

Machine learning for the prediction of drug-induced toxicity

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch- Naturwissenschaftlichen Fakultät

der Universität Basel

von

Verena Schöning

aus Deutschland

Basel, 2019

Originaldokument gespeichert auf dem Dokumentenserver der Universität Basel,
edoc.unibas.ch

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Prof. Dr. Jürgen Drewe, Erstbetreuer
(*Klinische Pharmakologie*)

und

Prof. Dr. Dr. Stephan Krähenbühl, Zweitbetreuer
(*Gruppenleiter Klinische Pharmazie und Head Klinische Pharmakologie & Toxikologie*)

und

Prof. Dr. Gert Fricker, externer Experte
(*Pharmazeutische Technologie und Pharmakologie*)

Basel, den 26.03.2019

Prof. Dr. Martin Spiess
Dekan

In memory of
Christel Schöning

Table of content

TABLE OF CONTENT.....	4	
ACKNOWLEDGEMENTS	9	
ABBREVIATIONS	11	
1	OVERALL INTRODUCTION	13
1.1	Machine learning	14
1.2	Training of a machine learning model.....	18
1.2.1	Data preparation	18
1.2.2	Modelling	24
1.2.3	Validation	27
1.3	Applicability domain	29
1.4	Limitations of machine learning.....	30
1.5	Summary and aim of conducted studies	31
2	PREDICTION OF CLINICALLY RELEVANT DRUG-INDUCED LIVER INJURY FROM STRUCTURE USING MACHINE LEARNING.....	34
2.1	Abstract.....	34
2.2	Introduction	35
2.3	Methods	37
2.3.1	Data collection and preparation.....	37
2.4	Results	41
2.4.1	Dataset	41
2.4.2	Interactions with bio-entities	44
2.4.3	Defined daily doses (DDD).....	45
2.5	Discussion.....	46
2.5.1	Predictive models	46
2.5.2	Interactions with bio-entities	48
2.5.3	Defined daily doses (DDD).....	51
2.6	Conclusions	51
3	IDENTIFICATION OF ANY STRUCTURE-SPECIFIC HEPATOTOXIC POTENTIAL OF DIFFERENT PYRROLIZIDINE ALKALOIDS USING RANDOM FOREST AND ARTIFICIAL NEURAL NETWORK.....	53

3.1	Abstract.....	53
3.2	Introduction	54
3.3	Materials and methods.....	57
3.3.1	Compilation of the PA dataset.....	57
3.3.2	Data pre-processing and feature selection.....	59
3.3.3	Random Forest model (RF).....	63
3.3.4	Artificial Neural Network model (aNN)	63
3.3.5	Prediction model and assessment of outcome	64
3.3.6	Validation of prediction model.....	64
3.4	Results	66
3.4.1	Validation	66
3.4.2	Prediction of the PA dataset	67
3.5	Discussion.....	72
4	PREDICTION OF THE MUTAGENIC POTENTIAL OF DIFFERENT PYRROLIZIDINE ALKALOIDS USING LAZAR, RANDOM FOREST, SUPPORT VECTOR MACHINES, AND DEEP LEARNING.....	77
4.1	Abstract.....	77
4.2	Introduction	78
4.3	Materials and Methods	80
4.3.1	Training dataset	80
4.3.2	Testing dataset	81
4.3.3	LAZAR.....	82
4.3.4	Random Forest, Support Vector Machines, and Deep Learning in R- project.....	85
4.3.5	Deep Learning in TensorFlow	88
4.4	Results	89
4.4.1	LAZAR.....	89
4.4.2	Random Forest, Support Vector Machines, and Deep Learning	92
4.5	Discussion.....	100
4.6	Conclusions	105
5	THE HEPATOTOXIC POTENTIAL OF PROTEIN KINASE INHIBITORS PREDICTED WITH RANDOM FOREST AND ARTIFICIAL NEURAL NETWORKS.....	106
5.1	Abstract.....	106
5.2	Introduction	107

