Artificial Intelligence in the Life Sciences 1 (2021) 100027

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Research Article

Identification of bile salt export pump inhibitors using machine learning: Predictive safety from an industry perspective

Raquel Rodríguez-Pérez*, Grégori Gerebtzoff

Novartis Institutes for Biomedical Research, Novartis Campus, CH-4002 Basel, Switzerland

ARTICLE INFO

Keywords:
Predictive safety
Machine learning
Model evaluation
Decision-making
Drug-induced liver injury (DILI)
Liver toxicity
Bile salt export pump

ABSTRACT

Bile salt export pump (BSEP) is a transporter that moves bile salts from hepatocytes into bile canaliculi. BSEP inhibition can result in the toxic accumulation of bile salts in the liver, which has been identified as a risk factor of drug-induced liver injury (DILI). Since DILI is a frequent cause of drug withdrawals from the market or failings in drug development, *in vitro* BSEP activity is measured with the [^{3}H]taurocholate uptake assay and a half-maximal inhibitory concentration (IC_{50}) higher than 30 μM is advised. Herein, a machine learning classification model was developed to accurately detect BSEP inhibitors and help in the prioritization of *in vitro* testing. Regression models for the numerical prediction of IC_{50} values were also generated. Classification and regression models for BSEP inhibition have been evaluated on realistic settings, which is critical prior to ML-based decision making in drug discovery programs. This work illustrates how predictive safety can help in early toxicity risk assessment and compound prioritization by leveraging Novartis historical experimental data.

1. Introduction

Drug-induced liver injury (DILI) is a main cause of drug withdrawals from the market and drugs failing in development [1,2]. Factors governing DILI have been studied to anticipate which compounds might lead to acute liver toxicity and reduce late stage attrition. Analysis of Liver Toxicity Knowledge Base (LTKB) [3] from the Food and Drug Administration (FDA) as well as further experimental data have provided insights into the main metabolic processes and proteins involved. Bile salt export pump (BSEP) inhibition has been identified as a main risk factor for DILI [4–6] and *in vitro* screening of BSEP inhibition is recommended by health authorities, e.g., European Medical Agency (EMA) [7]. BSEP is an efflux transporter that moves intracellular bile acids from hepatocytes into bile canaliculi (Fig. 1) [8]. This ATP-dependent membrane transport protein is placed in the apical domain of hepatocytes and is the rate-determining step for bile acids secretion [4,8]. Inhibition of this transporter causes an accumulation of bile salts in the liver potentially leading to cholestatic DILI. Interestingly, such build-up of salts might only cause detectable liver injury in humans due to different bile salt composition (more hydrophobic) compared to rodents and dogs, which further supports the importance of measuring *in vitro* human BSEP inhibition [4]. Many drugs causing DILI inhibit BSEP *in vitro* at concentrations relevant to *in vivo* human total plasma steady state drug concentrations [4]. Therefore, *in vitro* BSEP inhibition should be interpreted in context of the *in vivo* exposure and the BSEP safety margin, defined as IC_{50}/C_{max} in plasma, and used in risk assessment in later discovery phases [4]. The pre-clinical safety (PCS) department at Novartis monitors the BSEP safety margin. Low BSEP safety margin relates to higher risk of serious DILI, and conversely. At early stages when exposure is unknown, BSEP $IC_{50} > 30$ μM is recommended based on previous in-house analyses [9].

BSEP inhibition is only measured on a small fraction of promising compounds that have shown desirable physicochemical and biological activity. Thus, an accurate detection of BSEP inhibitors by *in silico* methods would help to prioritize compounds for synthesis and *in vitro* testing. Qualitative and quantitative structure-activity relationship (Q)SAR models have been reported for the prediction of BSEP inhibition [13–17]. Warner et al. proposed a support vector machine (SVM) model based on a data set of 624 compounds (437 for training) represented by physicochemical properties [10]. In the prediction of BSEP activity with a binary threshold of 300 μM, the model achieved an accuracy of 87% and a Cohen's kappa of 0.74 [10]. However, similarly to other studies, model performance was assessed on a random subset of compounds. In drug discovery, compounds are designed and optimized over time, and the chemical or property space under study is constantly evolving. Random data splitting often results in compound analogs being present in both training and test sets. In such cases, generalization ability of the model is not being evaluated and overoptimistic results are obtained compared to prospective model application [11,12]. In another work, a random forest (RF) based on 78 one-dimensional (1D) and two-dimensional (2D) descriptors was applied to classify of BSEP inhibitors ($IC_{50} < 10$ μM) and non−inhibitors ($IC_{50} > 50$ μM) with a larger data set of 838 compounds [13]. The model yielded an accuracy of 88%, with both precision and recall of 77% on an external test set. An impor-

* Corresponding author.
E-mail address: raquel.rodriguez_perez@novartis.com (R. Rodríguez-Pérez).

https://doi.org/10.1016/j.ailsci.2021.100027
Received 29 November 2021; Accepted 29 November 2021
Available online 1 December 2021

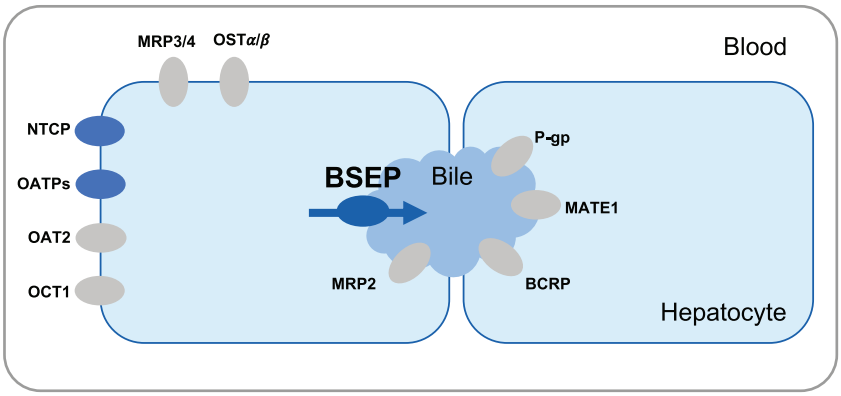

Fig. 1. Role of bile salt export pump (BSEP) in the transport of bile salts in the human liver. The scheme reports the principal transporters of bile salts from blood to hepatocytes (sodium-taurocholate cotransporting polypeptide, NTCP; organic anion-transporting polypeptides, OATPs), and from hepatocytes to the bile canaliculi (BSEP). The ATP-dependent BSEP transporter moves bile salts through the canalicular plasma membrane.

tant practical limitation is that compounds falling into the intermediate activity range were excluded, which simplifies the modeling task [13]. In a recent study, McLoughlin et al. reported an extensive benchmark of predictive models for BSEP inhibition using 1149 compounds [14]. Authors evaluated over 15,500 models using different algorithms, hyperparameters, descriptors, and data splitting approaches, e.g., random and scaffold-based. Superiority of deep learning methods was not observed, and the preferred approach was RF with Molecular Operating Environment (MOE) descriptors, similar to Montanari et al. [13,14].

