, four distinct modeling strategies were applied. Logistic regression-, support vector machine-, random forest- and k-nearest neighbor classifiers were used to train models with the datasets from LiverTox. This concept was used to validate our approach to use it for the actual model generation. The comparison of the performance indicated similar results as shown in the documentation of LiverTox. The support vector and random forest models performed overall better. For the improvement of the models, new datasets for all six transporter datasets were collected from ChEMBL and PubChem. Further, newly published data from Chembl28 and PubChem were used as external datasets, whereas the previous versions were implemented for the training of the four modeling approaches. Again, the three feature selection methods and a hyperparameter search were conducted for the optimal number of descriptors and parameters. Finally, we obtained for each transporter two models, support vector and random forest, with a very similar balanced accuracy ranging from 0.67 till 0.83 for SVM and 0.62 till 0.82 for RF via tenfold cross validation within the various transporters. Overall, we observed that training the models with a subset of all descriptors, addition of new chemical space and the application of the grid search can improve the model performance as compared to the LiverTox models, especially considering the balanced accuracy.

Discussion

Fast and facile model generation for binary classification tasks that are applicable for more than one transporter and additionally allow a retraining of the model is of great interest. Current ML model approaches that were developed are often based on one protein when trying to predict substances as transporter inhibitors or non-inhibitors [32, 53–55]. This makes it harder to generalize models when trying to predict substances for a group of transporters. As the selection of appropriate molecular descriptors can vary from one protein to another, this becomes a quite challenging step. Therefore, we established a Jupyter Notebook that allows the user to generate classification models for six transporter proteins (BCRP, BSEP, OATP1B1, OATP1B3, MRP3, P-gp) without intensive descriptor analysis and

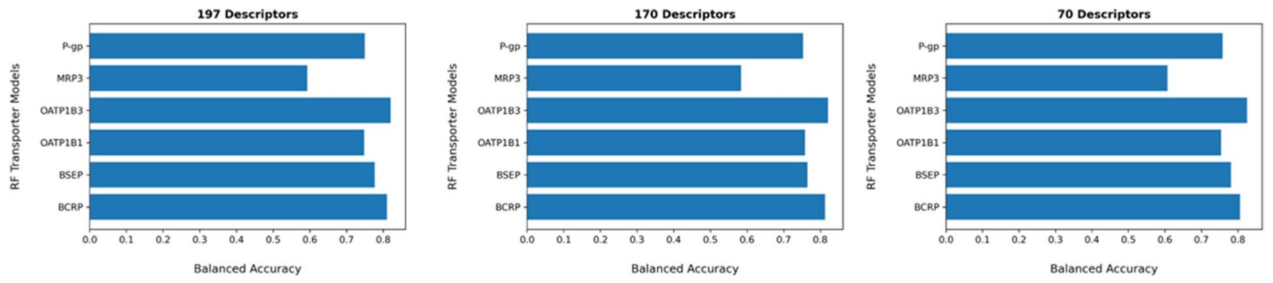

Fig. 3 Comparison of the performances (balanced accuracy) of random forest models using 197, 170 and 70 descriptors

Table 4 Applicability domain estimation for all six transporter protein test sets

Endpoint	LOF result	
	Compounds In-domain	Compounds Out of domain
Breast cancer resistance protein (BCRP)	156	31
Bile salt export pump (BSEP)	9	1
Organic anion transporting polypeptide 1B1 (OATP1B1)	15	6
Organic anion transporting polypeptide 1B3 (OATP1B3)	14	3
Multidrug resistance associated protein (MRP3)	3	0
P-glycoprotein (Pgp)	201	35