5.3	Materials and methods.....	108
5.3.1	PKI dataset.....	108
5.3.2	DILI dataset and model training.....	109
5.3.3	Model validation.....	110
5.4	Results and discussion.....	111
5.4.1	Model validation and predictor importance.....	111
5.4.2	Overall acute hepatotoxic probability of PKIs	112
5.4.3	Target-specific hepatotoxic probability of PKIs	114
5.4.4	Similarity of PKIs.....	115
5.4.5	Limitation of the study	117
5.5	Conclusion.....	117
6	DEVELOPMENT OF AN <i>IN VITRO</i> SCREENING METHOD OF ACUTE CYTOTOXICITY OF THE PYRROLIZIDINE ALKALOID LASIOCARPINE IN HUMAN AND RODENT HEPATIC CELL LINES BY INCREASING SUSCEPTIBILITY	118
6.1	Abstract.....	118
6.2	Introduction	120
6.3	Materials and Methods	123
6.3.1	Chemical and reagents.....	123
6.3.2	Cells.....	123
6.3.3	Treatment conditions	124
6.3.4	WST-1 assay.....	125
6.4	Results	126
6.4.1	Susceptibility of cells to PAs without pre-treatment.....	126
6.4.2	Enhancement of susceptibility by induction of metabolic activation (rifampicin).....	127
6.4.3	Enhancement of susceptibility by changes in the medium (high-glucose <i>versus</i> galactose).....	128
6.4.4	Enhancement of susceptibility by inhibition of detoxification (carboxylesterases and glutathione formation).....	129
6.5	Discussion.....	133
6.6	Conclusions	136
7	OVERALL DISCUSSION.....	137
8	OVERALL CONCLUSION	139
9	SOFTWARE	141
10	ANNEX.....	142

11	SUPPLEMENTARY MATERIAL	146
12	REFERENCES	147
13	CURRICULUM VITAE	156

Table of figures

Figure 1:	General procedure of generating a predictive model using machine learning methods.....	15
Figure 2:	Under- and overfitting in machine learning.	21
Figure 3:	Principles of the different machine learning techniques.	24
Figure 4:	Confusion matrix and ROC (<i>Receiver Operating Characteristics</i>)-curve.	28
Figure 5:	Decision tree model for hepatotoxic ('DILI') and non-hepatotoxic ('NoDILI') compounds.....	42
Figure 6:	Fraction of drugs interacting with the 15 most common enzymes, carriers, transporters, and targets, grouped by hepatotoxicity.	45
Figure 7:	Distribution of defined daily doses (DDD) is different for hepatotoxic and non-hepatotoxic compounds ($P < 0.001$).....	46
Figure 8:	Common structural features of PAs.....	55
Figure 9:	Flowchart of the creation and validation of the Random Forest and the artificial Neural Network (aNN) models.....	61
Figure 10:	Correlation of the hepatotoxic potential of single PAs as predicted by the RF and the aNN model.....	67
Figure 11:	Cumulative number of PA (in percent) in structural feature groups versus the probability of hepatotoxicity.	69
Figure 12:	Boxplots of the combined PA-structures, the necine base is indicated above the boxplot, the necic acid below.	71
Figure 13:	Flowchart of the generation and validation of the models generated in R-project	88
Figure 14:	Genotoxic potential of the different PA groups as predicted by LAZAR, using the similarity threshold of 0.5	90
Figure 15:	Genotoxic potential of the different PA groups as predicted by LAZAR, using the similarity threshold of 0.2	91
Figure 16:	Genotoxic potential of the different PA groups as predicted by RF model	93
Figure 17:	Genotoxic potential of the different PA groups as predicted by SVM model	94
Figure 18:	Genotoxic potential of the different PA groups as predicted by DL model (R-project)	96
Figure 19:	Six-fold cross-validation of TensorFlow DL model show an average area under the ROC-curve (ROC-AUC; measure of accuracy) of 68%.....	98