Overall, most BSEP modeling studies have been characterized by limited data set size or evaluation set-ups that do not demonstrate the possibility of having a generalizable model applicable in the pharmaceutical industry [15–17]. Herein, machine learning (ML) models are developed and validated for the prediction of BSEP inhibitors from a pharmaceutical industry perspective. Emphasis is given to the ML set-up, including data splitting, evaluation metrics or defined categories, which highly depend on the model application and influence decision making.

2. Materials and methods

2.1. Data set description

The in-house BSEP assay measures [^{3}H]taurocholate uptake by inverted vesicles prepared from HEK293 cells over-expressing human recombinant BSEP (hrBSEP) [18]. BSEP inhibition is quantified with the half-maximal inhibitory concentration (IC_{50}). *In vitro* BSEP data was assembled and pre-processed. Measurements without associated structure or report date were discarded, and salts were removed. The data set consisted of 2754 molecules with their corresponding IC_{50} values. Logarithmic transformation was applied to calculate $pIC_{50} = -\log(IC_{50})$ values. The data set had a range of 4.6 log units and a mean pIC_{50} of 4.47. It consisted of 45% BSEP inhibitors ($IC_{50} \leq 30$ μM) and 42% non-inhibitors ($IC_{50} \geq 60$ μM), whereas 13% of the compounds had IC_{50} values between 30 and 60 μM (medium class).

2.2. Molecular representations

Compounds were represented using RDKit 2D descriptors and/or Morgan fingerprints (MFP) [19]. Descriptor vectors had a length of 200 (RDKit version 2019.03.3) [20] and included molecular weight, calculated LogP, topological polar surface area, number of hydrogen donors and acceptors, count of specific substructures, among others. MFP is a hashed fingerprint that encodes circular atom environments up to a given diameter. Herein, the count version of MFP was utilized, where the number of occurrences of a given atom environment or substructure is encoded. MFP had a length of 2048 with radius 2.

2.3. Modeling strategy

ML can be applied in a classification setting, where BSEP inhibitor/non-inhibitor categories need to be defined based on the IC_{50} value, or in a regression setting, where numerical IC_{50} predictions are obtained. For binary classification, the activity threshold was set to 30 μM, which is the BSEP IC_{50} value associated with higher DILI risk according to previous work [9]. To avoid boundary effects due to experimental variability, compounds were divided into three categories: inhibitor ($IC_{50} \leq 30$ μM), non-inhibitor ($IC_{50} \geq 60$ μM), and medium. Final model performance was evaluated on the prediction of these three classes, without excluding the medium class. For regression modeling, pIC_{50} values were used.

Training and test sets were generated according to the measurement date. For classification models, training (60%), calibration (15%), and validation (25%) sets were considered. Models were initially built with the training set and evaluated on the calibration set. Prediction reliability was assessed using output probabilities from binary ML classifiers. High and low probabilities were used to classify compounds into inhibitors and non-inhibitors, respectively. For probabilities associated to lower confidence (closer to the decision boundary), predictions were set to inconclusive. After model optimization and selection, the final model was retrained using 75% of the data, and performance was estimated on the external validation set, which consisted of the most recent BSEP experiments. For regression models, more time-split scenarios were considered with different proportions of training and test data: (i) 75/25%, (ii) 80/20%, (iii) 85/15%, (iv) 90/10%, and (v) 95/5%.

2.4. Machine learning methods

ML methods were applied for classification and regression models. Linear regression methods included Partial least squares (PLS) [21] and Lasso regression [22]. PLS reduces feature dimensionality focusing on covariance, prior to a linear regression model. PLS is applicable to multicollinear features. Lasso regression utilizes shrinkage or L1 regularization and generates sparser models by adding a penalty to rely on less parameters. Non-linear support vector regression (SVR) [23] with a radial basis function was also applied. SVR is an extension of support vector machines (SVM). SVR maps data into a higher dimensional space and derives a regression function of the form $f(\mathbf{x}) = \langle \mathbf{w},\mathbf{x} \rangle + b$. Optimization does not depend on the dimensionality of the input space and only a subset of training compounds is utilized, the so-called support vectors. Cost factors and slack variables also enable SVR regularization. Prior to regression modeling with these methods, feature scaling was applied.

Finally, ensemble models based on decision trees were applied in classification and regression settings. Random forest (RF) [24] and extremely randomized trees (ExtraTrees) [25] create decision trees in parallel, whereas extreme gradient boosting (XGB) [26] trains them sequentially so that one tree learns the errors of the previous one. RF is based on bootstrapping and feature bagging, in order to create trees using distinct data subsets as well as random feature subsets as candidates for node splitting. Compared to RF, ExtraTrees applies more randomization by selecting the feature threshold in node splitting randomly, and does not use bootstrapping.

Table 1
Classification methods comparison. Binary classification performance for models based on 2D descriptors and/or MFP, and tree-ensembles (ExtraTrees, RF, XGB) or k-NN. Reported are the AUC, MCC, and BA for all calibration set compounds and for reliable/conclusive predictions only. The percentage of inconclusive predictions is also shown.