hyperparameter search. Moreover, these models can be shared between two instances for additional training. Our analysis indicated that 70 molecular descriptors from the RDKit module can be used for the creation of well-performing predictive models when random forest and support vector classifiers are used. The comparison of the feature selection methods implemented in the scikit-learn Python library showed to be useful for the reduction of descriptors by maintaining a good performance for most of the transporter models and establishing a general set of 70 descriptors for all six transporter proteins. However, in the case of MRP3 an overall low performance was obtained due to a low amount of available data points. The data gathering step revealed that only two transporter proteins, namely BCRP and P-gp, covered a well-balanced number of actives and inactives. This can be visualized when comparing the resulted precision with the remaining datasets that are unbalanced. Both, P-gp and BCRP models, predict correctly 76 to 80% of the cases when a random forest classifier was chosen. Interestingly, for all except MRP3, both good sensitivity and specificity values were retrieved, although the other transporters possess an unbalanced dataset. Only for OATP1B1 a sensitivity lower than 70% was obtained. Best performance was retrieved using the BCRP dataset with a balanced accuracy of 80%, precision of 85%, and sensitivity and specificity values from 79 to 83%. This can be explained by the high number of well-curated data points and the balanced number of actives and inactives in the dataset. Nevertheless, these models can be used for re-training and therefore the performance can increase once more data is available. For each transporter protein a tenfold cross-validation was performed and an external dataset was used for a thorough evaluation, after the final model was trained. A reasonable amount of test compounds was collected for BCRP and P-gp transporters. In the case of OATP1B1 and OATP1B3 more than 17 compounds were retrieved, and less than 10 compounds were obtained for BSEP and MRP3. Therefore, an external validation was meaningful when BCRP and P-gp test sets were evaluated. In both cases, the balanced accuracy, specificity decreased by more than 20% compared to the cross-validation, which still indicated a moderate performance. This could be explained by the fact that 31 compounds from the BCRP and 35 compounds from the P-gp test set were out of domain, when local outlier factor algorithm was used for the applicability domain estimation (Table 4) [52]. Using the same approach for OATP1B1 and OATP1B3, indicated a total of 9 outliers and similar decrease in performance. For the remaining test sets no results could have been retrieved due to the low number of data points obtained from ChEMBL28. Nevertheless, a tenfold cross validation was carried out for each transporter dataset indicating performances close to 80% for five out of six transporter datasets, making it a valuable and feasible tool for the prediction of new data related to both ABC and SLC transporters. Additionally, this approach benefits from the model's ability to be updated and shared in a facile manner using Jupyter Notebook.

Conclusion

In this study, we present a JN which enables the user to generate classification models for six transporter proteins (BCRP, BSEP, OATP1B1, OATP1B3, MRP3, P-gp) based on four different classifiers with pre-selected descriptors and without extensive hyperparameter search. In addition, the notebook can further be used to create models for additional transporters as well as retraining of the existing prediction models using pre-defined descriptors as well as hyperparameters with an extended/novel dataset. The JN can be as well used for educational purposes, especially for the ones interested in the creation of predictive ML models for inhibitory activity predictions.

Abbreviations

BCRP: Breast cancer resistance protein; BSEP: Bile salt export pump; MRP3: Multidrug resistance-associated protein 3; OATP1B1: Organic anion transporting polypeptide 1B1; OATP1B3: Organic anion transporting polypeptide 1B3; P-gp: P-glycoprotein; SMILES: Simplified molecular input line entry system; InChI: International chemical identifier; LOF: Local outlier factor; AUC: Area under the curve; MCC: Matthews correlation coefficient.

Acknowledgements

The authors acknowledge the eTRANSAFE consortium for the support. Disclaimer. This work reflects only the author's views and the JU is not responsible for any use that may be made of the information it contains.

Author contributions

AS developed, implemented, and validated the method. MG and GE: were involved in designing the method and writing the paper. All authors read and approved the final manuscript.

Funding

This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777365. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. The Pharmacoinformatics Research Group (Ecker lab) acknowledges funding provided by the Austrian Science Fund FWF AW012321 MolTag.

Availability of data and materials

The Jupyter Notebook as well as relevant files for the model building can be found in our GitHub repository at https://github.com/PharminfoVienna/Retraining_Notebook.

Declarations

Competing interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 26 December 2021 Accepted: 31 July 2022
Published online: 13 August 2022