Figure 20:	Genotoxic potential of the different PA groups as predicted by DL model (TensorFlow)	99
Figure 21:	Correlation of the hepatotoxic potential of single PKIs as predicted by the RF and the aNN model	113
Figure 22:	Hepatotoxic probability of PKIs in relation to their target.	116
Figure 23:	Metabolic pathways of retronecine-type PAs	121
Figure 24:	Results of WST-I assay in H-4-II-E und HepG2 cells	127
Figure 25:	Results of WST-I assay in H-4-II-E und HepG2 cells	128
Figure 26:	Results of WST-I assay in H-4-II-E cells	129
Figure 27:	Results of WST-I assay in H-4-II-E cells	130
Figure 28:	Results of WST-I assay in HepG2 cells	131
Figure 29:	Results of WST-I assay in H-4-II-E cells	132
Figure 30:	Results of WST-I assay in HepG2 cells	133

Table of tables

Table 1:	Examples for representation of molecular structures by SMILES	18
Table 2:	Strength and weaknesses of different machine learning approaches (modified from (Blower 2006)).....	27
Table 3:	Confusion matrix of the RF model	92
Table 4:	Confusion matrix of the SVM model	94
Table 5:	Confusion matrix of the DL model (R-project).....	95
Table 6:	Confusion matrix of the DL model (TensorFlow).....	97
Table 7	Results of the cross-validation of the four trained models and after y-randomisation	100

Acknowledgements

I would like to express my deepest appreciation and gratefulness to the following people, who contributed in this research:

Prof. Dr. Jürgen Drewe, who believed in me, supported me, spent countless hours to discuss the approaches used and the results obtained, critically reviewed and commented each manuscript to keep the highest standard, and was never shy of new research ideas and projects. You not only help me with and during this “big project” to evolve as scientist, but also as human being! There are no words to express my gratitude, as all attempts seem to belittle my feelings. Thank you so much for everything you have done for me!

Prof. Dr. Dr. Stephan Krählenbühl, who gave me the opportunity to pursue this project and contributed to this research with his ideas for additional studies and his professional insight in the matter!

Dr. Dr. Felix Hammann, who came up with new ways and solutions for problems, let me participate in his work and follow his path, commented and revised manuscripts thoroughly. majQa'!

Kristina Forsch, who has such a nurturing attitude that she needs to take extreme measures to “kill” her cells. Thank you for letting me participate in your lab work, discussing and interpreting these results, and making manuscript writing so much more fun! In addition, for knowing when it was time for retail therapy.

Mark Peinl, who has been a regular in this section during my whole scientific life. Thank you for helping me with all my software, hardware and scripting issues on a 24/7 basis, and for being a great friend.

Last, but not least, from the bottom of my heart, I would like to thank two people, who suffered through endless hours of me explaining my research (with only minor yawning) and never complained too much about the time, I did not spend with them: my wonderful daughter *Lysanne Nierula* and my companion and partner in crime, *Urs Lenz*.

Abbreviations

AD	Applicability domain
aNN	Artificial Neural Networks
ATP	Adenosine triphosphate
AUC	Area Under the Curve
BNPP	Bis(4-nitrophenyl)phosphate
BSEP	Bile salt export pump
BSO	Buthionine sulphoximine
CCR	Correct classification rate
CES	Carboxylesterase
CFSS	Correlation-based feature subset selection
DDD	Defined daily dose
DHP	Dehydropyrrolizidine
DILI	Drug-induced liver injury
DL	Deep Learning
DTI	Decision tree induction
FBS	Foetal bovine serum
FDA	Food and Drug Agency, US Health Authority
GSH	Glutathione
HEPES	2-(4-(2-hydroxyethyl)-1-piperazinyl)-ethansulphoic acid

IQR	Interquartile range
kNN	k-nearest neighbour
LASSO	Least Absolute Shrinkage and Selection Operator
LAZAR	Lazy Structure-Activity Relationships
N/A	Not applicable, missing value
OXPHOS	Mitochondrial oxidative phosphorylation
PA(s)	Pyrrolizidine alkaloid(s)
PCA	Principal component analysis
PK(I)	Protein kinase (inhibitor)
QSAR	Quantitative structure–activity relationship
RF	Random Forest
RFE	Recursive Feature Elimination
ROC	Receiver Operating Characteristic
SMILES	Simplified molecular-input line-entry system
SVM	Support vector machine
TKI	Tyrosine kinase inhibitors
VS	Virtual Screening

