Article

Computational Models Using Multiple Machine Learning Algorithms for Predicting Drug Hepatotoxicity with the DILIrank Dataset

Robert Ancuceanu[1], Marilena Viorica Hovanet[1]*, Adriana Iuliana Anghel[1], Florentina Furtunescu[2], Monica Neagu[3,4,5], Carolina Constantin[3,4], and Mihaela Dinu[1,]

[1] Faculty of Pharmacy, Carol Davila University of Medicine and Pharmacy, 020956 Bucharest, Romania; e-mail: robert.ancuceanu@umfcd.ro
[2] Faculty of Medicine, Carol Davila University of Medicine and Pharmacy, 020956 Bucharest, Romania; e-mail: florentina.furtunescu@umfcd.ro
[3] Immunology Laboratory, Victor Babes National Institute of Pathology, 050096 Bucharest, Romania;
[4] Department of Pathology, Colentina University Hospital, Bucharest 020125, Bucharest, Romania;
[5] Faculty of Biology, University of Bucharest, 76201 Bucharest, Romania.

* Correspondence: marilena.hovanet@umfcd.ro.

Abstract: Drug induced liver injury (DILI) remains one of the challenges in the safety profile of both authorized drugs and candidate drugs and predicting hepatotoxicity from the chemical structure of a substance remains a challenge worth pursuing, being also coherent with the current tendency for replacing non-clinical tests with *in vitro* or *in silico* alternatives. In 2016 a group of researchers from FDA published an improved annotated list of drugs with respect to their DILI risk, constituting "the largest reference   drug list ranked by the risk for developing drug-induced liver injury in humans", DILIrank. This paper is one of the few attempting to predict liver toxicity using the DILIrank dataset. Molecular descriptors were computed with the Dragon 7.0 software, and a variety of feature selection and machine learning algorithms were implemented in the R computing environment. Nested (double) cross-validation was used to externally validate the models selected. A number of 78 models with reasonable performance have been selected and stacked through several approaches, including the building of multiple meta-models. The performance of the stacked models was slightly superior to other models published. The models were applied in a virtual screening exercise on over 100,000 compounds from the ZINC database and about 20% of them were predicted to be non-hepatotoxic.

Keywords: DILIrank; DILI; drug hepatotoxicity; QSAR; nested cross-validation; virtual screening; *in silico*;

1. Introduction

Drug induced liver injury (DILI) has been stated as being the most common single cause of drug withdrawal or major regulatory action with respect to a medicinal product (such as a labeling change, "black box" etc) [1,2]. Over 1100 products used by humans on a relatively frequent basis, such as drugs, herbal and other natural products, minerals, "recreational" or illicit chemical substances have been identified as potentially causing liver injuries, although the frequency for some of them is low or very low [3]. The clinical image may be varied, from an increase in the level of liver enzymes to hepatitis, cholestasis or liver cirrhosis, and the diagnosis may be very

challenging [4]. Two distinct ways of DILI occurrence have been described: one direct, intrinsic, for which the risk increases proportionally with the dose (e.g. paracetamol) and one idiosyncratic, which only affects susceptible individuals, is not dose-dependent and is consequently not predictable [5] (e.g. nonsteroidal antiinflammatory drugs [6]). Because of the important impact that DILI may have on patient life, as well as of the regulatory impact it has on a drug, early detection of DILI remains a key concern across all phases of a drug development and substantial efforts are geared towards this goal [7].

The attempts to predict hepatotoxicity based on cell culture tests are prone to failure, because many compounds that *in vivo* exert liver toxicity do not kill hepatocytes *in vitro* or if they do, they cause their death only at unrealistically high concentrations, and this is related to the variability in gene expression of different hepatocyte cell lines [8]. Non-clinical studies performed in animals also have limitations that preclude certainty with respect to their ability to predict liver toxicity in humans. The majority of compounds causing idiosyncratic liver injuries in humans could not be detected as doing so in toxicology studies required by the regulatory framework for new drugs [9]. Current computational methods not only are able to provide at least similar performance to the cell culture or animal methods, but they are considerably cheaper, faster and circumvent ethical issues related to animal models. Moreover, using a computational approach coheres with the current tendency for replacing non-clinical tests with *in vitro* or *in silico* alternatives, mandated by the implementation of the "3R" principle[10], and they are actively encouraged by public authorities such as the European Chemicals Agency (ECHA) or international organizations such as OECD [11]. Furthermore, the use of computational models allows rapid prediction of the activity of a large number of substances in virtual screening exercises, a feat that even with the most sophisticated and automated high-throughput technologies is simply only partially feasible and at huge costs, considering the expensive targets and ligands necessary [12].

Although the number of computational models attempting to predict DILI published up to now is impressive, many were not based on a reference drug list, and developing such a reference list is a daunting task. In the absence of a "gold standard" defining the DILI risk, the different schema and data sources used to compile previously published annotations has been declared by FDA researchers as being "of concern"[13]. A first annotated data set on DILI originated at FDA and was limited to a small number of 287 drugs; more recently (2016), a group of researchers from FDA published an improved annotated list of drugs with respect to their DILI risk, constituting "the largest reference   drug list ranked by the risk for developing drug-induced liver injury in humans", DILIrank [13]. Few Quantitative Structure-Activity Relationship (QSAR) studies have focused exclusively on the DILIrank data set up to date [14];   published studies either have used other data sets [15], only a subset of the DILIrank [16] or have pooled DILIrank with other data sources [17], an approach that may have advantages (increasing the data set and allowing the development of more robust models), but also disadvantages (misclassification bias due to different criteria in annotating drugs from different lists). We have developed a number of QSAR models using a variety of descriptors and machine learning algorithms, and have assembled them to increase the performance. While the majority of the models published up to date use only conventional cross-validation, we have assessed the performance of our models using the state-of-the art method of nested (double) cross-validation, which has better performance and better control of overfitting [18]. Finally, we

have used the models to virtually screen over 100,000 compounds of the ZINC 15 database [19] (the compounds having a name in ZINC) and examined bottom compounds (with the lowest probability of being hepatotoxic) to explore the validity of the models on unknown chemical compounds.

2. Results

2.1. Data set analysis

The final data set included 694 organic molecules with a molecular weight varying from 76 (hydroxyurea) to 7055 Da (ecallantide), with a median value of 329.5 Da and 75% of the molecular weight values less than 430 Da.   The number of atoms per molecule varied between 9 and 934, the median being 43. The number of ring system per molecule ranged between 0 (aliphatic molecules) and maximum 12 (with a median of 2). Of the 694 molecules, only 99 satisfied the Lipinsky's rule of five, divided roughly equally between those of concern and those of no concern (50 and 49, respectively). The variability of the data set by several simple constitutional descriptors or molecular properties is illustrated in Figure 1.
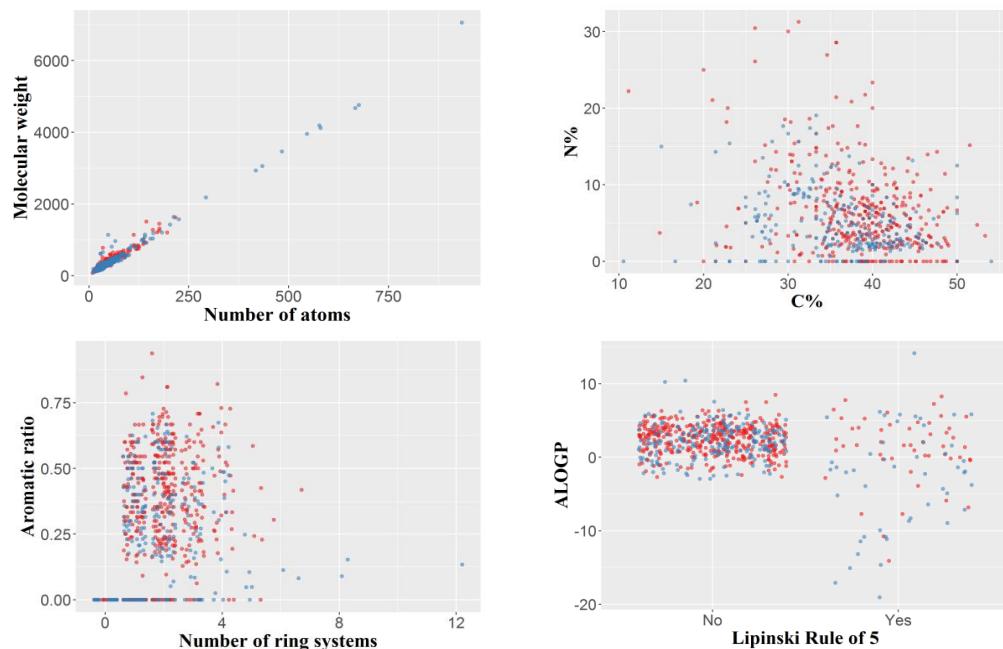

Figure 1. Variability of the data set illustrated by several simple constitutional descriptors or molecular properties. Blue – compounds of no concern; red – compounds of hepatotoxicity concern.

To estimate the chemical diversity in the 694 drugs constituting the data set, a dissimilarity matrix based was computed using the rescaled Gower distance (this metric being appropriate for a combination of continuous and binary data), using 708 most relevant descriptors (obtained by removing auto-correlated and quasi-constant features) (Figure 2). Most compounds have other constituents from the same data set that they resemble (scaled distances under 0.25), but also that they are quite unlike other compounds from the same data set (scaled distances larger than 0.55) (Supplementary Figures S1-S3). The majority of median scaled distance values were about 0.2-0.3, suggesting that the chemical diversity in the data set was somewhat limited (but since all the substances in the data set are approved drugs, this should not be very surprising).