ML algorithm	Features	AUC	MCC	BA	AUC (conclusive)	MCC (conclusive)	BA (conclusive)	% Inconclusive
k-NN	2D	0.609	0.218	60.9	–	–	–	–
Extra Trees	2D	0.769	0.376	67.9	0.841	0.611	76.9	61
Extra Trees	MFP	0.724	0.259	58.3	0.832	0.422	70.7	69
Extra Trees	2D+MFP	0.770	0.384	67.1	0.828	0.436	72.3	53
RF	2D	0.768	0.394	69.4	0.823	0.557	72.8	66
RF	MFP	0.747	0.284	59.0	0.835	0.380	67.9	62
RF	2D+MFP	0.774	0.440	71.3	0.797	0.549	72.4	57
XGB	2D	0.775	0.382	68.9	0.856	0.643	80.9	55
XGB	MFP	0.734	0.394	67.6	0.849	0.682	82.2	59
XGB	**2D+ MFP**	**0.782**	**0.392**	**69.4**	**0.872**	**0.643**	**81.1**	**50**

In classification models, a cross-validation based on the training data was carried out. For RF and ExtraTrees, grid search included number of trees (100, 500), maximum depth (10, 20, 40), maximum features to consider in node splitting (square root or logarithm in base 2 of the number of features), and minimum samples in leaf (2, 4, 8) and split (2, 8, 16) nodes. For XGB models, the number of trees (500, 1000), learning rate (0.01, 0.1, 0.2), maximum depth (6, 12), subsample (0.1, 0.3) hyperparameters were optimized. The optimized XGB model with 2D descriptors and MFP used 500 decision trees, a learning rate of 0.01, maximum depth of 12, and a subsample of 0.3.

2.5. Performance metrics

A variety of metrics were calculated for model evaluation. For classification into BSEP inhibitors and non-inhibitors, relevant metrics include positive predictive value (PPV), negative predictive value (NPV), balanced accuracy (BA), [27] and Matthew's correlation coefficient (MCC) [28].

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$$

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}}$$

$$\mathrm{BA} = \frac{1}{2}\cdot\left(\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}+\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}\right)$$

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

where TP, TN, FP, and FN refer to true positive, true negative, false positive, and false negative, respectively.

For numerical pIC_{50} value predictions, mean absolute error (MAE), mean square error (MSE), and coefficient of determination (R^2) were calculated.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right|$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$$

$$R^2 = 1-\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}_i\right)^2}$$

where y and $\hat{y}$ are the measured and predicted pIC_{50} values, respectively, $\bar{y}$ is the average experimental pIC_{50} value on the set, and n is the number of molecules.

3. Results and discussion

3.1. Classification models

Binary classification models for the prediction of BSEP inhibition were generated, where the IC_{50} value associated with higher DILI risk (IC_{50} = 30 μM) was utilized as the classification threshold [9]. Final model performance was evaluated on the prediction of three BSEP inhibition classes: inhibitor, $IC_{50} \leq 30$ μM; non-inhibitor, $IC_{50} \geq 60$ μM; and medium values.

3.1.1. Binary classification

Different molecular representations and ML algorithms were compared for the prospective prediction of calibration set compounds. Table 1 reports classification model performance using three different types of tree-ensembles: RF, ExtraTrees, and XGB. Compounds were represented using RDKit 2D descriptors, Morgan count fingerprints (MFP) with radius 2, or their combination as a single vector. Standard binary performance metrics such as AUC, MCC or BA were calculated. Different tree ensemble strategies resulted in similar performance, except for RF and ExtraTrees models with only MFP as molecular representation, which resulted in 8–12% less BA. Overall, tree-based ensembles gave promising ranking (AUC ~ 0.78) and binary classification results (MCC ~ 0.44, BA ~ 71%). As a control, k-Nearest neighbors (k-NN) performance is also reported. K-NN results show that the tree-based ensembles have better performance than simply predicting the class of the closest training compounds.

3.1.2. Addition of inconclusive predictions

To improve model performance, only conclusive classification predictions are reported (Fig. 2) [29]. Reliable ML model predictions are typically associated with probability values close to zero and one for non-inhibitors and inhibitors, respectively. By focusing on such predictions and disregarding the rest, BA values increased up to 15%. MCC values improved on average 0.18, with a maximum increase of 0.29 for the XGB model with MFP. Table 1 also reports the AUC, MCC, and BA values for the conclusive predictions only, and indicates the percentage of inconclusive (non-reported) predictions for each model. The model with the lowest number of inconclusive predictions (50%) was based on XGB and 2D descriptors and MFP. It achieved a BA of 81% and an MCC of 0.64. Due to the higher number of reliable predictions compared to other methods with similar performance, this XGB model was the preferred approach. High precision values of 80% and 87% were obtained for the inhibitors (positive predictive value, PPV) and non-inhibitors (negative predictive value, NPV). These metrics are relevant for the practical application of the model. They show that when the model predicted that a compound had a BSEP IC_{50} higher or lower than 30 μM, it was correct 80% and 87% of the times, respectively.

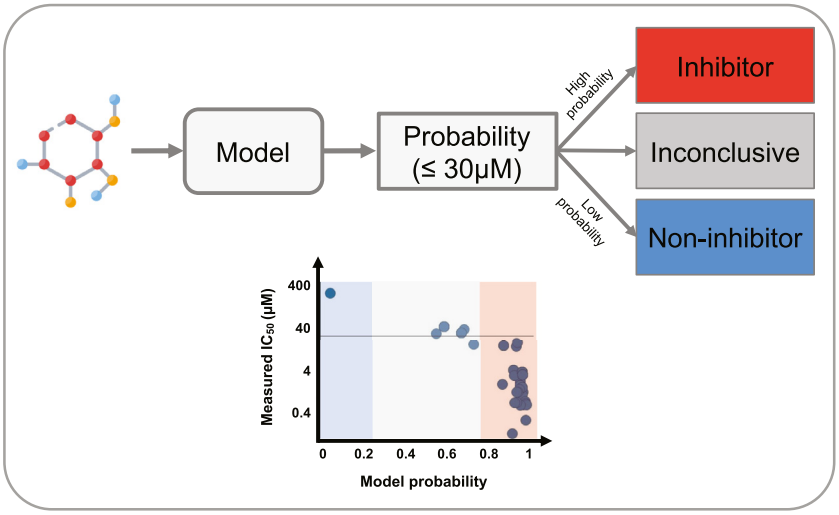

Fig. 2. Workflow for three-class classification. A binary classification model is built with a threshold of 30 μM. Predicted probabilities are utilized for compound classification. Only reliable predictions (probabilities close to zero or one) are reported, and the rest are classified as inconclusive. The plot shows the predicted model probability (x-axis) vs. the measured IC_{50} values for an exemplary chemical series.