References

1. Yang H, Sun L, Li W et al (2018) In silico prediction of chemical toxicity for drug design using machine learning methods and structural alerts. Front Chem 6:30. https://doi.org/10.3389/fchem.2018.00030
2. Klambauer G, Hochreiter S, Rarey M (2019) Machine learning in drug discovery. J Chem Inf Model 59:945–946. https://doi.org/10.1021/acs.jcim.9b00136
3. Vo AH, Van Vleet TR, Gupta RR et al (2020) An overview of machine learning and big data for drug toxicity evaluation. Chem Res Toxicol 33:20–37. https://doi.org/10.1021/acs.chemrestox.9b00227
4. Cases M, Briggs K, Steger-Hartmann T et al (2014) The eTOX data-sharing project to advance in silico drug-induced toxicity prediction. Int J Mol Sci 15:21136–21154. https://doi.org/10.3390/ijms151121136
5. Pastor M, Quintana J, Sanz F (2018) Development of an infrastructure for the prediction of biological endpoints in industrial environments. Lessons learned at the eTOX Project. Front Pharmacol. https://doi.org/10.3389/FPHAR.2018.01147
6. Gaulton A, Hersey A, Nowotka ML et al (2017) The ChEMBL database in 2017. Nucleic Acids Res 45:D945–D954. https://doi.org/10.1093/nar/gkw1074
7. Kim S, Chen J, Cheng T et al (2021) PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 49:D1388–D1395. https://doi.org/10.1093/nar/gkaa971
8. Idakwo G, Luttrell J, Chen M et al (2019) A review on machine learning methods for in silico toxicity prediction. J Environ Sci Health Part C 36:169–191. https://doi.org/10.1080/10590501.2018.1537118
9. Kruhlak NL, Benz RD, Zhou H, Colatsky TJ (2012) (Q)SAR modeling and safety assessment in regulatory review. Clin Pharmacol Ther 91:529–534. https://doi.org/10.1038/clpt.2011.300
10. König J, Müller F, Fromm MF (2013) Transporters and drug-drug interactions: Important determinants of drug disposition and effects. Pharmacol Rev 65:944–966. https://doi.org/10.1124/pr.113.007518
11. Padda MS, Sanchez M, Akhtar AJ, Boyer JL (2011) Drug-induced cholestasis. Hepatology 53:1377–1387. https://doi.org/10.1002/hep.24229
12. Nicolaou M, Andress EJ, Zolnerciks JK et al (2012) Canalicular ABC transporters and liver disease. J Pathol 226:300–315. https://doi.org/10.1002/path.3019
13. Attili AF, Angelico M, Cantafora A et al (1986) Bile acid-induced liver toxicity: relation to the hydrophobic-hydrophilic balance of bile acids. Med Hypotheses 19:57–69. https://doi.org/10.1016/0306-9877(86)90137-4
14. Meier PJ, Stieger B (2002) Bile salt transporters. Annu Rev Physiol 64:635–661. https://doi.org/10.1146/annurev.physiol.64.082201.100300
15. Keppler D (2014) The roles of MRP2, MRP3, OATP1B1, and OATP1B3 in conjugated hyperbilirubinemia. Drug Metab Dispos 42:561–565. https://doi.org/10.1124/dmd.113.055772
16. Briz O, Serrano MA, Macias RIR et al (2003) Role of organic anion-transporting polypeptides, OATP-A, OATP-C and OATP-8, in the human placenta-maternal liver tandem excretory pathway for foetal bilirubin. Biochem J 371:897–905. https://doi.org/10.1042/BJ20030034