Dissimilarity matrix
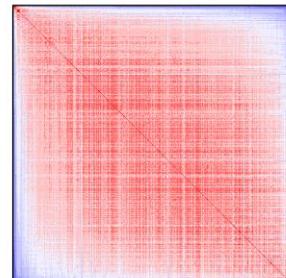
Figure 2. Dissimilarity matrix (based on Gower distance) offering a synthetic image of the chemical diversity in the data set.
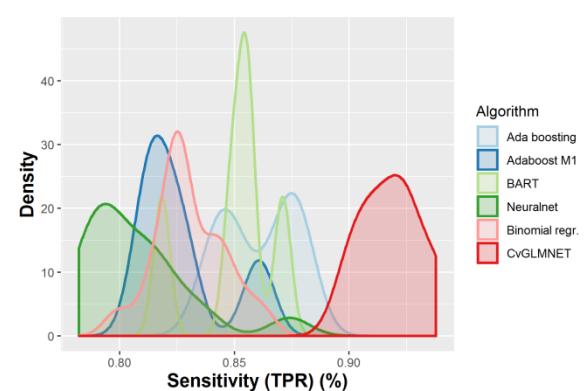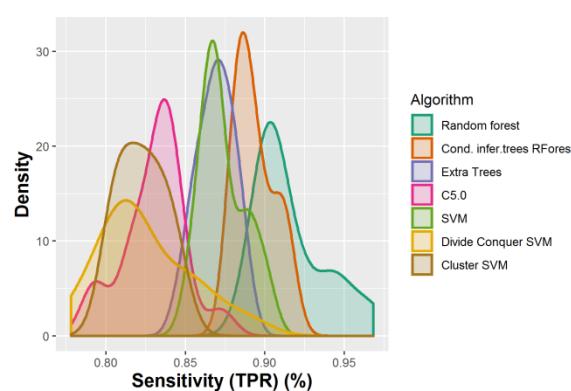

Figure 3. Performance of 165 QSAR models in terms of sensitivity.
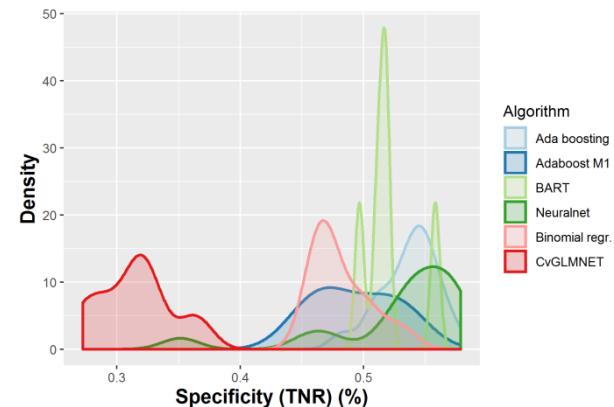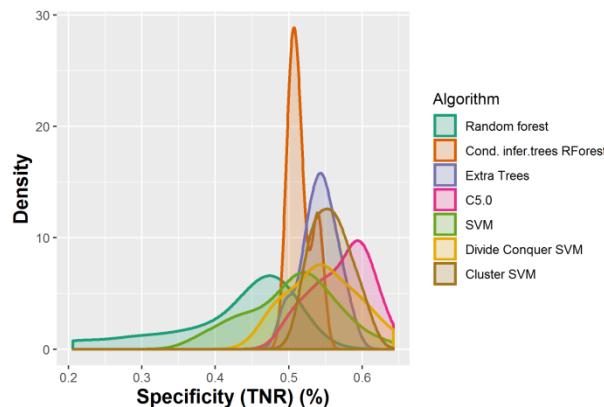
Figure 4. Performance of 165 QSAR models in terms of specificity.
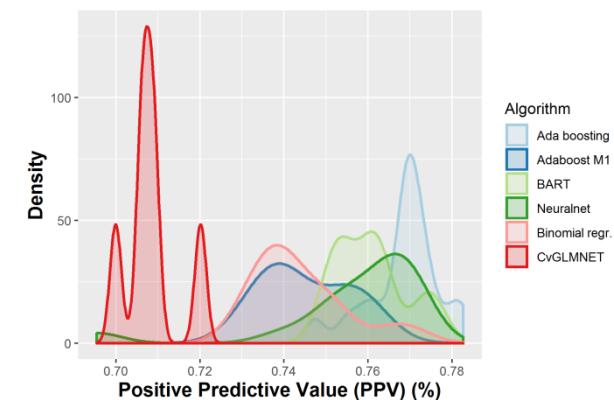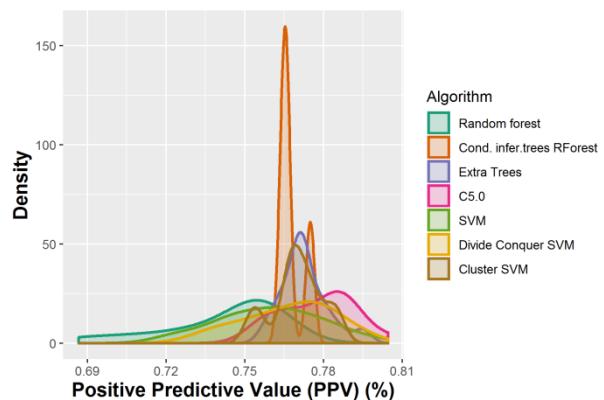
Figure 5. Performance of the 165 QSAR models in terms of positive predictive value.
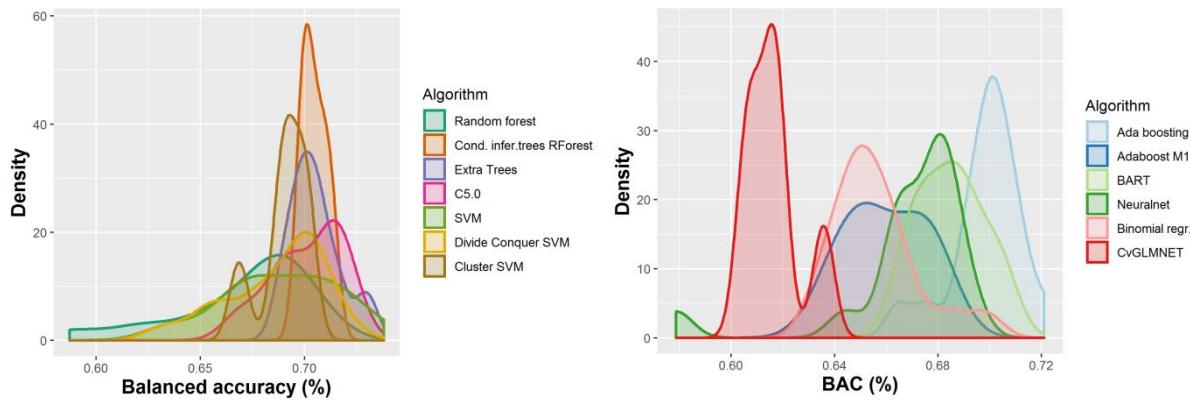
Figure 6. Performance of 165 QSAR models in terms of balanced accuracy.

2.2. Performances of models

A number of 267 different QSAR models were built, with different feature selection algorithms and machine learning techniques, of which 79 were selected for assembling by stacking. The performance of the majority of algorithms (in 165 models, each using 50 features for classification) in terms of sensitivity, specificity and positive predictive values, in nested cross-validation, is shown in Figures 3-6. Whereas certain algorithms (logistic regression, gradient boosting machine) did not manage to model well the hepatotoxicity, many of the algorithms attempted were useful in building models with a reasonable good performance.

The majority of those models had good or very good performance in terms of sensitivity, which was in most cases over 80% and up to 95%; this good sensitivity, however, came at a cost in specificity, which varied mostly around 50% and in best cases reached or slightly overpassed 60%. This means that about half of the non-hepatotoxic substances are likely to be predicted as hepatotoxic, although they are not so. However, we preferred to sacrifice to a certain extent specificity for sensitivity, because our purpose was not to mislabel a hepatotoxic substance as innocuous, rather than the vice versa; on the other hand, there is considerably high confidence that a substance is not hepatotoxic if predicted not to be so. The positive predictive value (PPV) was reasonably good, for most models around 77%. The balanced accuracy (BA) was also reasonably good, 79 models (those selected for stacking and prediction use) having a BA higher than 70%, but in only six models was BA higher than 72% and in all cases it was lower than 73%.

We first assembled the best models through a simple majority vote of binary predictions; this ensured a balanced accuracy of 72.8%, a sensitivity of 89.0% and a specificity of 56.5%. Assembling the models based on the mean probability of all models and a decision threshold of 50% resulted in a balanced accuracy of 72.2%, a sensitivity of 88.3%, and a specificity of 56.1%. Using the same 50% threshold and the median probability values slightly improved the performance, but it was not better than that based on the majority vote (balanced accuracy 72.6%, sensitivity 88.8%, and specificity 56.5%). Changing the probability threshold value to 0.67 (instead of 0.5) and using median predicted probabilities leads to the best performance in terms of balanced accuracy (74.6%), with a lower sensitivity (76.0%) and improved specificity (73.2%).

The two best-performing meta-models built by applying the random forest classifier to the binary predictions of 50 models (selected by applying the same feature selection algorithms) and the maximum daily dose, had a balanced accuracy of 74.38% and 74.20%, respectively; both these two meta-models had high sensitivity (89.68% and 89.71%) and thus low false negative rates (10.32% and 10.29%), being therefore particularly useful to ascertain whether an unknown compound is devoid of liver toxicity properties. Assessing each of these meta-models with different random seed numbers slightly decreased the performance for one of them and increased it for the other (mean balanced accuracies for five repeated runs with different seed numbers were 73.56% and 74.27%, respectively; standard deviations 0.47% and 0.44%). The inclusion of dose among the predictors in the meta-models only slightly (if at all) increased the performance compared with the meta-models built without the dose, but we preferred to include it on the basis of domain knowledge [20,21]. Meta-models built similarly with SVM, knn, IBk and naïve Bayes algorithms had slightly lower performance in terms of both balanced accuracy and sensitivity than those build with random forests.