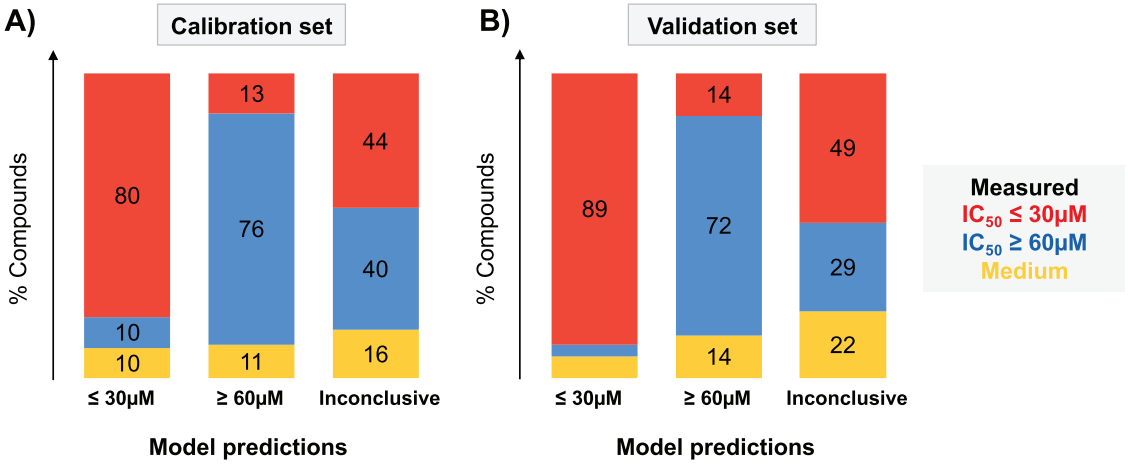

Fig. 3. Prospective validation of the BSEP classification model. Precision values and misclassification errors are reported for the (a) calibration and (b) validation sets. Colors indicate the BSEP inhibitor (≤ 30 μM, red), non-inhibitor (≥ 60 μM, blue) or medium (between 30 and 60 μM, yellow) categories according to the measured IC_{50} values. Model predictions are shown in the x-axis, and the percentage (%) of compounds from each experimental category is reported in the y-axis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.1.3. Three-class predictions

So far, the selected XGB model with 2D descriptors and MFP was evaluated on a binary BSEP inhibition classification with a 30 μM threshold. Some predictions were classified as inconclusive to improve the prediction performance for binary classification. However, experimental values are categorized into three classes to account for experimental variability. BSEP inhibitors and non-inhibitors are defined to have IC_{50} ≤ 30 μM and IC_{50} ≥ 60 μM, respectively, whereas the 'medium' category compounds have IC_{50} values between 30 and 60 μM. Fig. 3 reports the model evaluation on the prediction of these three defined classes. Fig. 3a shows the prospective calibration set predictions from the XGB model with 2D descriptors and MFP. The distribution of BSEP categories based on measured IC_{50} values (≤ 30 μM in red, ≥ 60 μM in blue, medium in yellow) is reported for each predicted category (x-axis). Precision values are 80% and 76% for the prediction of BSEP inhibitors and non-inhibitors, respectively. Misclassification errors were 10% for the inhibitor predictions (false positives) and 13% for the non-inhibitor predictions (false negatives). Since high performance is desired for the prediction of BSEP inhibitors, predicted probabilities strongly indicative of a compound having IC_{50} ≤ 30 μM are the major focus. In that case, when the model predicted that a compound had a BSEP IC_{50} ≤ 30 μM, it was correct 80% of the times, with only 10% of misclassifications.

3.1.4. Prospective evaluation

In the external validation set, the XGB binary classifier trained on a threshold of 30 μM gave a successful ranking with AUC = 0.82. Classification results achieved a BA of 72% and MCC of 0.43. The calibration set, which was previously utilized for model selection, was added to the training set. Such model retraining positively impacted model performance on the validation set, with AUC, BA, and MCC values being 0.84, 76%, and 0.49, respectively. With the addition of inconclusive predictions, BA was 74% but MCC improved up to 0.60. In this case, the model provided more reliable predictions than for the calibration set, since only 34% were inconclusive. Furthermore, the PPV was 89% and the NPV was 86%, indicating that binary BSEP inhibition predictions with a threshold of 30 μM were highly precise.

For the three-class classification predictions, Fig. 3b shows the prospective evaluation on the most recent experiments corresponding to the validation set. The model achieved ~90% precision in the prediction of BSEP inhibitors, with a misclassification error of only 4%, and correctly detected 74% of them (recall). Predictions of non-inhibitors had a precision of 72% (≥60 μM), and only 14% of the BSEP non-inhibitors predictions were incorrect.

Some control calculations were carried out to assess structural diversity of the validation set compared to the training and calibration sets, and the statistical significance of the results. First, the structural diver-

sity of the data set was assessed using the Bemis-Murcko scaffold (BMS) [30] definition. The number of scaffolds and compounds per scaffold were calculated and statistics showed that there is a high percentage of singletons in the data set and only few large clusters, in particular 14 (4) clusters with at least 10 (20) compounds. From the external validation set compounds, 88% contained BMSs that were not contained in compounds from the training and calibration sets. As an alternative way of assessing structural diversity, the maximum and mean Tanimoto coefficient (Tc) [31] values with respect to the training and calibration sets were calculated based on MFP. On the validation set, there was only a small fraction of compounds very similar to the training set. The overall mean similarity was low, with a median Tc = 0.12, and 57% of the compounds had Tc values < 0.35. BMS and Tc results indicate that the model is not only successful in predicting compounds very similar to training data or from the same chemical series (scaffolds). Moreover, permutation tests [32] were carried out to assess the significance of obtained results. XGB binary classification models were built with shuffled labels and the performance was assessed in independent trials. The model trained on data with random labels had a mean AUC of 0.55 and BA of 50%, and did not achieve the performance of the model based on true labels in any case. Therefore, results showed that the obtained AUC and MCC values from the XGB model are not likely obtained by chance (p-value<0.05).

Taken together, the reported BSEP classification model has a high precision in the prediction of BSEP inhibitors. When the model flags a compound as a BSEP inhibitor, there is a high probability that this is correct. Moreover, the model is generally applicable since it has been prospectively evaluated on a relatively large and representative set of compounds, which includes different chemical series.