1 Overall introduction

The knowledge of toxicological properties of compounds (e.g. drugs, chemicals, and contaminants) is crucial for drug development, definition of toxicological thresholds and exposure limits. However, toxicological testing, either *in vitro* or *in vivo*, is time-consuming, labour intensive and expensive. Furthermore, the 3R-principle¹ aims to reduce and replace animal testing *in vivo*. Therefore, especially for natural occurring contaminants, not all substances are tested, but only a few selected. To complicate the matter even more, this selection is often not based on a toxicological point of view, but on commercial availability. Based on these few results, the notion is often to deduce toxicological limits for the whole substance group from data of few substances, without considering that not all substances from this substance group have the same toxicity.

Another problem, which is more often encountered within the pharmaceutical industry, is that, even if toxicological testing was performed, this information is usually not available in the public domain. Additionally, more often than not, especially when considering more complex endpoints, the results are not comparable to each other due to different experimental set-ups. One way to overcome the two afore mentioned issues, is the use of computational (*in silico*) approaches, such as machine learning. For machine learning, it is assumed that substances with comparable structure or molecular features also exhibit the comparable pharmacological or toxicological action. Based on the comparison of substances with known pharmacological or toxicological action to substances with unknown properties, models, which were generated using machine learning methods, are able to predict the action of the latter substances. The prediction of toxicity by machine learning complements the traditional *in vitro* and *in vivo*

¹ The 3R-principle aims to refine, reduce and replace animal experiments.

experiments in several ways. On one hand, huge datasets may be analysed in a short period and, as the prediction is done by the same model, the results are comparable to each other. This helps to establish a rank order between the different substances and identify substances, which might be interesting to be selected to study *in vitro* or *in vivo*, e.g. the most toxic ones. Furthermore, as many substances are analysed, relationships between e.g. specific structural features and the toxic potential may be established. This might contribute to the elucidation of the mode of action or dependencies.

1.1 Machine learning

Machine learning is a branch of artificial intelligence (AI). A computer learns, using a machine learning method, based on substances with known pharmacological/ toxicological properties (outcome) what features of these substances contribute to the specific outcome. The dataset, which is used for the training of the computer, is often referred to as *training dataset*. The computer is then able to apply the resulting predictive model to a new or unseen dataset, also referred to as *testing dataset*, and predict the outcome of the substances thereof.

To create a predictive model, using machine learning methods, with a good predictivity, several steps need to be undertaken, which are explained shortly in the following section and in, greater detail, in the following chapters. The general steps, which need to be undertaken for the training and validation of a predictive model using machine learning, are shown in Figure 1.