Using the predicted probabilities to build meta-models with random forests in a similar way with 50 models and the maximum daily dose as features did not improve the performance in terms of both balanced accuracy and sensitivity, but the computing time increased about five times. Meta-models built with naïve Bayes on output probabilities had the highest balanced accuracy (mean of five runs with different seeds 74.64%, standard deviation 0.20%), but lower sensitivity (83.0%) and higher specificity (66.3%). Using other algorithms (knn, IBk, svm, ksvm, C5.0) for the construction of meta-models had very similar results, with resembling performances for each selection algorithm (70.69-74.83% balanced accuracy, 82.81-90.32% sensitivity, 54.90-64.47% specificity). Adding the typical duration of treatment (in days) as an additional feature had a minimal effect on the meta-model performance; the number of days up to the first occurrence of liver toxicity might be more relevant, but we could not collect data for each drug substance in the data set for this variable. The feature selection algorithms applied to the 72 models outputting probabilities identified as the most important for prediction the following classification algorithms: ksvm, svm, Adabag Boosting, kknn, IBk, random forest based on conditional inference trees, and ada, mostly as applied to the subsets selected by the OneR association rule and the "randomForestSRC_var.select" of the randomForestSRC package [22].

2.3. y-randomization test

The y-randomization test showed that in all cases the performance was considerably worse after the scrambling of the response variable. In all cases the balanced accuracy was close to 50% in the case of the scrambled data sets, whereas it was generally over 70% in the case of the genuine models (figure S4). The other parameters were also considerably worse when compared with the genuine models: the AUC was close to 50%, whereas in many cases the true positive rate (TPR) and false positive rate (FPR) were 100% (which means that all compounds were classified into a single class). The repeated similar performance of the randomized data sets, for different classification algorithms and feature selection algorithms confirms that our models are not the result of mere chance, but reflect a genuine

relationship between the chemical structure as measured through the molecular descriptors used and the degree of DILI concern.

2.4. Descriptors associated with hepatotoxicity

Although multiple QSAR models have been developed for DILI substances, often the articles published were focused more on the performance of the models than on the discussion of the descriptors that are associated with an increased or lower risk of liver toxicity. In order to understand the most important features/descriptors associated with DILI, we examined the first five descriptors identified by each selection algorithm. We therefore computed the frequency with which descriptors occurred among the most important 5 features for each of the selection algorithm and those that occurred at least twice are shown in Table 2. A higher mean atomic polarizability tended to associate with a higher DILI concern; similarly, a lower percentage of hydrogen in the molecule and a lower Geary autocorrelation of lag 1 weighted by mass tended to associate with a higher risk of hepatotoxicity.

Table 2. The most important molecular descriptors associated with DILI by the 17 feature selection algorithms used

Descriptor	Interpretation	Descriptor block (group)	Frequency occurring among the first 5 most important features
Mp	mean atomic polarizability (scaled on Carbon atom)	Constitutional indices	12 (70.59%)
H%	percentage of H atoms	Constitutional indices	12 (70.59%)
GATS1m	Geary autocorrelation of lag 1 weighted by mass	2D autocorrelations	12 (70.59%)
SpPosA_B(m)	normalized spectral positive sum from Burden matrix weighted by mass	2D matrix-based descriptors	10 (58.82%)
MLOGP	Moriguchi octanol-water partition coeff. (logP)	Molecular properties	4 (23.53%)
PCR	ratio of multiple path count over path count	Walk and path counts	3 (17.65%)
totalcharge	total charge	Constitutional indices	2 (11.76%)
SM1_Dz.m.	spectral moment of order 1 from Barysz matrix weighted by mass	2D matrix-based descriptors	2 (11.76%)
SIC1	Structural Information Content index (neighborhood symmetry of 1-order)	Information indices	2 (11.76%)

2.5. Virtual screening

Besides using the nested cross-validation procedure, which offers considerably stronger safeguards than merely using an external hold-out sample for external validation, we would have liked to test our stacked models on an independent external data set. For reasons shown in the Discussion section, this was an almost impossible mission; we therefore chose to apply the models in a virtual screening exercise on 104,619 compounds from the ZINC database, so as to identify compounds having a high probability of being devoid of liver toxicity properties. 19.92% of the whole data set (20,835 substances) were predicted by the assemble of 72 models based on the mean probability to be non-hepatotoxic and 20.08% (21,012 substances) were predicted by the 72 models based on the median probability to be non-hepatotoxic. The false negative rate of the stacked models, using the average probability of hepatotoxicity for the 72 models was 11.7%, whereas the negative predictive value was 72.6%, which means that if the same proportions between hepatotoxic and non-hepatotoxic compounds was present in the tested data set, one should expect that 72.63% of the compounds predicted to be non-toxic should indeed be non-toxic. We show in Table S1 the first 2000 such compounds predicted to be devoid of liver toxicity, sorted by the predicted probability (i.e. the first 2000 compounds with the lowest probabilities of being hepatotoxic). As discussed in the next section, all or almost all of these 2000 compounds were in the applicability domain for at least a fraction of the models used. A short look over these compounds shows that they include many oses, polyols, short peptides, vitamins, various hydrosoluble compounds, and this makes credible those predictions, at least for their majority.

2.6. Outliers, applicability domain and wrongly classified drugs

An outlier may be defined as "an observation in a data set which appears to be inconsistent with the remainder of that set of data" [23]. Outliers may lead to wrongly specified models and wrong results, but they may be as well carriers of important information [23], therefore a decision to remove them should be based on well-founded reasons (e.g. obvious recording error), and not on the mere intention of having well-performing models. A variety of model-dependent (parametric) and model-independent (non-parametric) methods have been advanced in the literature, although the comparative performance is difficult to estimate, and different benchmarking studies compared different algorithms and sometimes with inconsistent results [24,25]. Because in one complex benchmarking study [24] the isolation forest (IFOREST) and subspace outlier detection (SOD) algorithms were among the best (IFOREST had the highest performance), we used the two algorithms to examine the potential outliers and how well they were predicted by the models. The IFOREST algorithm identified no obvious outlier (for this algorithm potential outliers have scores close to 1, whereas for all observations of the data set the scores varied between 0.30 and 0.55). For the SOD algorithm we used a 5% threshold for the definition of outliers (thus identifying 35 outliers), and over three quarters of them (78.49% on average, s.d. 5.16%) were correctly classified by the assemble of models selected.

The "applicability domain" (AD) is a tool used to assess whether a QSAR model may be employed to predict in a valid manner the class label of a test compound; such a "prediction" is only valid if the assumptions on which the model was built are still met for the test compound [26]. If the prediction exercise involves an extrapolation from the feature space, the result of this exercise cannot be relied

upon. Hence, evaluating the AD for a specific model is of key significance if that model is to be put to use for prediction purposes, and a great variety of methods have been advanced in the literature in this connection, each with its own strengths and shortcomings [27]. Our approach of normalization and capping extreme values to two standard deviations was meant to ensure a broad AD for untested substances and we assessed this on a subset consisting of the first 2000 substances from the ZINC database, for which there is a high probability that they are not hepatotoxic (the first 2000 compounds with the lowest computed probabilities of being hepatotoxic). The method of F. Sahigara et al. (2013), which uses different decision thresholds for each test compound, identified only a small number of compounds that were out of AD, and only for a subset of the models. The application of this method, therefore, led only to small changes in the average probability predicted for each compound, with no change in the predicted label. The same was true for the use of the INFLO and COF methods. Although leaving out the models that were outside of the AD did not change the classification of any of the 2000 compounds analyzed, in some cases, particularly in the case of using the COF method, leaving aside some of the models (that were outside AD) tended to slightly increase the average probability of being hepatotoxic, but in no case did it reach the 50% so as to change the classification.

We analyzed the hepatotoxic compounds from the DILIrank data set that were wrongly classified by the majority of models and identified a number of 49 such substances out of 447 compounds with different levels of hepatotoxicity concern (i.e. 10.96% false negative rate). Among these, ethambutol was predicted wrongly by all models (as non-toxic, although it is labeled as of "most concern"); in other words, all models predict ethambutol as non-hepatotoxic, although it has been categorized in the DILI Rank data set as being of high concern. This is in line with some medical papers that consider ethambutol as non-toxic [28]. According to LiverTox, despite over a half of century of use, ethambutol has been connected to clinically evident liver injury in a very small number of case reports, but some of these were convincing (re-occurrence by rechallenge) [29]. This prompts us to hypothesize that such rare cases of confirmed hepatotoxicity might have been related to an impurity present in the active substance or one of the excipients, although it may as well be possible that the models were just unable to classify it correctly. Perhexiline, daptomycin, and amphetamine, were correctly classified by a single model out of 79, terbutaline by two models, and acarbose by only three. All other compounds misclassified by the majority of the models are shown in Table S2. Only 14 of these 49 compounds (28.57%) were detected as outliers for at least one of the models, a fact that indicates that in the majority of cases the wrong classification is not the consequence of different chemical features but rather of the model limitations.

3. Discussion

Predicting hepatotoxicity of chemical substances from their chemical structure is an intimidating task, because the mechanisms by which different substances cause liver toxicity may be varied and they are often not understood at all. Moreover, the hepatotoxic substances differs in their liver toxicity profile with respect to the doses or duration needed for the toxic effect to get manifested, as well as in the clinical severity of the hepatotoxicity (varying from slight increases in transaminase levels to fulminant hepatitis requiring an emergency liver transplant). Dactinomycin, for instance, may manifest its liver toxicity at total doses of about 1 mg/day (10 to 15 mcg/kg) [30], whereas

4-aminosalicylic acid has been used at daily doses 12,000 higher [31]. Hepatic injury may be started by paracetamol in 24 to 72 hour after a single overdose [32], in the case of allopurinol hepatitis usually develops within the first month of treatment, whereas clinically symptomatic liver injury is seen with perhexilin only after several months or years of treatment [33]. Drug-induced hepatotoxicity may be predictable (dose-dependent, replicable in animal experiments, and with a short onset), but in most cases it is idiosyncratic (non-predictable) [34]. The latter is often classified in three main mechanistic patterns hepatocellular, cholestatic, and mixed [35]. Other mechanistic classifications are considerably more nuanced, distinguishing between immune mediated and non-immune mediated hepatitis, non-alcoholic steatohepatitis, immune and non-immune mediated cholestatic injury, fibrosis/cirrhosis, granulomas (allergic in nature), microvesicular steatosis, vascular lesions, phospholipidosis, and neoplasms [36]. Considering this diversity of doses, latencies and mechanisms, the chemical diversity of the drugs currently used in therapy, as well as their limited number, predicting hepatotoxicity from the structure alone is fraught with difficulties.