3.1.5. Comparison to external BSEP models

Model performance was compared to other external BSEP models available at Novartis, such as ADMET Predictor (version 10, Simulations Plus, Lancaster, CA, USA) and eTRANSAFE (version 2020, European consortium). ADMET Predictor provides a binary classification model with an IC_{50} threshold of 60 μM and was trained on a public data set from Morgan et al. [33], which contains 632 compounds. On the other hand, eTRANSAFE consortium built a series of models in an open-source framework named Flame [34], including BSEP activity with an IC_{50} threshold of 20 μM. Training data was collected from three publications, including Morgan et al. [33], Dawson et al. [5], and Warner et al. [10], and comprised 743 compounds. A standard RF as well as a conformal predictor were implemented at eTRANSAFE. Using conformal RF [35], predictions associated with higher uncertainty are discarded aiming at improving the overall accuracy. A non-conformity score based on predicted probabilities is used to estimate uncertainty. However, decrease in compound recall was not accompanied by a substantial improvement in predictive performance, especially for BSEP inhibitors.

As mentioned above, due to the in-house analysis that correlates BSEP safety margin and DILI, in-house models aim at detecting BSEP inhibitors with $IC_{50} \leq 30$ μM, and preferably also $IC_{50} \geq 60$ μM. However, other available models have different definitions. Classification performance was evaluated using the same definition of inhibitors/non-inhibitors that was used during each model building. The real or observed classes were defined according to the 30 μM threshold used for the in-house model, and the external thresholds (60 μM in ADMET Predictor; 20 μM in eTRANSAFE). Under these evaluation conditions, the ADMET predictor model is expected to be superior for the BSEP inhibition threshold of 60 μM and eTRANSAFE for the threshold of 20 μM.

Table 2 reports the validation set performance using the in-house XGB model as well as the ADMET Predictor and eTRANSAFE BSEP models. Some conclusions can be drawn from these binary classification results. Overall, BSEP inhibitors are predicted with more precision than non-inhibitors. Moreover, a threshold of 60 μM would improve the precision for the inhibitors, but the definition would not take into consideration the preferred BSEP inhibition flag from the in-house DILI risk assessment analysis nor a medium class. Despite being trained on public or external data sources, promising results were obtained for ADMET Predictor and eTRANSAFE models. Nevertheless, the in-house XGB model reached higher performance than these models, regardless of the applied thresholds for the definition of inhibitors and non-inhibitors. When only conclusive predictions were considered, performance differences substantially increased. Higher PPV, NPV, MCC, and BA values were obtained for all comparisons. PPV and NPV values for the in-house model reflect the high utility of the model in practical applications since predictions are very likely to be true, particularly 89% and 84% for inhibitors and non-inhibitors, respectively.

Table 2
Evaluation of BSEP models in binary classification. The in-house XGB model based on 2D descriptors and MFP was compared to ADMET Predictor and eTRANSAFE BSEP models. Performance metrics are reported for the external validation set and three different thresholds to define the real BSEP inhibition class: in-house BSEP activity threshold (30 μM) and the thresholds utilized for the other models (60 μM in ADMET Predictor, and 20 μM in eTRANSAFE).

Threshold	Model	PPV	NPV	MCC	BA
30 μM	In-house	87.1	59.7	0.492	**75.9**
	In-house (conclusive)	**88.5**	**84.3**	**0.589**	73.8
	ADMET Predictor	80.5	61.6	0.404	69.4
	eTRANSAFE	81.1	53.9	0.356	68.1
60 μM	In-house	96.1	38.3	0.445	78.8
	In-house (conclusive)	**96.3**	**68.6**	**0.660**	**83.6**
	ADMET Predictor	92.4	43.0	0.416	74.4
20 μM	In-house	79.0	72.6	0.510	**75.2**
	In-house (conclusive)	**80.7**	**90.2**	**0.515**	68.7
	eTRANSAFE	72.9	68.0	0.389	68.5

When using external models, scientists need to be informed about the thresholds that were used to define the categories and assess whether these predictions (even if accurate) are useful for decision making in the context of a given project. If not, model predictions should be evaluated with the desired thresholds, in this case 30 and 60 μM. Fig. 4 reports precision and misclassification errors for the three-class classification setting and the validation set. Precision values with the in-house model improved by 8–9% and 29–34% in the inhibitors and non-inhibitors prediction, respectively. Moreover, less misclassification errors were made, especially for non-inhibitors.

3.2. Regression models

Regression models were also built and evaluated for the prediction of BSEP pIC_{50} values. As schematized in Fig. 5a, models were generated with increasing amounts of training data, and evaluated on test sets composed of the molecules measured next. The most recent 25% of experimental data were divided in subsets of 5% and a letter (from A-E) was assigned to each. Models were evaluated on their corresponding test sets. For instance, model 1 (trained with 75% of the data) can be validated on all test sets (A-E), whereas model 5 (trained with 95% of the data) can only be evaluated on the last test set (E).

3.2.1. Model selection

A set of ML algorithms of distinct complexity were compared with 2D descriptors and MFP as molecular features. Table 3 reports regression model performance on the test set for the distinct methods, considering an 85/15% data split. The best performance was obtained with a non-linear SVR with scaled 2D descriptors, with a coefficient of determination of 0.7, and a mean absolute error (MAE) of 0.363, which corresponds to 2.3-fold on a linear scale. In the prospective test set, 50% of the predictions were within 2-fold of the measured value, and 72% within 3-fold.