17. Raies AB, Bajic VB (2016) In silico toxicology: computational methods for the prediction of chemical toxicity. Wiley Interdiscip Rev Comput Mol Sci 6:147–172. https://doi.org/10.1002/wcms.1240
18. Kluyver T, Ragan-Kelley B, Pérez F, et al (2016) Jupyter Notebooks—a publishing format for reproducible computational workflows. Position Power Acad Publ Play Agents Agendas—Proc 20th Int Conf Electron Publ ELPUB 2016 87–90. https://doi.org/10.3233/978-1-61499-649-1-87
19. Jupyter Notebook (2022) What is the Jupyter Notebook?—Jupyter/IPython Notebook Quick Start Guide 0.1 documentation. In: Online. https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html. Accessed 22 Jul 2022
20. Sydow D, Morger A, Driller M, Volkamer A (2019) TeachopenCadd: a teaching platform for computer-aided drug design using open source packages and data. J Cheminform 11:1–7. https://doi.org/10.1186/s13321-019-0351-x
21. Krauze A, Grinberga S, Krasnova L et al (2014) Thieno[2,3-b]pyridines—a new class of multidrug resistance (MDR) modulators. Bioorganic Med Chem 22:5860–5870. https://doi.org/10.1016/j.bmc.2014.09.023
22. Montanari F, Ecker GF (2014) BCRP inhibition: from data collection to ligand-based modeling. Mol Inform 33:322–331. https://doi.org/10.1002/minf.201400012
23. Hirano H, Kurata A, Onishi Y et al (2006) High-speed screening and QSAR analysis of human ATP-binding cassette transporter ABCB11 (bile salt export pump) to predict drug-induced intrahepatic cholestasis. Mol Pharm 3:252–265. https://doi.org/10.1021/mp060004w
24. Pinto M, Trauner M, Ecker GF (2012) An in silico classification model for putative ABCC2 substrates. Mol Inform 31:547–553. https://doi.org/10.1002/minf.201200049
25. Winter E, Lecerf-Schmidt F, Gozzi G et al (2013) Structure-activity relationships of chromone derivatives toward the mechanism of interaction with and inhibition of breast cancer resistance protein ABCG2. J Med Chem 56:9849–9860. https://doi.org/10.1021/jm401649j
26. Warner DJ, Chen H, Cantin LD et al (2012) Mitigating the inhibition of human bile salt export pump by drugs: opportunities provided by physicochemical property modulation, in silico modeling, and structural modification. Drug Metab Dispos 40:2332–2341. https://doi.org/10.1124/dmd.112.047068
27. De Bruyn T, Van Westen GJP, IJzerman AP et al (2013) Structure-based identification of oatp1b1/3 inhibitorss. Mol Pharmacol 83:1257–1267. https://doi.org/10.1124/mol.112.084152
28. Karlgren M, Vildhede A, Norinder U et al (2012) Classification of inhibitors of hepatic organic anion transporting polypeptides (OATPs): influence of protein expression on drug–drug interactions. J Med Chem 55:4740–4763. https://doi.org/10.1021/JM300212S
29. Juvale K, Stefan K, Wiese M (2013) Synthesis and biological evaluation of flavones and benzoflavones as inhibitors of BCRP/ABCG2. Eur J Med Chem 67:115–126. https://doi.org/10.1016/j.ejmech.2013.06.035