Multiple studies have concluded that the agreement between the results of animal liver toxicity models and human outcomes is low, and this may contribute to a high number of cases where hepatotoxic potential of a drug is only detected in the late stages of clinical development [37]. To further complicate modeling, classifying a known drug as being hepatotoxic or not is not a simple exercise, because for the majority of medicines the liver toxicity is idiosyncratic and if the toxic events are relatively rare, they cannot be detected in clinical trials, but only through post-marketing studies, helped by the spontaneous reporting pharmacovigilance systems, which in their turn tend to suffer from underreporting [13]. Taking these factors into account, as well as the diversity of clinical forms of DILI and the ambiguities affecting the causality assessment, the fact that data from animal studies may not necessarily be relevant for the classification of a drug as hepatotoxic or not, different sources may classify differently specific drugs (i.e. the same drug may be considered hepatotoxic in one study/list, but non-hepatotoxic in another). For instance, leuprolide has been classified as of "Less DILI concern" in DILIrank, whereas the LiverTox website (produced by the National Institute of Diabetes and Digestive and Kidney Diseases, NIDDK) states that "Despite use for several decades, leuprolide has not been linked to convincing cases of clinically apparent liver injury. Routine monitoring of patients for liver test abnormalities is not recommended. Likelihood score: E (unlikely cause of clinically apparent liver injury)" [38]. Morphine and codeine are listed by DILIrank as of no DILI concern, whereas a recently published list of (mostly herbal) ingredients with potential hepatotoxicity or hepatoprotection, labels both of them as a hepatotoxic [39]. The same list labels lamivudine and metformin as hepatoprotective [39], whereas DILIrank classifies them as hepatotoxic of less concern.

The DILIrank data set is based on a "refined annotation schema by weighing evidence of causality to overcome inherent deficits in drug labeling and improve the accuracy of DILI annotation" [13]. When we finished the modeling exercise, it was still the best annotated list of drugs classified by their hepatotoxic potential, based on clinical considerations, that was publicly available. When we were in the final stages of drafting this paper, an improved version with a larger number of drugs was just published by the same group of FDA-affiliated authors, under the name DILIst [37].

Predicting liver toxicity from chemical structure has been a preoccupation of over two decades, first starting somewhat timidly with local models [40,41], to move later also to global models [42,43].

Many of the previous models used relatively small size datasets (less than 400 compounds) [44,45]; studies based on a data sets larger than DILIrank have also been published [11,17,46–48], but the authors of DILIrank used a methodology that in theory at least, was superior and more consistent with the totality of available clinical data. The total DILIrank data set includes 1036 compounds, but 254 were classified as of "ambiguous DILI-concern", because the causality evidence was limited; besides, it took great care in defining DILI negatives, which varied among four large sources previously studied, labeling some previously compounds considered of no DILI concern as of less DILI concern. Thus, although somewhat smaller than other published DILI datasets, the DILIrank had a smaller probability of misclassification, and this prompted us to prefer it to other larger data sets available in the literature.

All or almost all studies published up to date have used a training and a testing data set, and most often also a holdout (external validation) data set, in most cases quite small. None of the previously published studies have used a nested-cross validation approach. It has been the tradition, in the field of machine learning, to divide the whole data set in a training subset (about 70-80%) and a test set (about 20-30%); the models were developed, hyperparameters tuned and performance evaluated on the training subset, and the selected model was then tested on the test set (holdout). In this approach, it is still possible to have a model with good performance on the holdout test by mere chance. The nested cross-validation actually splits the data set in training and testing sets multiple times, each time the best model being tested on an external data set that has not been contaminated by the training samples. The main difference is that whereas in the traditional approach a single holdout test was kept for assessment, in the nested CV the process is repeated multiple time (in our case 10 times). The performance on the training data set is of little interest, because it is known that the available algorithms can overfit and have apparent very high performances. The performance on the test data sets should also be of little interest, because it is also quite common for the algorithms used to get good performance on the test set by mere chance. The performance as measured on external validation data sets is the most important, and it is interesting that whereas models published up to now have reported good and very good performances on the training or test subsets, the performance on external data sets was much lower. A review found in 2014 that in models published up to that time point the external validation data sets had been "quite small" (20-50 drugs), and that for the larger external data sets the model performance seemed to be "less favorable" [49]. One of the best performing ensemble models recently published used three external validation data sets and had a pooled balanced accuracy of 71.60% [17]. Although based on a different data set (DILIrank), our models compares favourably, with a balanced accuracy in the nested cross-validation (i.e. average of 10 external data sets) slightly superior for several meta-models attempted, higher than 74% (as shown in the results section).

We were interested to compare our results with those of the PROTOX II (http://tox.charite.de/protox_II/index.php?site=compound_input), but such a comparison proved not to be realistic. Our predictions were all as part of an external data set (we used the predictions in the outer test set of the nested-cross-validation loop), whereas in the case of PROTOX II, based on a training set of 850 compounds and an external data set of 95 compounds [50], the majority of compounds are likely to be part of the training set, not really predictions. Moreover, the labels used in the training set of PROTOX II seems to be very different from those used in the DILIrank; whereas

there is a large agreement between our results and the PROTOX II for the compounds classified as of most concern by DILIrank, the large majority of compounds labeled by DILIrank as of less concern, are "predicted" by the PROTOX II to be "inactive". This is obviously not due to a failure of the PROTOX II in prediction, but rather in the different labels used in its training dataset. Purely for information purposes we show the results of the PROTOX II in comparison with our predictions in the external loop in Table S2.

Among the most important chemical descriptors associated with liver toxicity in our study were the mean atomic polarizability (Mp), the percentage of hydrogen atoms (H%), the Geary autocorrelation of lag 1 weighted by mass (GATS1m), normalized spectral positive sum from Burden matrix weighted by mass (SpPosA_B(m)), and Moriguchi octanol-water partition coefficient (MLOGP). Atomic polarizability, identified as important by the majority of selection algorithms used, was previously shown to associate with renal toxicity of drugs [51], but it was also used in other liver toxicity models, as was logP [17,52], the latter measuring lipophilicity, which was shown to correlate with DILI [21]. GATS1m is less intuitive than the constitutional descriptors, but it has also been reported in a different publication as an important descriptor for the liver toxicity of drugs [53]. SpPosA_B(m) is also less easily interpretable, and we did not find any previous use in other QSAR models of hepatotoxicity.

Many papers have developed QSAR DILI models, but very few if any applied those models to a large number of substances in a virtual screening exercise, so as to estimate the chances of finding substances devoid of hepatotoxicity. Our virtual screening of over 100,000 substances from the ZINC database found that about 20% of the substances from this data set have a high probability of being devoid of any liver toxicity. Because our models tended to have a relatively high rate of false positives, and the specificity was only about 56%, the proportion of non-hepatotoxic substances is probably much higher in the data set than 20%.

The DILIrank data set is very limited (after excluding non-modelable entities, such as those with ambiguous DILI concern, mixtures or inorganic compounds, only 694 chemical structures), and this leads to small chemical diversity and small learning power for the machine learning algorithms. The very recent publication of the DILIst [37] opens up the possibility of using a larger data set, which is expected to allow better performance of QSAR models built with its help. Our preliminary data show that including the daily dose and duration of therapy may slightly increase the performance of the models, and we intend in the future not only to use the larger DILIst data set, but also attempt the improvement of the models by using such variables as dose and duration of treatment, because it is obvious that they influence the liver toxicity of a product (if paracetamol is used at low doses it is very unlikely to be hepatotoxic, as it is if perhexilin were to be only used for a few days). Albeit using only chemical descriptors for model building in the case of a defined molecular target makes full sense, in the case of a heterogeneous and more or less black box effect such as liver toxicity, it stands to reason that additional considerations related to the context of use (dose, duration) should also be taken into account and explored in future research.

4. Materials and Methods

Dataset

The dataset (Table S1) was downloaded from the FDA website (https://www.fda.gov/science-research/liver-toxicity-knowledge-base-ltkb/drug-induced-liver-injury-rank-dilirank-dataset) and included 1036 drugs, classified in four groups: 192 labeled as "Most-DILI-concern", 287 "Less-DILI-concern", 312 "No-DILI-concern", and 254 "Ambiguous-DILI-concern". Because "ambiguous" is not an actual outcome, but rather a category for which there is uncertainty on their DILI-inducing potential, this last category was excluded from our analysis, leaving the size of the data set to 791. Biological medicines (e.g. abatacept, abciximab etc), drugs that are not definite chemical entities (mixtures – e.g. divalproex sodium, which is a coordination compound of sodium valproate and valproic acid, polymers – e.g. polyethylene glycol 3350), and a small number of simple compounds (e.g. "sterile water", "calcium acetate", "cisplatin") were also eliminated, remaining a final dataset of 694 chemical compounds, of which 179 of most DILI concern, 268 of less DILI concern, and 247 of no concern. We collapsed the most DILI concern and less DILI concern in a single category (of DILI concern), so as to apply binary classification algorithms. The corresponding smiles were included in the data set provided by FDA; we converted them to 2D chemical structures (sdf) using Discovery Studio Visualizer v16.1.0.15350 (Dassault Systèmes BIOVIA, San Diego, CA, USA). We checked the correctness of the resulted formulas with the Chemaxon Structure Checker v. 18.8.0 (ChemAxon, Budapest, Hungary) (no errors found) and used the Chemaxon Standardizer v. 18.8.0 to neutralize, tautomerize, aromatize, and clean 2D the formulas (in this order).

Descriptors

Dragon 7 software (version 7.0, https://chm.kode-solutions.net; Kode SRL, Milano, Italy) was used to compute 3839 molecular descriptors (2D), based on the sdf structures of the chemical compounds in the data set. 19 blocks of molecular descriptors were computed: constitutional descriptors (n=47), ring descriptors (n=32), topological indices (n=75), walk and path counts (n=46), connectivity indices (n=37), information indices (n=50), 2D matrix-based descriptors (n=607), 2D-autocorrelations (n=213), Burden eigenvalues (n=96), P-VSA-like descriptors (n=55), ETA indices (n=23), edge adjacency indices (n=324), functional groups count (153), atom-centred fragments (n=115), atom-type E-state indices (n=172), CATS 2D (n=150), 2D atom pairs (n=1596), molecular properties (n=20), and drug-like indices (n=28).