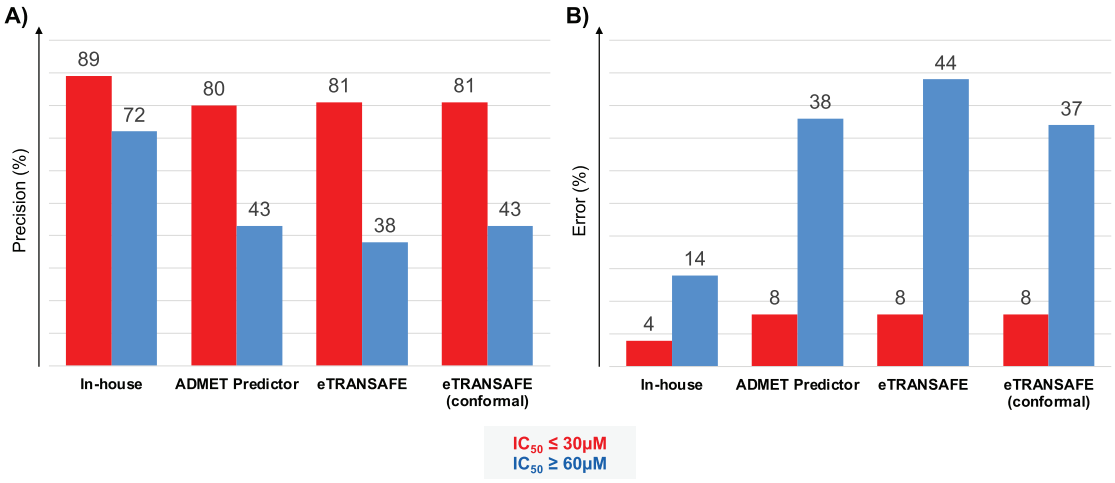

Fig. 4. Comparison to other BSEP models in three-class predictions. Precision (a) and misclassification errors (b) are reported on the validation set. The in-house XGB classification model is compared to BSEP models from ADMET Predictor software and the eTRANSAFE project (with and without conformal predictor). Colors indicate the BSEP inhibitor (≤ 30 μM, red), and non-inhibitor (≥ 60 μM, blue) categories according to the experimental measurement. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Regression methods comparison. Validation set performance is reported for five ML algorithms (Lasso, PLS, RF, XGB, and SVR), and 2D descriptors or MFP. Mean absolute error (MAE), mean squared error (MSE), coefficient of determination (R2), percentage of predictions with 2-Fold and 3-Fold. This validation set consists of 15% of the data.

ML algorithm	Features	MAE	MSE	R2	% 2-Fold	% 3-Fold
Lasso	2D	0.491	0.380	0.471	40	55
PLS	2D	0.435	0.302	0.580	42	63
RF	2D	0.376	0.232	0.678	50	69
RF	2D+MFP	0.381	0.237	0.670	50	68
XGB	2D	0.373	0.230	0.680	49	71
XGB	2D+MFP	0.375	0.232	0.677	49	72
SVR	**2D**	**0.363**	**0.219**	**0.696**	**50**	**72**
SVR	2D+MFP	0.426	0.302	0.580	45	66

3.2.2. Long-term vs. short-term validation

The trade-off between training and test set splitting is a classical issue in ML. More training data generally implies a better model and, depending on the data composition, might lead to higher accuracy or less variance [36]. Similarly, more test or validation compounds improve performance estimation, which is critical to assess the generalization ability of a ML model. Herein, SVR models were evaluated at different points in time with the data not used for model training.

Figs. 5b and 5c report the MAE on each test subset corresponding to approximately 5% of the total data (130–160 compounds), and in the complete test set for each model, respectively. Fig. 5b shows that model performance is positively impacted from model retraining. After the addition of more training compounds, prediction errors were either equivalent or lower for the same test set. Some compound predictions were associated with higher error, regardless of the model used (trained with varying amounts of compound data). For instance, the test set D had lower MAE values for models 1 and 2 compared to test set B. Hence, model errors might vary greatly across test sets and do not always increase with time. Splitting date also has a strong influence on observed performance and different data splits might change the conclusions about data 'modelability' (Fig. 5c). With a conservative chronological data split simulating a long-term validation of 75–25% (to have enough data to assess performance), a regression model might not be prioritized. However, models trained with at least ~85% data provided predicted with MAE lower than 0.4, corresponding to 2.5-Fold in linear scale. Since a single hold-out set might lead to over/under-optimistic estimation of predictive performance, a representative prospective test set and the consideration of both long and short-term validation (i.e. multiple splits) were essential to assess SVR generalization ability in scenarios that resemble future model usage.

Dependency on the data split and benefit of model retraining is also illustrated in Fig. 6, where the predicted versus measured pIC_{50} values are reported for model 1 (evaluated on test sets A-E, and only on test set E) and model 5 (evaluated on test set E). Note that the first model's setting of 75/25% data is equivalent to the global BSEP classification model presented above. Fig. 6a shows that there were some trends between predicted and observed, but the potency of strong BSEP inhibitors is underestimated. The MAE was 0.50, which corresponds to 3.2-fold on a linear scale. Predictions have an associated error that essentially prevents more prediction granularity than a three-class classification. Same trends are observed when the model 1 is evaluated on the most recent 5% of test data. In contrast, Fig. 6c reports a lower MAE of 0.39.

Despite deviations from measured values, Spearman rank correlation values ranged between 0.74 and 0.78. Current literature strategies for reducing BSEP inhibition focus on modifications of molecular weight and lipophilicity. However, the Spearman rank correlation between these properties and BSEP inhibition ranged from 0.56 to 0.61, being slightly lower for LogP. These results suggest that ML models provide a better ranking and more effective guidance for compound modifications than considering these two properties individually.