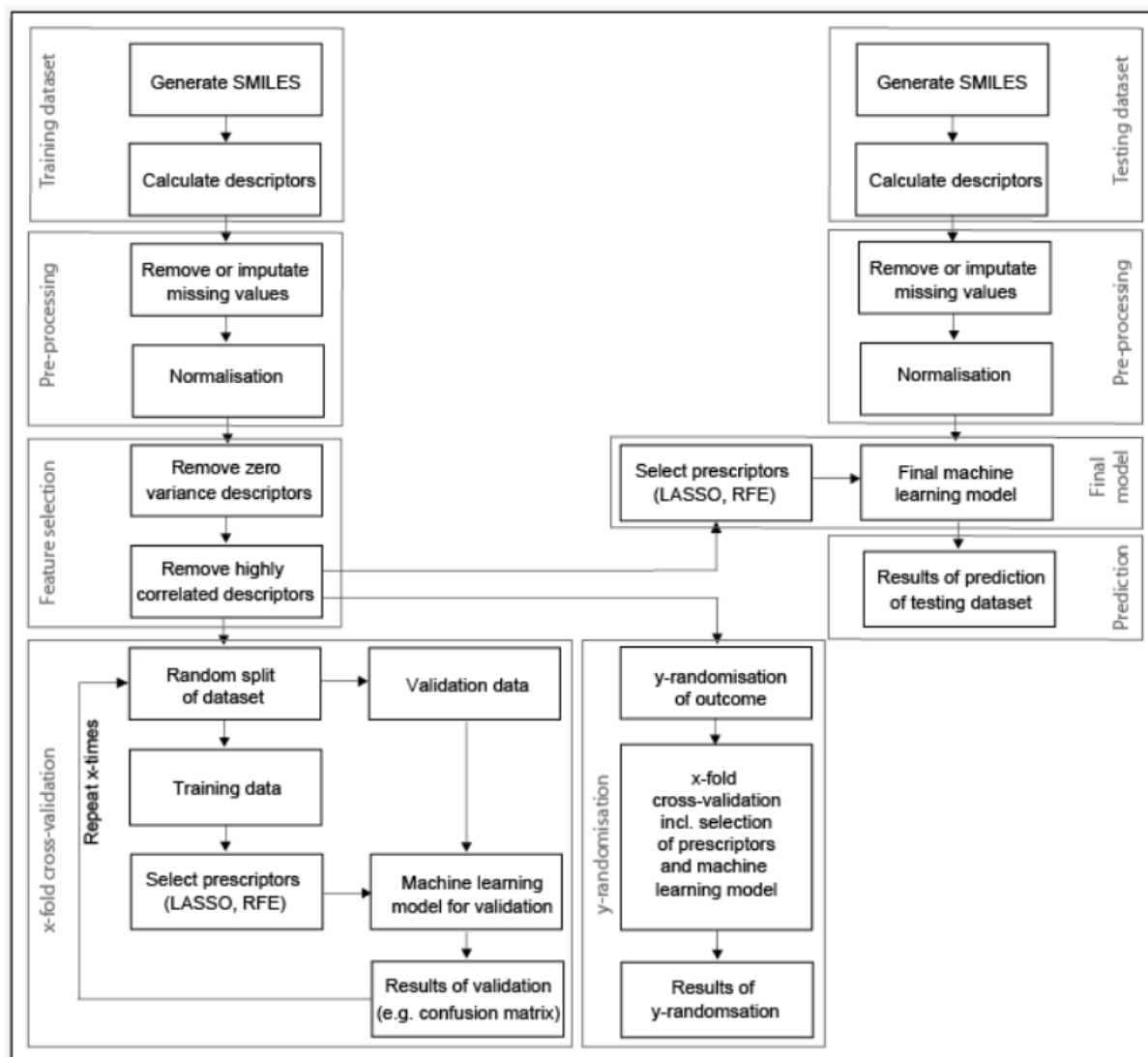


Figure 1: General procedure of generating a predictive model using machine learning methods.

RFE: recursive feature elimination

Before a predictive model can be generated, the substances of interest and their specific structural features need to be converted in a computer-readable form. One common way is the use of unambiguous alphanumeric strings (e.g. SMILES, see chapter 1.2.1.1), which describe the structure of the substance. Based on the structure, numerical values describing the substance, also called descriptors, can be calculated. Each numerical descriptor relates to a different property of a substance, e.g. pH, lipophilicity, amounts of nitrogen atoms, amount of double bonds, or molecular size (see chapter 1.2.1.2).

For the training dataset, not only the descriptors, but also the pharmacological/ toxicological action, the outcome, of the substances is important. The outcome defines the problem that the machine learning model has to solve. When the outcome separates the substances into two or more classes and has therefore a categorical value (e.g. '1' (toxic) and '0' (not toxic)), a *classification problem* needs to be solved. The model will try to sort the substances of the testing dataset into one of the available classes. For *regression problems*, the outcome is a real or continuous variable (e.g. IC₅₀). Based on the data, the models try to predict the actual values for the substances of the testing dataset.