Feature selection

High dimensions of the data (and "high" may refer to billions of observations, but even 691 x 3891 data points is not exceedingly small), pose a challenge for data analysis, because it is likely that – in our cases - among the several thousand of descriptors not all are correlated with the activity and there is likely much noise or redundancy that should be removed in order to build parsimonious models that are not overfitted and are useful for prediction purposes [54]. Many if not most machine learning algorithms have actually been developed for a fairly low number of variables, and using a very large number of features will likely result in overfitting [55]. It is therefore a requirement to remove noisy and redundant features with the help of one or more feature selection algorithms [56]. A variety of such algorithms have been published, but few comparative performance data with respect to these algorithms are available.

We removed the constant and quasi-constant features (those with less than 1% variation from the statistical mode value) and auto-correlated features (correlation coefficient > 0.9). On the data set thus reduced we have applied 17 distinct feature selection approaches. Some of these are available through the R "mlr" package [57] directly: "anova.test" (based on ANOVA), "auc" (based on the area under the curve), "kruskal.test" (based on a Kruskal-Wallis test rank sum test), and "permutation.importance" (based on the aggregate difference between predictions performed with the unmodified and permuted features), "univariate.model.score" (based on resampling a recursive partitioning learner with each separate feature), and a simple test based on variance. Others are available through the FSelector R package [58]: "FSelector_chi.squared" (uses a chi-squared test of independence between each variable and the outcome), "FSelector_gain.ratio", "FSelector_information.gain", and "FSelector_symmetrical.uncertainty (entropy-based filters), FSelector_oneR (applies the OneR algorithm), "FSelector_relief" (based on the RELIEF algorithm as updated by Kononenko et al.). Other feature selection algorithms included "party_cforest.importance" (permutation importance implemented in the "party" package [59]), "ranger_permutation" (permutation importance implemented in the "ranger" package [60]), "ranger_impurity" (based on ranger impurity importance), and three selection algorithms based on random forests (implemented in the "randomForest" [61] and "randomForestSRC" packages [22]).

Although it is a frequent practice to remove outliers when building QSAR/QSPR models, based on the wrong prediction by the majority of models [17] or other approaches [62], we preferred to eliminate no value, because although this practice might be justified, it might also lead to overestimation of the performance of the models, because the fact that the majority of models are not able to correctly classify an observations does not necessarily imply "outlierness" for that value. We prefer to identify and discuss outliers but to build models being aware of their presence and the limitations they bring upon the models in terms of performance.

Classification algorithms

We have applied the following algorithms to build classification models for the DILIrank dataset: binomial regression; regularized regression; C5.0 decision trees and rule-based models; random forests, regularized random forests, and random forests based on conditional inference trees; rotation forests; extremely randomized trees; Bayesian additive regression trees; support vector machines, clustered support vector machines and divided-conquer support vector machines; Ada boosting; regularized and shrinkage discriminant analysis; neural networks (in three different implementations). All algorithms were applied in the computing and programming environment R, v. 3.6.1 [63], under the unified interface provided by the "mlr" R package [57] coupled with "parallelMap" [64] for parallel computing. For data pre-processing the "caret" package [65] was also used.

Binomial regression (logistic regression), is a fairly simple classification algorithm that models the probability that a certain instance belongs to one of two classes in a linear manner [66]. In essence, logistic regression estimates the probability $P = 1/(1 + e^{-t})$, where $t = a_0 + a_1x_1 + a_2x_2 + ... + a_nx_n$ [67]. *Regularized regression* is a slightly more sophisticated form of conventional regression, where the loss function besides minimizing the sum of squares uses a penalty term, and depending on its value, the regularized regression takes three different shapes (variants): ridge regression, lasso regression, and

elastic net regression, each with its own strengths and weaknesses [68]. We implemented different forms of regularized regression using the "glmnet" R package [69].

C5.0 decision trees and rule-based models have been first advanced by R. Quinlan in 1992, under the name "C4.5", which was an extension of a previous algorithm called Iterative Dichotomizer 3   (ID3); C4.5 was later improved into the new C5.0 classifier, which has superior efficiency [70]. C5.0 decision trees are versatile, swift and easy to use, and their use for QSAR modeling is seen as a reasonable option [71]. In R, the C5.0 algorithm is implemented in the C50 R package [72].

Random forests (RF) are a widespread classification algorithm in a variety of fields, including QSAR [73,74]. They assemble a large number of decision trees with the help of a simple majority vote to resolve the most probable class for each data point. The trees are built using random subsets of both the instances from the training set and of the features in building the individual trees [75]. We applied the algorithm as implemented in the "randomForest" R package [61]. *Regularized random forests* are an adjustment of the conventional random forests geared towards improving the feature selection process by penalizing the introduction of new features in comparison with the previous trees, and thus adding new features only if they provide substantially new information gain/predictive information [76]. It is implemented in R by the original author (H. Deng) in the "RRF" R package [77]. A particular types of trees assembled in random forests are the *conditional inference trees*, that are developed within a so-called "conditional inference framework". Shortly, a global test of independence between response and features is applied, and if the null hypothesis is rejected the variable with the strongest association with the response is selected, a binary split is performed in that variable and the process is repeated recursively [78]. It is implemented in R by its author, T. Hothorn, in the R package "party" [59]. *Rotation forests* are a more sophisticated form of random forests, where feature extraction (e.g. principal component analysis) is applied to subsets of features in an attempt to build "accurate and diverse classifiers" [79]. It is implemented in the R package "rotationForest" [80]. *Extremely randomized trees* (ERT) algorithm is similar to the RF but tends to lessen the variance of the model by the use of a more pronounced randomization component; it differs from RF in two main aspects: each tree is built with all training sample (instead of random subsets) but each node split is chosen randomly in building each tree, instead of using the best split [81].   ERTs are implemented in the R package "extraTrees" [82].

First proposed by H.A. Chipmann et al. in 2010 [83], the *Bayesian additive regression trees* (BART) represent a form offlexible Bayesian non-linear regression which has been shown to have a similar level of performance with other machine learning approaches, for instance with random forests [84]. It is regarded as a successful blend of the advantages of the Bayesian approach with the efficiency of random forests [85]. BART algorithm was used in the R implementation of the package "bartMachine" [86]. The *naïve Bayes* classifier uses the Bayes theorem to compute probabilities and assumes the independence of all variables conditioned on the class; this assumption rarely holds true in the real life (justifying the "naïve" label), but the performance of the algorithm may be surprisingly good in a wide range of classification tasks [87]. In this paper we have used the naïve Bayes classifier only in the building of meta-models, as implemented in the "e1071" R package [88].

The *support vector machines* (SVM) algorithm employs a range of kernel functions (e.g. linear, polynomial, radial etc) to maximize the decision boundary between classes and to define a hyperplane able to best discriminate the classes [89]. It is an algorithm apt of dealing with a large number of features and has been used with good results to solve a diverse range of classification and

regression tasks, including QSAR investigations [90,91]. We used the implementation of the algorithm in the R package "kernlab"[88]. Clustered support vector machines (clusteredSVM) is an algorithm proposed in 2013 by Quanquan Gu and Jiawei Han, as a solution for nonlinear data, with "considerably lower time complexity than nonlinear classifiers"[92]. Briefly, it first partitions the data in clusters (e.g. by k-means), and in each cluster is trained a linear SVM, and a global regularization is applied to prevent over-fitting; the authors reported superiority over linear SVM and similar or superior performance over kernel SVM with better computational efficiency [92]. Although SVM has been described as "probably the most widely used classifier", the kernel SVM has difficulties when the sample size becomes very large because of the huge computational costs [93]. The divide-and-conquer SVM (DC-SVM) approach, proposed in 2014 by C.J. Hsie et al. manages to break through the sample size barrier, ensuring faster computation speeds and prediction accuracy superior to the approximate solvers used by the conventional kernel SVM algorithms [93]. We used both clusterSVM and DC-SVM as implemented in the R package "SwarmSVM" [94].

Boosting is a concept developed gradually in the field of machine learning. After Kearns and Vazirani were the first to ask whether a "weak" classifier (one performing only marginally better than random) may be "boosted" into a successful, "strong" classification model, Robert E. Schapire developed the first such working algorithm in 1989, whereas his colleague Yoav Freund proposed a more efficient one in 1990 [95].   *Adaboost* was put forward by the two researchers in 1995; it iterates the application of a weak (base) algorithm, at each iteration adjusting the weights of the wrongly classified instances, thus "forcing" the algorithm to correctly classify those instances [95]. We applied the algorithm in two implementations, one from the R package "ada"[96], the other from the RWeka package [97].

Regularized discriminant analysis (RDA) is an improvement of the linear discriminant analysis (LDA), the first statistical classifier, which was proposed by R.A. Fisher in 1936 [98]. LDA uses a simple discriminant function to classify instances of the sample and its essence consists in maximizing the between classes sum of squares (SSbetween) and minimizing the within class SS (SSwithin) [99]. RDA ensures better performance (at least in certain cases in which the LDA and quadratic discriminant analysis fail) by using two regularization parameters in the discriminant function [100]. We have use the algorithm as implemented in the R package "klaR" [101].

Artificial neural networks (ANNs) are computational tools used in the prediction of continuous variables or in classification, inspired by the functioning of neurons in the human brain [102]. W. S. McCulloch and W. Pitts are the first who, in 1943, have described an algorithmic neuron, today known after their name (the McCulloch and Pitts neuron). Several years later F. Rosenblatt described another algorithm inspired by the neuron functioning, the perceptron, which automatically "learns" optimal weighting coefficients, which are multiplied by the input variables to decide on the emission (or not) of an output by the neuron [103]. ANNs are particularly apt to model complex and non-linear relationships, as are often those found in chemistry, and thus they are seen as perfectly suitable for QSAR modeling [104]. We have used neural networks in three different R implementations: "neuralnet" [105], "nnet" [106], and "deepnet"[107].

k nearest neighbor (kNN) is one of the simplest and efficient classification algorithms, being based on the idea of assigning an unknown sample to the class to which belong the k most similar compounds (k nearest neighbors); the similarity is assessed through the distance between each data

point of the training sample and the unknown sample [108]. It has often been used successfully in QSAR applications [103, 104]. We have used two R implementations of the algorithm: kknn [111] and the RWeka version (IBk) [97].