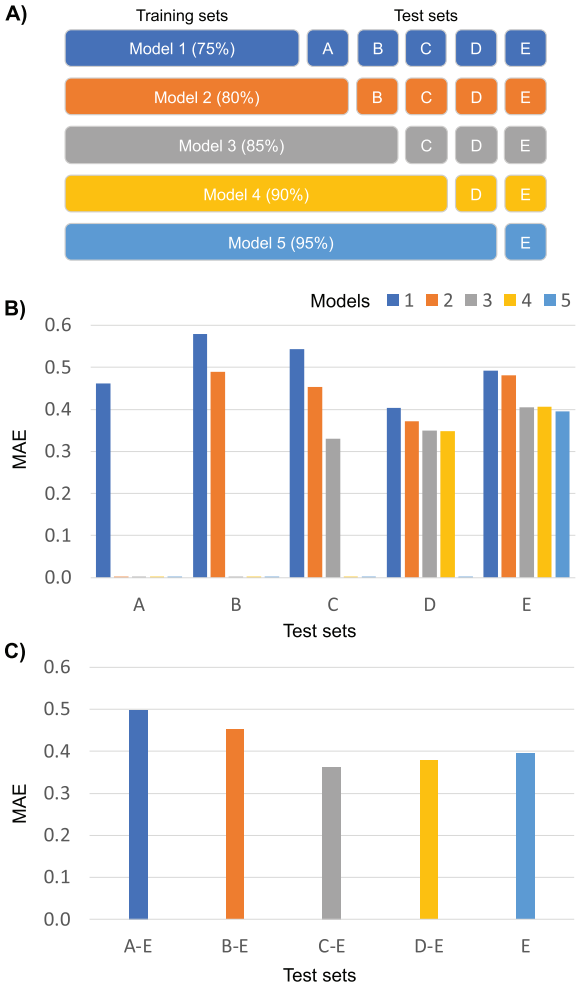

Fig. 5. Regression models evaluation over time. (a) ML set-up for time-gated validation is schematized. Five SVR models (1–5) are trained with different time-splits: Model 1 (75/25%, dark blue), model 2 (80/20%, orange), model 3 (85/15%, gray), model 4 (90/10%, yellow), and model 5 (95/5%, cyan). Test sets are divided into subsets corresponding to 5% of the data and are labeled with a letter (from A to E) according to the measurement date. Models are evaluated on their corresponding test sets, both (b) considering individual subsets of 5%, and (c) all test compounds. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Overall, regression model performance indicates that pIC_{50} predictions might be useful for ranking and compound prioritization. Nevertheless, long-term predictivity was not ensured and there was a tendency to underpredict BSEP inhibition for the most potent compounds. Thus, model application should be limited to short-term predictions or specific projects.

4. Conclusions

Herein, ML classification and regression models were generated for the prediction of BSEP inhibition, an important DILI risk factor. ML models can make predictions for any compound, even prior to synthesis, solely from chemical structure. This provides a valuable opportunity. However, robust model building and evaluation are key for supporting correct decisions in pharmaceutical industry. A predictive safety modeling work has been reported from a practical drug discovery perspective, focusing on the importance of defined classification categories and ML set-up, which greatly influence the quality of ML-based decisions.

A global classification model aiming at detecting BSEP inhibitors with high precision was developed. The model distinguishes between BSEP inhibitors ($IC_{50} \leq 30$ μM) and non-inhibitors ($IC_{50} \geq 60$ μM). To ensure that performance retains the desired precision over time and on new chemical series, unreliable predictions were excluded and classified as inconclusive. Focusing on compounds predicted with higher confidence (lower/higher probabilities) resulted in improved model performance. This aspect is often disregarded in the literature, but it is highly relevant for ML-based decision making in drug discovery. The classification model was prospectively evaluated and achieved ~90% precision in the prediction of BSEP inhibitors with a misclassification error of 4% (false positives). Therefore, when the global model predicts that a compound is a BSEP inhibitor, there is a high probability that it is correct, and is generally applicable across projects or disease areas. Novartis global models are considered as *in silico* assays and predictions are automatically computed and stored for all in-house compounds, including virtual molecules. Consequently, retraining generally occurs when model performance degrades and long-term predictivity as well as general applicability is highly desirable. Although the classification model was successfully evaluated on 22 months of BSEP inhibition experiments, regression models were not able to accurately predict numerical values for this large and representative validation set. Nevertheless, regression models built with at least ~85% of available data were applicable to short-term predictions with errors lower than 0.4 log units. Numerical predictions can be complementary to the global classification results and used for compound ranking, especially for chemical series with high correlation on previously measured data.

Taken together, ML-based predictions can be used as an additional factor providing decision support when progressing compounds during early stages of drug discovery and *in vitro* experiment prioritization. *In vitro* experiments can be limited to those compounds for which models do not give reliable predictions and the most promising compounds for which a numerical value is preferred. Even at discovery phases when safety or liver toxicity are still not being considered, BSEP classification predictions are made available, which helps to de-risk compounds or series at early stages.

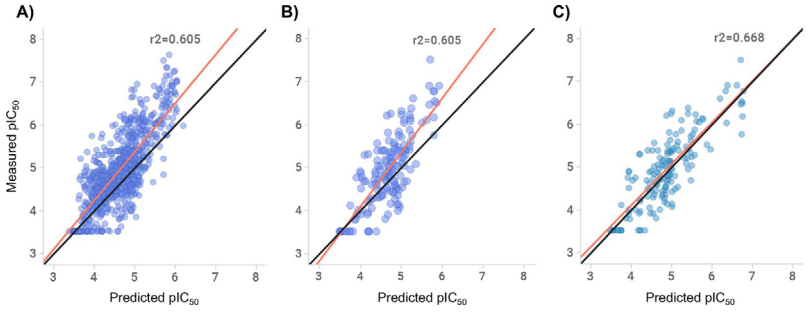

Fig. 6. Predicted versus observed pIC_{50} BSEP. Predictions vs. experimental BSEP pIC_{50} values are shown for model 1 (dark blue) and model 5 (cyan). Model 1 was trained on 75% of the data and predictions are shown for the remaining (a) 25% (test sets A-E), and (b) 5% (test set E), whereas (c) model 5 was trained on 95% of the data and predictions are shown for 5% (test set E). Black lines show the unity line, whereas red lines show the fitting line (with reported coefficient of determination, r2). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Author contributions

All authors contributed to designing and conducting the study, analyzing the results, and preparing the manuscript.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgement

The authors thank Hong Jin and Laszlo Urban for data generation, Jeff Sutherland for scientific discussions, and Seid Hamzic for a review of the manuscript and valuable suggestions. The study was in part carried out using proprietary data. In this case, the Editor has granted a waiver on general data release requirements for Original Research Articles. Such waivers might occasionally be granted if it is judged that a study of comparable magnitude providing new insights and reaching equivalent conclusions could not be carried out solely on public domain data.