Furthermore, the knowledge of the outcome of the training dataset also determines the general approach to the learning strategy of the model. If the outcome for all substances in the training dataset is known (labelled data), a *supervised learning* strategy can be pursued. The model will try to find a correlation between the outcome and the substances of the dataset. This enables the user to uncover relationships between the outcome and physico-chemical or structural properties of the substances studied and, furthermore, to make predictions about new substances, where the respective outcome is unknown. However, sometimes, the outcome variables of the training dataset are unknown (unlabelled data). In these cases, an *unsupervised learning* strategy can be employed. While unsupervised learning cannot directly be used for a classification or regression problem, as the outcome variables are unknown, it is, however, able to detect the underlying structures or patterns in the dataset. A combination of supervised and unsupervised learning strategy is *semi-supervised learning*. In these cases, the training datasets contains substances with known and substances with unknown outcome (labelled and unlabelled data). This is advantageous in cases, where not enough labelled data are present. The inclusion of unlabelled data increases the size of the training dataset and might help to

better define the border between the different classes of the labelled dataset, as underlying pattern become more pronounced.

Before the actual training of the model, the training dataset need to be pre-processed. As mentioned above, usually, hundreds to a few thousand descriptors (also called features) are calculated per substance. Some of these descriptors are actually related to the outcome and others are not. Descriptors, which are related to the outcome can be used for its prediction, and are therefore often called predictors. The other, unrelated descriptors are noise in the dataset, and need to be identified and eliminated prior to the training of the model. This process is referred to as *feature selection* (see chapter 1.2.1.4).

Another equally important part of data pre-processing apart from the feature selection is data preparation. Missing and incomplete values in the dataset need to be identified, as not all machine learning models are able to deal with missing data. Depending on the size of the training dataset, these values might be replaced by other values, such as the descriptor mean or median. This procedure of replacement is called *imputation*. Otherwise, the whole substance or whole descriptor, which contains missing data, might be deleted. Furthermore, improperly formatted records need to be reformatted.

The selection of the machine learning method, the algorithm, depends largely on the question at hand and the available data. For example, the results of classification algorithms such as *Decision Trees* and *Random Forest* (see chapter 1.2.2.1) could be easily be used to interpret the importance of used variables (descriptors). This might be especially useful if the mechanism is also in the focus of the study. *Deep Learning Networks* (see chapter 1.2.2.2) are best suited for highly complex problems where sufficient amount of data is available. For simpler problems and smaller datasets, Deep Learning Networks tend to adapted too much to the training dataset (*overfitting*) and consequently show poor generalisation on new data. A

detailed description of different machine learning algorithms and their strengths and weaknesses is provided in chapter 1.2.2.

1.2 Training of a machine learning model

The training of a predictive model includes two main steps, data pre-processing, which includes data cleaning and descriptor (feature) selection, and the actual modelling with the corresponding validation. These steps are described in more detail in the following sections.

1.2.1 Data preparation

1.2.1.1 Computational description of molecular structures

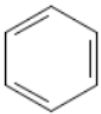

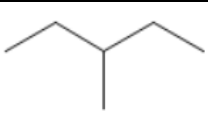
Molecules and chemical structures need to be translated into an alphanumeric string to be interpretable for a computer. One of the most encountered form is the *Simple Molecular Input Line Entry Specification* (SMILES). To simplify the string, hydrogens are usually omitted.

As SMILES strings are unambiguous, they are used as index keys in chemical databases (i.e. PubChem). Furthermore, SMILES can be used in cheminformatics for the calculation of molecular properties (descriptors, see chapter 1.2.1.2). There are two different forms of SMILES, *canonical SMILES*, which **do not** contain stereochemical information, and *isomeric SMILES*, which **do** contain stereochemical information of the molecule.

Some examples for SMILES are provided in the table below (Weininger 1988).

Table 1: Examples for representation of molecular structures by SMILES

Common name	Structure	SMILES
Water	H ₂ O	O
Oxygen	O ₂	O=O
Methane	CH ₄	C

Common name	Structure	SMILES
Ethane	CH ₃ CH ₃	CC
Ethylene	CH ₂ =CH ₂	C=C
Ethanol	CH ₃ CH ₂ OH	CCO
Benzene		C1=CC=CC=C1
Pentane		CCCCC
3-methylpentane		CCC(C)CC

Another, very abstract way of encoding the structure of a molecule, is called molecular fingerprints. Most commonly they are a string of binary numbers (0 and 1) that indicate the presence or the absence of a particular substructure in a molecule (Open Babel community 2011). The similarity of small molecules can be assessed using molecular fingerprints through bit string comparison. It is assumed that structurally similar molecules also exhibit a similar biological activity. Therefore, the comparison of the fingerprint of a target molecule with unknown activity to molecules with known activity can be used to predict the biological activity of the target molecule. This process commonly referred to as virtual screening (VS) (Muegge & Mukherjee 2016).