Performance evaluation

To assess the performance of the models used in the study we applied nested cross-validation with 5 folds in the inner loop and 10 folds in the outer one; in the case of BART models, which took long times for computing, we used 5 folds for both loops. Cross-validation has a number of strengths that makes it superior to hold-out external validation [112], and nested cross-validation (also known as double cross-validation) is superior to the conventional, simple cross-validation, extending the concept of external validation to the whole data set [113]. Nested cross-validation is therefore considered the state-of-the art approach for the validation of computational models such as QSAR [112], although up to date only a very small number of published studies have used it (e.g. using "nested cross validation QSAR" as keywords in Medline returns only seven papers).

The following metrics were computed and assessed by nested cross-validation: balanced accuracy (BA), mean misclassification error (MMCE), sensitivity (true positive rate, TPR), specificity (true negative rate, TNR), positive predictive value (PPV), and area under the Receiver Operating Characteristics curve (AUC) with their widely known definitions and equations [114,115]. We were interested in predicting whether a particular substance may induce DILI, therefore we were focused on increasing the balanced accuracy, as well as PPV. For this reason, we selected only those models that had both PPV and BA higher than 70%.

Y-randomization test. To examine to what extent models are the result of mere chance we applied a typical y-scrambling test [116] by permuting the toxicity concern label of the drugs from the data set and re-developing the models following the same procedure as the one used for the selected models. We applied the process ten times, each time with different permutations and a different classification algorithm, the models thus build being were evaluated for their performance using the same metrics as for the ones selected. If the models selected for use are not a mere gambling result, the performance of the models built with the permuted data should be (considerably) worse than that of the performance of the selected models.

Virtual screening

In order to explore the way in which the models selected perform on real-world data we assembled the 79 selected QSAR models and applied them to predict the DILI potential of a data set of 104619 compounds of the ZINC database (all compounds having names in the said database [117]). The 79 models were stacked in three different ways:

a) By a majority vote applied to the classification performed by each model;
b) By computing the average of the probabilities outputted by each model and then applying the 50% threshold to classify the compound as being of concern or of no concern (only 72 models outputted probabilities, 6 only made binary predictions); and
c) By developing meta-models using the predictions of the best 50 models (selected with the help of the same selection algorithms as for the building of the individual models) as independent variables for the final classification. We evaluated models based exclusively

on the 50 best-performing individual models, as well as models that additionally included the dose and duration of treatment as supplementary features for the improvement of the performance.

Outliers and applicability domain

For the detection and analysis of outliers in the DILIrank data set we applied the isolation forest (IFOREST) and subspace outlier detection (SOD) algorithms, as implemented in the "solitude"[118] and "HighDimOut" R packages [119]. The applicability domain was assessed by three different methods, as detailed elsewhere [120].

SVM, knn, IBk

Abbreviations

AD	Applicability Domain
ANNs	Artificial Neural Networks
AUC	Area Under the Receiver Operating Characteristics Curve
BA	Balanced Accuracy
BART	Bayesian Additive Regression Trees
CV	Cross-Validation
DC-SVM	Divide-and-Conquer SVM
DILI	Drug induced liver injury
ECHA	European Chemicals Agency
ERT	Extremely Randomized Trees
FPR	False Positive Rate
IBk	an RWeka implementation of the knn algorithm
IFOREST	Isolation Forest
knn	k-nearest neighbours
LDA	Linear Discriminant Analysis
MMCE	Mean MisClassification Error
NIDDK	National Institute of Diabetes and Digestive and Kidney Diseases
OECD	The Organisation for Economic Co-operation and Development
PPV	Positive Predictive Value
QSAR	Quantitative Structure-Activity Relationship
RDA	Regularized Discriminant Analysis
RF	Random Forests
SOD	Subspace Outlier Detection
SVM	Support Vector Machines

TNR	True Negative Rate
TPR	True Positive Rate

References

1. Temple, R.J.; Himmel, M.H. Safety of newly approved drugs: implications for prescribing. *JAMA* **2002**, *287*, 2273–2275.
2. Albrecht, W.; Kappenberg, F.; Brecklinghaus, T.; Stoeber, R.; Marchan, R.; Zhang, M.; Ebbert, K.; Kirschner, H.; Grinberg, M.; Leist, M.; et al. Prediction of human drug-induced liver injury (DILI) in relation to oral doses and blood concentrations. *Archives of Toxicology* **2019**, *93*, 1609–1637.
3. Reuben, A.; Koch, D.G.; Lee, W.M. Drug-induced acute liver failure: Results of a U.S. multicenter, prospective study. *Hepatology* **2010**, *52*, 2065–2076.
4. Ye, H.; Nelson, L.J.; Gómez Del Moral, M.; Martínez-Naves, E.; Cubero, F.J. Dissecting the molecular pathophysiology of drug-induced liver injury. *World J. Gastroenterol.* **2018**, *24*, 1373–1385.
5. Iorga, A.; Dara, L. Cell death in drug-induced liver injury. In *Advances in Pharmacology*; Elsevier, 2019; Vol. 85, pp. 31–74 ISBN 978-0-12-816759-5.
6. Boelsterli, U.A.; Zimmerman, H.J.; Kretz-Rommel, A. Idiosyncratic liver toxicity of nonsteroidal antiinflammatory drugs: molecular mechanisms and pathology. *Crit. Rev. Toxicol.* **1995**, *25*, 207–235.
7. Lewis, J.H. Drug-Induced Liver Injury Throughout the Drug Development Life Cycle: Where We Have Been, Where We are Now, and Where We are Headed. Perspectives of a Clinical Hepatologist. *Pharmaceutical Medicine* **2013**, *27*, 165–191.
8. Godoy, P.; Hewitt, N.J.; Albrecht, U.; Andersen, M.E.; Ansari, N.; Bhattacharya, S.; Bode, J.G.; Bolleyn, J.; Borner, C.; Böttger, J.; et al. Recent advances in 2D and 3D in vitro systems using primary hepatocytes, alternative hepatocyte sources and non-parenchymal liver cells and their use in investigating mechanisms of hepatotoxicity, cell signaling and ADME. *Archives of Toxicology* **2013**, *87*, 1315–1530.
9. Ballet, F. Preventing Drug-Induced Liver Injury: How Useful Are Animal Models? *Dig Dis* **2015**, *33*, 477–485.
10. Negres, S.; Dinu, M.; Ancuceanu, R.; Olaru, T.O.; Ghica, M.V.; Seremet, O.C.; Zbarcea, C.E.; Velescu, B.S.; Stefanescu, E.; Chirita, C. Correlations in silico/in vitro/in vivo regarding determinating acute toxicity in non-clinical experimental trial, according to bioethic regulations inforced by the European Union. *Farmacia* **2015**, *63*, 877–885.
11. Toropova, A.P.; Toropov, A.A. CORAL: Binary classifications (active/inactive) for drug-induced liver injury. *Toxicology Letters* **2017**, *268*, 51–57.
12. Leelananda, S.P.; Lindert, S. Computational methods in drug discovery. *Beilstein Journal of Organic Chemistry* **2016**, *12*, 2694–2718.
13. Chen, M.; Suzuki, A.; Thakkar, S.; Yu, K.; Hu, C.; Tong, W. DILIrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discovery Today* **2016**, *21*, 648–653.

14. Hong, H.; Thakkar, S.; Chen, M.; Tong, W. Development of Decision Forest Models for Prediction of Drug-Induced Liver Injury in Humans Using A Large Set of FDA-approved Drugs. *Scientific Reports* **2017**, *7*, 17311.

15. Liu, L.; Fu, L.; Zhang, J.-W.; Wei, H.; Ye, W.-L.; Deng, Z.-K.; Zhang, L.; Cheng, Y.; Ouyang, D.; Cao, Q.; et al. Three-Level Hepatotoxicity Prediction System Based on Adverse Hepatic Effects. *Molecular Pharmaceutics* **2019**, *16*, 393–408.

16. Wang; Xiao; Chen; Wang In Silico Prediction of Drug-Induced Liver Injury Based on Ensemble Classifier Method. *International Journal of Molecular Sciences* **2019**, *20*, 4106.

17. He, S.; Ye, T.; Wang, R.; Zhang, C.; Zhang, X.; Sun, G.; Sun, X. An In Silico Model for Predicting Drug-Induced Hepatotoxicity. *International Journal of Molecular Sciences* **2019**, *20*, 1897.

18. Roy, K.; Ambure, P. The "double cross-validation" software tool for MLR QSAR model development. *Chemometrics and Intelligent Laboratory Systems* **2016**, *159*, 108–126.

19. Sterling, T.; Irwin, J.J. ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling* **2015**, *55*, 2324–2337.

20. Palomba, D.; Martínez, M.J.; Ponzoni, I.; Díaz, M.F.; Vazquez, G.E.; Soto, A.J. QSPR models for predicting log P(liver) values for volatile organic compounds combining statistical methods and domain knowledge. *Molecules* **2012**, *17*, 14937–14953.

21. Chen, M.; Borlak, J.; Tong, W. High lipophilicity and high daily dose of oral medications are associated with significant risk for drug-induced liver injury. *Hepatology* **2013**, *58*, 388–396.

22. Ishwaran, H.; Kogalur, U.B. *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*; manual, 2019;

23. Ben-Gal, I. Outlier Detection. In *Data Mining and Knowledge Discovery Handbook*; Maimon, O., Rokach, L., Eds.; Springer-Verlag: New York, 2005; pp. 131–146 ISBN 978-0-387-24435-8.

24. Domingues, R.; Filippone, M.; Michiardi, P.; Zouaoui, J. A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition* **2018**, *74*, 406–421.