30. Köhler SC, Wiese M (2015) HM30181 derivatives as novel potent and selective inhibitors of the breast cancer resistance protein (BCRP/ABCG2). J Med Chem 58:3910–3921. https://doi.org/10.1021/acs.jmedchem.5b00188
31. Pedersen JM, Matsson P, Bergström CAS et al (2013) Early identification of clinically relevant drug interactions with the human bile salt export pump (BSEP/ABCB11). Toxicol Sci 136:328–343. https://doi.org/10.1093/toxsci/kft197
32. Kotsampasakou E, Brenner S, Jäger W, Ecker GF (2015) Identification of novel inhibitors of organic anion transporting polypeptides 1B1 and 1B3 (OATP1B1 and OATP1B3) using a consensus vote of six classification models. Mol Pharm 12:4395–4404. https://doi.org/10.1021/ACS.MOLPHARMACEUT.5B00583
33. Li XQ, Wang L, Lei Y et al (2015) Reversal of P-gp and BCRP-mediated MDR by tariquidar derivatives. Eur J Med Chem 101:560–572. https://doi.org/10.1016/j.ejmech.2015.06.049
34. Contino M, Zinzi L, Cantore M et al (2013) Activity-lipophilicity relationship studies on P-gp ligands designed as simplified tariquidar bulky fragments. Bioorgan Med Chem Lett 23:3728–3731. https://doi.org/10.1016/j.bmcl.2013.05.019
35. Morgan RE, Trauner M, van Staden CJ et al (2010) Interference with bile salt export pump function is a susceptibility factor for human liver injury in drug development. Toxicol Sci 118:485–500. https://doi.org/10.1093/toxsci/kfq269
36. Hayashi D, Tsukioka N, Inoue Y et al (2015) Synthesis and ABCG2 inhibitory evaluation of 5-N-acetylardeemin derivatives the paper is dedicated to Professor Amos B. Smith, III on the occasion of his 70th birthday. Bioorganic Med Chem 23:2010–2023. https://doi.org/10.1016/j.bmc.2015.03.017
37. Köck K, Ferslew BC, Netterberg I et al (2014) Risk factors for development of cholestatic drug-induced liver injury: inhibition of hepatic basolateral bile acid transporters multidrug resistance-associated proteins 3 and 4. Drug Metab Dispos 42:665–674. https://doi.org/10.1124/DMD.113.054304
38. Ochoa-Puentes C, Bauer S, Kühnle M et al (2013) Benzanilide-biphenyl replacement: a bioisosteric approach to quinoline carboxamide-type ABCG2 modulators. ACS Med Chem Lett 4:393–396. https://doi.org/10.1021/ml4000832
39. Zinzi L, Contino M, Cantore M et al (2014) ABC transporters in CSCs membranes as a novel target for treating tumor relapse. Front Pharmacol. https://doi.org/10.3389/fphar.2014.00163
40. Orlandi F, Coronnello M, Bellucci C et al (2013) New structure-activity relationship studies in a series of N, N-bis(cyclohexanol)amine aryl esters as potent reversers of P-glycoprotein-mediated multidrug resistance (MDR). Bioorgan Med Chem 21:456–465. https://doi.org/10.1016/j.bmc.2012.11.019
41. Dawson S, Stahl S, Paul N et al (2012) In vitro inhibition of the bile salt export pump correlates with risk of cholestatic drug-induced liver injury in humans. Drug Metab Dispos 40:130–138. https://doi.org/10.1124/dmd.111.040758
42. Capparelli E, Zinzi L, Cantore M et al (2014) SAR studies on tetrahydroisoquinoline derivatives: the role of flexibility and bioisosterism to raise potency and selectivity toward P-glycoprotein. J Med Chem 57:9983–9994. https://doi.org/10.1021/jm501640e
43. Reis M, Ferreira RJ, Santos MMM et al (2013) Enhancing macrocyclic diterpenes as multidrug-resistance reversers: structure-activity studies on jolkinol D derivatives. J Med Chem 56:748–760. https://doi.org/10.1021/jm301441w
44. Contino M, Zinzi L, Perrone MG et al (2013) Potent and selective tariquidar bioisosters as potential PET radiotracers for imaging P-gp. Bioorgan Med Chem Lett 23:1370–1374. https://doi.org/10.1016/j.bmcl.2012.12.084
45. Winter E, Devantier Neuenfeldt P, Chiaradia-Delatorre LD et al (2014) Symmetric bis-chalcones as a new type of breast cancer resistance protein inhibitors with a mechanism different from that of chromones. J Med Chem 57:2930–2941. https://doi.org/10.1021/jm401879z
46. Baumert C, Günthel M, Krawczyk S et al (2013) Development of small-molecule P-gp inhibitors of the N-benzyl 1,4-dihydropyridine type: novel aspects in SAR and bioanalytical evaluation of multidrug resistance (MDR) reversal properties. Bioorgan Med Chem 21:166–177. https://doi.org/10.1016/j.bmc.2012.10.041
47. Montanari F, Knasmüller B, Kohlbacher S et al (2020) Vienna LiverTox workspace—a set of machine learning models for prediction of interactions profiles of small molecules with transporters relevant for regulatory agencies. Front Chem 7:899. https://doi.org/10.3389/fchem.2019.00899
48. ChEMBL26.https://doi.org/10.6019/CHEMBL.database.26. Accessed 22 Jul 2022
49. ChEMBL27. https://doi.org/10.6019/CHEMBL.database.27. Accessed 22 Jul 2022
50. ChEMBL28. https://doi.org/10.6019/CHEMBL.database.28. Accessed 22 Jul 2022
51. Landrum G "RDKit: Open-source cheminformatics," can be found under http://www.rdkit.org/. Accessed 22 Jul 2022
52. Breunig MM, Kriegel H-P, Ng RT, Sander J (2000) LOF. Proc 2000 ACM SIGMOD Int Conf Manag data—SIGMOD '00 93–104. https://doi.org/10.1145/342009.335388
53. Jain S, Grandits M, Richter L, Ecker GF (2017) Structure based classification for bile salt export pump (BSEP) inhibitors using comparative structural modeling of human BSEP. J Comput Aided Mol Des 31:507–521. https://doi.org/10.1007/s10822-017-0021-x
54. Prachayasittikul V, Worachartcheewan A, Shoombuatong W et al (2015) Classification of p-glycoprotein-interacting compounds using machine learning methods. EXCLI J 14:958–970. https://doi.org/10.17179/excli2015-374
55. Belekar V, Lingineni K, Garg P (2015) Classification of breast cancer resistant protein (BCRP) inhibitors and non-inhibitors using machine learning approaches. Comb Chem High Throughput Screen 18:476–485. https://doi.org/10.2174/1386207318666150525094503

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.