1.2.1.2 Descriptors

For machine learning studies, chemical substances are characterized in numerical form by different types of descriptors. *Physicochemical descriptors* describe physical and chemical properties of a molecule estimated by examination of its two-dimensional (2D) structure. Examples for physicochemical descriptors are lipophilicity and molecular weight. *Topological descriptors* represent the 2D connectivity of atoms in molecules, whereas *geometrical*

descriptors capture the three-dimensional (3D) information regarding the molecular size, shape, and atoms distribution (Khan & Khan 2016). Those descriptors, which are actually used for the generation of the predictive model, are often referred to as predictors.

1.2.1.3 Data pre-processing

Before the actual training of a potentially predictive model, the training dataset has to be prepared. This pre-processing has a huge influence on the model. First, textual content needs to be converted into a numeric system (e.g. 'toxic' to '1', 'not toxic' to '0'). Missing values ('N/A' values) need to be identified and the approach to handle these instances (imputation, deletion) defined. If the training dataset is large and N/A values are very common in some descriptors or substances, one approach might be to delete these (redundant) descriptors or substances. However, when the training dataset is rather small or N/A values are more or less equally distributed over the whole dataset, deletion would adversely affect the size of the training dataset and therewith the predictive power of the model. In these cases, it might be feasible to impute the missing value e.g. replace these N/A values with the most common value, the mean or median value of the descriptor column, or by applying machine learning methods which can handle N/A values, to calculate the most probable value.

The values of different descriptor have very different ranges, e.g. between '-1' and '1' or between '0' and '10000'. Models might incorrectly overestimate the importance of descriptors with large ranges or numerical values and underestimate descriptors with small ranges or numerical values. Therefore, for some machine learning methods, a normalization of the descriptor column might yield better results for the predictive model. Different approaches are possible. During *range transformation*, the range of all descriptors is harmonised, e.g. all descriptors have only values between '0' and '1'. During *centre transformation*, the mean of the descriptor is subtracted from each descriptor value. The division of each descriptor value

by the standard deviation is called *scaling*. A further procedure is called *rank transformation*, in which the descriptor values of each descriptor are assigned rank numbers, e.g. the smallest descriptors is assigned the rank '1', the second smallest the rank '2', and so on. In each case of normalization, it has to be kept in mind that the same approach needs to be applied to the testing dataset.

1.2.1.4 Feature selection

For most machine learning methods, it is important to reduce the number of descriptors (features) to those who are actually related to the specific outcome and thus contribute to the accuracy of the model. On the one hand, irrelevant descriptors, meaning descriptors, which are not related to the outcome, are noise in the dataset and adversely affect the calculation time and performance of the computer. On the other hand, the major problem is that irrelevant descriptors might generate overly complex models. These overly complex models have often a very poor generalisation performance as the model adapts too much to the noise (unrelated descriptors) in the training dataset. This phenomenon is called *overfitting*, which reduces the predictive power of the model. Overfitting results in an excellent performance on training data but a poor performance on unseen test data (see Figure 2).