25. Campos, G.O.; Zimek, A.; Sander, J.; Campello, R.J.G.B.; Micenková, B.; Schubert, E.; Assent, I.; Houle, M.E. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Mining and Knowledge Discovery* **2016**, *30*, 891–927.

26. Berenger, F.; Yamanishi, Y. A Distance-Based Boolean Applicability Domain for Classification of High Throughput Screening Data. *J Chem Inf Model* **2019**, *59*, 463–476.

27. Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17*, 4791–4810.

28. Ramappa, V.; Aithal, G.P. Hepatotoxicity Related to Anti-tuberculosis Drugs: Mechanisms and Management. *Journal of Clinical and Experimental Hepatology* **2013**, *3*, 37–49.

29. Bethesda (MD): National Institute of Diabetes and Digestive and Kidney Diseases LiverTox: Clinical and Research Information on Drug-Induced Liver Injury [Internet]. Ethambutol. 2012.

30. Bethesda (MD): National Institute of Diabetes and Digestive and Kidney Diseases LiverTox: Clinical and Research Information on Drug-Induced Liver Injury [Internet]. Dactinomycin. 2012.

31. Kibleur, Y.; Brochart, H.; Schaaf, H.S.; Diacon, A.H.; Donald, P.R. Dose regimen of para-aminosalicylic acid gastro-resistant formulation (PAS-GR) in multidrug-resistant tuberculosis. *Clin Drug Investig* **2014**, *34*, 269–276.

32. Bethesda (MD): National Institute of Diabetes and Digestive and Kidney Diseases LiverTox: Clinical and Research Information on Drug-Induced Liver Injury [Internet]. Acetaminophen. 2012.

33. Pessayre, D.; Larrey, D. Acute and chronic drug-induced hepatitis. *Baillieres Clin. Gastroenterol.* **1988**, *2*, 385–422.

34. Gunawan, B.K.; Kaplowitz, N. Mechanisms of drug-induced liver disease. *Clin Liver Dis* **2007**, *11*, 459–475, v.

35. Katarey, D.; Verma, S. Drug-induced liver injury. *Clin Med (Lond)* **2016**, *16*, s104–s109.

36. Verma, S.; Kaplowitz, N. Diagnosis, management and prevention of drug-induced liver injury. *Gut* **2009**, *58*, 1555–1564.

37. Thakkar, S.; Li, T.; Liu, Z.; Wu, L.; Roberts, R.; Tong, W. Drug-induced liver injury severity and toxicity (DILIst): binary classification of 1279 drugs by human hepatotoxicity. *Drug Discov. Today* **2019**.

38. Bethesda (MD): National Institute of Diabetes and Digestive and Kidney Diseases LiverTox: Clinical and Research Information on Drug-Induced Liver Injury [Internet]. Leuprolide. 2012.

39. He, S.; Zhang, C.; Zhou, P.; Zhang, X.; Ye, T.; Wang, R.; Sun, G.; Sun, X. Herb-Induced Liver Injury: Phylogenetic Relationship, Structure-Toxicity Relationship, and Herb-Ingredient Network Analysis. *Int J Mol Sci* **2019**, *20*.

40. Prins, B.; Dartee, W.P.; Verboom, W.; Reinhoudt, D.N.; Koster, A.S. Quantitative structure activity relationship for the acute cytotoxicity of 13 (bis)aziridinyl-benzoquinones: relation to cellular ATP depletion. *Arch. Toxicol.* **1994**, *68*, 255–260.

41. Siraki, A.G.; Chevaldina, T.; O'Brien, P.J. Application of quantitative structure-toxicity relationships for acute NSAID cytotoxicity in rat hepatocytes. *Chem. Biol. Interact.* **2005**, *151*, 177–191.

42. Rodgers, A.D.; Zhu, H.; Fourches, D.; Rusyn, I.; Tropsha, A. Modeling liver-related adverse effects of drugs using knearest neighbor quantitative structure-activity relationship method. *Chem. Res. Toxicol.* **2010**, *23*, 724–732.

43. Low, Y.; Uehara, T.; Minowa, Y.; Yamada, H.; Ohno, Y.; Urushidani, T.; Sedykh, A.; Muratov, E.; Kuz'min, V.; Fourches, D.; et al. Predicting drug-induced hepatotoxicity using QSAR and toxicogenomics approaches. *Chem. Res. Toxicol.* **2011**, *24*, 1251–1262.

44. Huang, S.-H.; Tung, C.-W.; Fülöp, F.; Li, J.-H. Developing a QSAR model for hepatotoxicity screening of the active compounds in traditional Chinese medicines. *Food Chem. Toxicol.* **2015**, *78*, 71–77.

45. Toropov, A.A.; Toropova, A.P.; Rasulev, B.F.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J. CORAL: binary classifications (active/inactive) for Liver-Related Adverse Effects of Drugs. *Curr Drug Saf* **2012**, *7*, 257–261.

46. Matthews, E.J.; Ursem, C.J.; Kruhlak, N.L.; Benz, R.D.; Sabaté, D.A.; Yang, C.; Klopman, G.; Contrera, J.F. Identification of structure-activity relationships for adverse effects of pharmaceuticals in humans: Part B. Use of (Q)SAR systems for early detection of drug-induced hepatobiliary and urinary tract toxicities. *Regulatory Toxicology and Pharmacology* **2009**, *54*, 23–42.

47. Zhang, C.; Cheng, F.; Li, W.; Liu, G.; Lee, P.W.; Tang, Y. In silico Prediction of Drug Induced Liver Toxicity Using Substructure Pattern Recognition Method. *Mol Inform* **2016**, *35*, 136–144.

48. Ai, H.; Chen, W.; Zhang, L.; Huang, L.; Yin, Z.; Hu, H.; Zhao, Q.; Zhao, J.; Liu, H. Predicting Drug-Induced Liver Injury Using Ensemble Learning Methods and Molecular Fingerprints. *Toxicol. Sci.* **2018**, *165*, 100–107.

49. Chen, M.; Bisgin, H.; Tong, L.; Hong, H.; Fang, H.; Borlak, J.; Tong, W. Toward predictive models for drug-induced liver injury in humans: are we there yet? *Biomark Med* **2014**, *8*, 201–213.

50. Banerjee, P.; Dehnbostel, F.O.; Preissner, R. Prediction Is a Balancing Act: Importance of Sampling Methods to Balance Sensitivity and Specificity of Predictive Models Based on Imbalanced Chemical Data Sets. *Frontiers in Chemistry* **2018**, *6*.

51. Satalkar, V.; Kulkarni, S.; Joshi, D. QSAR based analysis of fatal drug induced renal toxicity. *Journal of Computational Methods in Molecular Design* **2015**, *5*, 24–32.

52. Kotsampasakou, E.; Ecker, G.F. Predicting Drug-Induced Cholestasis with the Help of Hepatic Transporters-An in Silico Modeling Approach. *J Chem Inf Model* **2017**, *57*, 608–615.

53. Jiang, L.; He, Y.; Zhang, Y. Prediction of hepatotoxicity of traditional Chinese medicine compounds by support vector machine approach. In Proceedings of the 2014 8th International Conference on Systems Biology (ISB); IEEE: Qingdao, China, 2014; pp. 27–30.

54. Cai, J.; Luo, J.; Wang, S.; Yang, S. Feature selection in machine learning: A new perspective. *Neurocomputing* **2018**, *300*, 70–79.

55. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature Selection: A Data Perspective. *ACM Computing Surveys* **2017**, *50*, 1–45.

56. Boopathi, V.; Subramaniyam, S.; Malik, A.; Lee, G.; Manavalan, B.; Yang, D.-C. mACPpred: A Support Vector Machine-Based Meta-Predictor for Identification of Anticancer Peptides. *International Journal of Molecular Sciences* **2019**, *20*, 1964.

57. Bischl, B.; Lang, M.; Kotthoff, L.; Schiffner, J.; Richter, J.; Studerus, E.; Casalicchio, G.; Jones, Z.M. mlr: Machine Learning in R. *Journal of Machine Learning Research* **2016**, *17*, 1–5.

58. Romanski, P.; Kotthoff, L. *FSelector: Selecting Attributes*; 2018;

59. Strobl, C.; Boulesteix, A.-L.; Kneib, T.; Augustin, T.; Zeileis, A. Conditional Variable Importance for Random Forests. *BMC Bioinformatics* **2008**, *9*.

60. Wright, M.N.; Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software* **2017**, *77*, 1–17.

61. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.

62. Veselinović, J.B.; Đorđević, V.; Bogdanović, M.; Morić, I.; Veselinović, A.M. QSAR modeling of dihydrofolate reductase inhibitors as a therapeutic target for multiresistant bacteria. *Structural Chemistry* **2018**, *29*, 541–551.

63. R Core Team *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019;

64. Bischl, B.; Lang, M. *parallelMap: Unified Interface to Parallelization Back-Ends*; 2015;

65. Wing, M.K.C. from J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z.; Kenkel, B.; Team, the R.C.; Benesty, M.; et al. *caret: Classification and Regression Training*; 2019;

66. Dieguez-Santana, K.; Pham-The, H.; Rivera-Borroto, O.M.; Puris, A.; Le-Thi-Thu, H.; Casanola-Martin, G.M. A Two QSAR Way for Antidiabetic Agents Targeting Using α-Amylase and α-Glucosidase Inhibitors: Model Parameters Settings in Artificial Intelligence Techniques. *Letters in Drug Design & Discovery* **2017**, *14*.

67. Raevsky, O.A.; Grigorev, V.Y.; Yarkov, A.V.; Polianczyk, D.E.; Tarasov, V.V.; Bovina, E.V.; Bryzhakina, E.N.; Dearden, J.C.; Avila-Rodriguez, M.; Aliev, G. Classification (Agonist/Antagonist) and Regression "Structure-Activity" Models of Drug Interaction with 5-HT6. *Cent Nerv Syst Agents Med Chem* **2018**.

68. Barrett, T.S.; Lockhart, G. Efficient Exploration of Many Variables and Interactions Using Regularized Regression. *Prevention Science* **2019**, *20*, 575–584.

69. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **2010**, *33*, 1–22.