References

[1] Chen M, Suziki A, Borlak J, Andrade RJ, Lucena MI. Drug-induced liver injury: interactions between drug properties and host factors. J Hepatol 2015;63:503–14.
[2] Schadt S;, Simon S, Kustermann S, Boess F, McGinnis C, Brink A, Lieven R, Fowler S, Youdim K, Ullah M, Marschmann M, Zihlmann C, Siegrist YM, Cascais AC, Di Lenarda E, Durr E, Schaub N, Ang X, Starke V, Singer T, Alvarez-Sanchez R, Roth AB, Schuler F, Funk C. Minimizing DILI risk in drug discovery - A screening tool for drug candidates. Toxicol In Vitro 2015;30:429–37.
[3] Thakkar S, Chen M, Fang H, Liu Z, Roberts R, Tong W. The Liver Toxicity Knowledge Base (LKTB) and drug-induced liver injury (DILI) classification for assessment of human liver injury. Expert Rev Gastroenterol Hepatol 2018;12:31–8.
[4] Kenna JG, Taskar KS, Battista C, Bourdet DL, Brouwer KLR, Brower KR, Dai D, Funk C, Hafey MJ, Lai Y, Maher J, Pak YA, Pedersen JM, Polli JW, Rodrigues AD, Watkins PD, Yang K, Yucha RW. Can bile salt export pump inhibition testing in drug discovery and development reduce liver injury risk? An international transporter consortium perspective. Clin Pharmacol 2018;104(5):916–32.
[5] Dawson S, Stahl S, Paul N, Barber J, Kenna JG. In vitro inhibition of the bile salt export pump correlates with risk of cholestatic drug-induced liver injury in humans. Drug Metab Dispos 2012;40:130–8.
[6] Thompson RA, Isin EM, Li Y, Weidolf L, Page K, Wilson I, Swallow S, Middleton B, Stahl S, Foster AJ, Dolgos H, Weaver R, Kenna JG. In vitro approach to assess the potential for risk of idiosyncratic adverse reactions caused by candidate drugs. Chem Res Toxicol 2012;25:1616–32.
[7] The European Medicines Agency (EMA) Guideline on the Investigation of Drug Interactions (Adopted 2012).
[8] Stieger B, Meier Y, Meier PJ. The bile salt export pump. Pflugers Arch 2007;453:611–20.
[9] Whitebread S, Fekete A, Jin H, Armstrong D, Urban L. Inhibition of bile salt export pump (BSEP) in relation to systemic exposure: a risk factor for drug-induced liver injury (DILI). J Pharmacol Tox Met 2017;88:215.
[10] Warner DJ, Chen H, Cantin L, Kenna JG, Stahl S, Walter CL, Noeske T. 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 2012;40(12):2332–41.
[11] Sheridan RP. Time-split cross-validation as a method for estimating the goodness of prospective prediction. J Chem Inf Model 2013;53:783–90.
[12] Rodríguez-Pérez R, Bajorath J. Evaluation of multi-target deep neural network models for compound potency prediction under increasingly challenging test conditions. J Comput Aided Mol Des 2021;35:285–95.
[13] Montanari F, Pinto M, Khunweeraphong N, Wlcek K, Sohail MI, Noeske T, Boyer S, Chiba P, Stieger B, Kuchler K, Ecker GF. Flagging drugs that inhibit the bile salt export pump. Mol Pharm 2015;13:163–71.
[14] McLoughlin KS, Jeong CG, Sweitzer TD, Minnich AJ, Tse MJ, Bennion BJ, Allen JE, Calad-Thomson S, Rush TS, Brase JM. Machine learning models to predict inhibition of the bile salt export pump. J Chem Inf Model 2021;61(2):587–602.
[15] Hirano H, Kurata A, Onishi Y, Sakurai A, Saito H, Nakagawa H, Nagakura M, Tarui S, Kanamori Y, Kitajima M, Ishikawa T. 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 2006;3(3):252–65.
[16] Pedersen JM, Matsson P, Bergström CAS, Hoogstraate J, Noren A, LeCluyse EL, Artursson P. Early identification of clinically relevant drug interactions with the human bile salt export pump (BSEP/ABCB11). Toxicol Sci 2013;136(2):328–43.
[17] Ritschel T, Hermans SMA, Schreurs M, van den Heuvel JJMW, Koenderink JB, Greupink R, Russel FGM. *In silico* identification and in vitro validation of potential cholestatic compounds through 3D ligand-based pharmacophore modeling of BSEP inhibitors. Chem Res Toxicol 2014;27:873–81.
[18] Morgan RE, Trauner M, van Staden CJ, Lee PH, Ramachandran B, Eschenberg M, Afshari CA, Qualls CW, Lightfoot-Dunn R, Hamadeh HK. Interference with bile salt export pump function is a susceptibility factor for human liver injury in drug development. Toxicol Sci 2010;118(2):485–500.
[19] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model 2010;50:742–54.
[20] RDKit: Open-source cheminformatics; http://www.rdkit.org
[21] Wold SSM, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst 2001;58:109–30.
[22] Tibshirani R. Regression shrinkage and selection via the lasso. J R Statist Soc 1996;58(1):267–88 Series B (methodological). Wiley.
[23] Drucker H, Burges CC, Kaufman L, Smola AJ, Vapnik VN. Support vector regression machines. In: Advances in neural information processing systems 9. MIT Press; 1996. p. 155–61. *NIPS*.
[24] Breiman L. Random forests. Mach Learn 2001;45:5–32.
[25] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006;63:3–42.
[26] Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2016. p. 785–94.
[27] Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. In: Proceedings of the 20th International Conference on Pattern Recognition (ICPR); 2010. p. 3121–4.
[28] Matthews B. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–51.
[29] Schuffenhauer A, Schneider N, Hintermann S;, Auld D, Blank J, Cotesta S, Engeloch C, Fechner N, Gaul C, Giovannoni J, Jansen J, Joslin J, Krastel P, Lounkine E, Manchester J, Monovich LG, Pelliccioli AP, Schwarze M, Shultz MD, Stiefl N, Baeschlin DK. Evolution of Novartis' small molecule screening deck design. J Med Chem 2020;63(23):14425–47.
[30] Bemis GW, Murcko MA. The properties of known drugs. 1. Molecular frameworks. J Med Chem 1996;39(15):2887–93.
[31] Willett P. Similarity methods in chemoinformatics. Ann Rev Inform Sci Technol 2009;43:3–71.
[32] Ojala M, Garriga GC. Permutation tests for studying classifier performance. J Mach Learn Res 2010;11:1833–63.
[33] Morgan RE, van Staden CJ, Chen Y, Kalyanaraman N, Kalanzi J, Dunn RT, Afshari CA, Hamadeh HK. A multifactorial approach to hepatobiliary transporter assessment enables improved therapeutic compound development. Tox Sci 2013;136(1):216–41.
[34] Pastor M;, Gomez-Tamayo JC, Sanz F. Flame: an open source framework for model development, hosting, and usage in production environments. J Cheminf 2021:13.
[35] Shafer G, Vovk V. A tutorial on conformal prediction. J Mach Learn Res 2008;9:371–421.
[36] Rodríguez-Pérez R, Bajorath J. Influence of varying training set composition and size on support vector machine-based prediction of active compounds. J Chem Inf Model 2017;57:710–19.