70. Ahmadi, E.; Weckman, G.R.; Masel, D.T. Decision making model to predict presence of coronary artery disease using neural network and C5.0 decision tree. *Journal of Ambient Intelligence and Humanized Computing* **2018**, *9*, 999–1011.

71. Schöning, V.; Hammann, F. How far have decision tree models come for data mining in drug discovery? *Expert Opinion on Drug Discovery* **2018**, *13*, 1067–1069.

72. Kuhn, M.; Quinlan, R. *C50: C5.0 Decision Trees and Rule-Based Models*; 2018;

73. Hdoufane, I.; Bjij, I.; Soliman, M.; Tadjer, A.; Villemin, D.; Bogdanov, J.; Cherqaoui, D. In Silico SAR Studies of HIV-1 Inhibitors. *Pharmaceuticals (Basel)* **2018**, *11*.

74. Ancuceanu, R.; Dinu, M.; Neaga, I.; Laszlo, F.; Boda, D. Development of QSAR machine learning-based models to forecast the effect of substances on malignant melanoma cells. *Oncology Letters* **2019**.

75. Alfaro, E.; Gámez, M.; García, N. Ensemble Classifiers Methods. In *Ensemble Classification Methods with Applicationsin R*; Alfaro, E., Gámez, M., García, N., Eds.; John Wiley & Sons, Ltd: Chichester, UK, 2018; pp. 31–50 ISBN 978-1-119-42156-6.

76. Deng, H.; Runger, G. Feature selection via regularized trees. In Proceedings of the The 2012 International Joint Conference on Neural Networks (IJCNN); IEEE: Brisbane, Australia, 2012; pp. 1–8.

77. Deng, H. Guided Random Forest in the RRF Package. *arXiv:1306.0237* **2013**.

78. Hothorn, T.; Hornik, K.; Zeileis, A. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* **2006**, *15*, 651–674.

79. Rodriguez, J.J.; Kuncheva, L.I.; Alonso, C.J. Rotation Forest: A New Classifier Ensemble Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2006**, *28*, 1619–1630.

80. Ballings, M.; Poel, D.V. den *rotationForest: Fit and Deploy Rotation Forest Models*; 2017;

81. Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G. AtbPpred: A Robust Sequence-Based Prediction of Anti-Tubercular Peptides Using Extremely Randomized Trees. *Computational and Structural Biotechnology Journal* **2019**, *17*, 972–981.

82. Simm, J.; Abril, I.M. de; Sugiyama, M. *Tree-Based Ensemble Multi-Task Learning Method for Classification and Regression*; The Institute of Electronics, Information and Communication Engineers, 2014; Vol. 97;.

83. Chipman, H.A.; George, E.I.; McCulloch, R.E. BART: Bayesian additive regression trees. *The Annals of Applied Statistics* **2010**, *4*, 266–298.

84. Feng, D.; Svetnik, V.; Liaw, A.; Pratola, M.; Sheridan, R.P. Building Quantitative Structure-Activity Relationship Models Using Bayesian Additive Regression Trees. *J Chem Inf Model* **2019**, *59*, 2642–2655.

85. Hernández, B.; Raftery, A.E.; Pennington, S.R.; Parnell, A.C. Bayesian Additive Regression Trees using Bayesian model averaging. *Statistics and Computing* **2018**, *28*, 869–890.

86. Kapelner, A.; Bleich, J. bartMachine: Machine Learning with Bayesian Additive Regression Trees. *Journal of Statistical Software* **2016**, *70*, 1–40.

87. Saritas, M.M.; Yasar, A. Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. *IJISAE* **2019**, *7*, 88–91.

88. Meyer, D.; Dimitriadou, E.; Hornik, K.; Weingessel, A.; Leisch, F. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*; 2019;

89. Idakwo, G.; Luttrell, J.; Chen, M.; Hong, H.; Zhou, Z.; Gong, P.; Zhang, C. A review on machine learning methods for in silico toxicity prediction. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev* **2018**, *36*, 169–191.

90. Sun, Y.; Shi, S.; Li, Y.; Wang, Q. Development of quantitative structure-activity relationship models to predict potential nephrotoxic ingredients in traditional Chinese medicines. *Food Chem. Toxicol.* **2019**, *128*, 163–170.

91. Chen, H.; Chen, L. Support Vector Machine Classification of Drunk Driving Behaviour. *Int J Environ Res Public Health* **2017**, *14*.

92. Gu, Q.; Han, J. Clustered Support Vector Machines. *Proceedings of the 16 th International Con- ference on Artificial Intelligence and Statistics (AISTATS)* **2013**, *31*, 307–315.

93. Hsieh, C.-J.; Si, S.; Dhillon, I. A Divide-and-Conquer Solver for Kernel Support Vector Machines. *Proceedings of the 31st International Conference on Machine Learning, PMLR* **2014**, *32*, 566–574.

94. He, T.; Demircioglu, A. *SwarmSVM: Ensemble Learning Algorithms Based on Support Vector Machines*; 2019;

95. Freund, Y.; Schapire, R.E. A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence* **1999**, *14*, 771–780.

96. Culp, M.; Johnson, K.; Michailidis, G. *ada: The R Package Ada for Stochastic Boosting*; 2016;

97. Hornik, K.; Buchta, C.; Zeileis, A. Open-Source Machine Learning: R Meets Weka. *Computational Statistics* **2009**, *24*, 225–232.

98. Bruce, P.C.; Bruce, A. *Practical statistics for data scientists: 50 essential concepts*; First edition.; O'Reilly: Sebastopol, CA, 2017; ISBN 978-1-4919-5296-2.

99. Zheng, S.; Ding, C.; Nie, F.; Huang, H. Harmonic Mean Linear Discriminant Analysis. *IEEE Transactions on Knowledge and Data Engineering* **2019**, *31*, 1520–1531.

100. Yang, X.; Elkhalil, K.; Kammoun, A.; Al-Naffouri, T.Y.; Alouini, M.-S. Regularized Discriminant Analysis: A Large Dimensional Study. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT); IEEE: Vail, CO, USA, 2018; pp. 536–540.

101. Weihs, C.; Ligges, U.; Luebke, K.; Raabe, N. klaR Analyzing German Business Cycles. In Proceedings of the Data Analysis and Decision Support; Baier, D., Decker, R., Schmidt-Thieme, L., Eds.; Springer-Verlag: Berlin, 2005; pp. 335–343.

102. Du, K.-L.; Swamy, M.N.S. Fundamentals of Machine Learning. In *Neural Networks and Statistical Learning*; Springer London: London, 2014; pp. 15–65 ISBN 978-1-4471-5570-6.

103. Rashka, S. *Python Machine Learning*; Packt Publishing: Birmingham-Mumbay, 2015;

104. Boudergua, S.; Alloui, M.; Belaidi, S.; Al Mogren, M.M.; Ellatif Ibrahim, U.A.A.; Hochlaf, M. QSAR Modeling and Drug-Likeness Screening for Antioxidant Activity of Benzofuran Derivatives. *Journal of Molecular Structure* **2019**, *1189*, 307–314.

105. Fritsch, S.; Guenther, F.; Wright, M.N. *neuralnet: Training of Neural Networks*; 2019;

106. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*; Fourth.; Springer: New York, 2002;

107. Rong, X. *deepnet: deep learning toolkit in R*; 2014;

108. Deng, Z.; Zhu, X.; Cheng, D.; Zong, M.; Zhang, S. Efficient k NN classification algorithm for big data. *Neurocomputing* **2016**, *195*, 143–148.

109.   Shameera Ahamed, T.K.; Rajan, V.K.; Sabira, K.; Muraleedharan, K. QSAR classification-based virtual screening followed by molecular docking studies for identification of potential inhibitors of 5-lipoxygenase. *Comput Biol Chem* **2018**, *77*, 154–166.

110.   Chitre, T.S.; Asgaonkar, K.D.; Patil, S.M.; Kumar, S.; Khedkar, V.M.; Garud, D.R. QSAR, docking studies of 1,3-thiazinan-3-yl isonicotinamide derivatives for antitubercular activity. *Comput Biol Chem* **2017**, *68*, 211–218.

111.   Schliep, K.; Hechenbichler, K. *kknn: Weighted k-Nearest Neighbors*; 2016;

112.   Baumann, D.; Baumann, K. Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *J Cheminform* **2014**, *6*, 47.

113.   Tetko, I.V.; Sushko, I.; Pandey, A.K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection. *J Chem Inf Model* **2008**, *48*, 1733–1746.

114.   Lagunin, A.A.; Romanova, M.A.; Zadorozhny, A.D.; Kurilenko, N.S.; Shilov, B.V.; Pogodin, P.V.; Ivanov, S.M.; Filimonov, D.A.; Poroikov, V.V. Comparison of Quantitative and Qualitative (Q)SAR Models Created for the Prediction of Ki and IC50 Values of Antitarget Inhibitors. *Front Pharmacol* **2018**, *9*, 1136.

115.   Capuzzi, S.J.; Sun, W.; Muratov, E.N.; Martínez-Romero, C.; He, S.; Zhu, W.; Li, H.; Tawa, G.; Fisher, E.G.; Xu, M.; et al. Computer-Aided Discovery and Characterization of Novel Ebola Virus Inhibitors. *J. Med. Chem.* **2018**, *61*, 3582–3594.

116.   Yang, H.; Du, Z.; Lv, W.-J.; Zhang, X.-Y.; Zhai, H.-L. In silico toxicity evaluation of dioxins using structure–activity relationship (SAR) and two-dimensional quantitative structure–activity relationship (2D-QSAR). *Archives of Toxicology* **2019**.

117.   Sterling, T.; Irwin, J.J. ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling* **2015**, *55*, 2324–2337.

118.   Srikanth, K.S. *solitude: An Implementation of Isolation Forest*; 2019;

119.   Fan, C. *HighDimOut: Outlier Detection Algorithms for High-Dimensional Data*; 2015;

120.   Ancuceanu, R.; Tamba, B.; Stoicescu, C.S.; Dinu, M. Use of QSAR Global Models and Molecular Docking for Developing New Inhibitors of c-src Tyrosine Kinase. *Int J Mol Sci* **2019**, *21*.