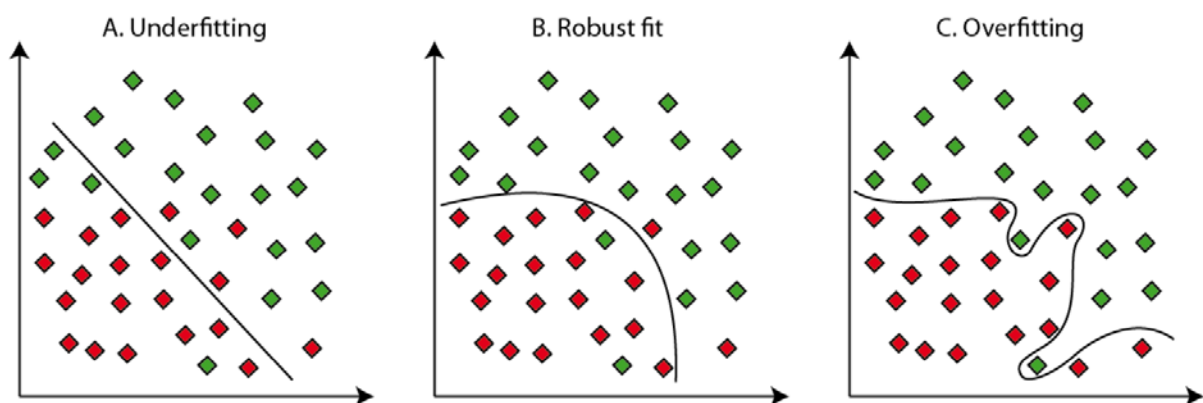


Figure 2: Under- and overfitting in machine learning.
Red and green diamonds symbolize instances of the training dataset belonging to different classes

Therefore, reduction of the descriptors to those actually related to the outcome is one important method to reduce overfitting. The reduction of the number of descriptors is of specific importance in datasets, where the number of substances in the training dataset is lower than the number of descriptors. The procedure of reducing the number of descriptors is called *feature selection*.

As a first step, descriptors, which show low or zero variance, may be deleted. Low or zero variance means that less than 10% of the values of one descriptor are unique or that the frequency of the most common value to the second most common value is more than 95% (e.g. 95 substances with the most common value versus 5 or fewer substances with the second most common value). Furthermore, descriptors that are highly correlated are redundant and may be removed. These two approaches do not take the outcome of the training dataset into account.

Further reduction of the number of descriptors is performed by considering the outcome. Different method can be used for this step. *Filter methods* try to rank the descriptors based on the usefulness to generate the model. These are usually statistical methods such as ANOVA or Chi-square test. *Wrapper methods* train the predictive model on different subset of descriptors and compare model performance. One example for this approach is *recursive feature elimination*. During recursive feature elimination, a predictive model is generated recursively on smaller and smaller subset of descriptors. First, a predictive model is trained with all descriptors, and the least important descriptor is excluded. Then a new predictive model is trained with the reduced subset of descriptors. This procedure is repeated until only a pre-defined number of descriptors are left. *Embedded methods* are a combination of filter and wrapper methods, such as LASSO (*Least Absolute Shrinkage and Selection Operator*): LASSO is a regression method, which performs regularisation and feature selection. The regularisation is done by putting a constraint on the sum of the absolute values of the regression coefficients

so that it becomes less than a fixed value (penalisation). That forces some coefficients to be set to zero. The larger the constraint, the more coefficients are shrunk to zero. For the feature selection process, only the descriptors are used, which have a coefficient of non-zero (Fonti 2017).

1.2.1.5 Balancing of the outcome

A dataset is considered unbalanced if the number of substances with specific outcomes are unequally distributed, e.g. 90% with outcome 'toxic' versus only 10% with outcome 'non-toxic'. Whereas minor unbalanced outcomes may not adversely affect the performance of the model, models based on highly imbalanced datasets tend to favour the majority outcome. This is due to the tendency of the model to reduce prediction error. In the above-mentioned examples, a prediction of all substances as 'toxic' would lead to a prediction error of only 10%. However, the usefulness of the predictions would be questionable, as the model would not be able to identify 'not toxic' compounds.

Different approaches may be used to address this problem, the most common being:

- Oversampling: add copies of substances from the minority outcome to the dataset
- Undersampling: delete substances from the majority outcome
- SMOTE (Synthetic Minority Oversampling Technique): generation of synthetic minority outcome substances based on real minority class (Chawla et al. 2002)
- a combination of over- and undersampling
- Penalisation: increase the influence of the substances from the minority outcome by putting a penalty on the model for wrong prediction of this outcome.

1.2.2 Modelling

Different computational algorithms may be used for the actual generation of a predictive model. A graphic representation of the general principal of each machine learning algorithm is provided in the figure below.

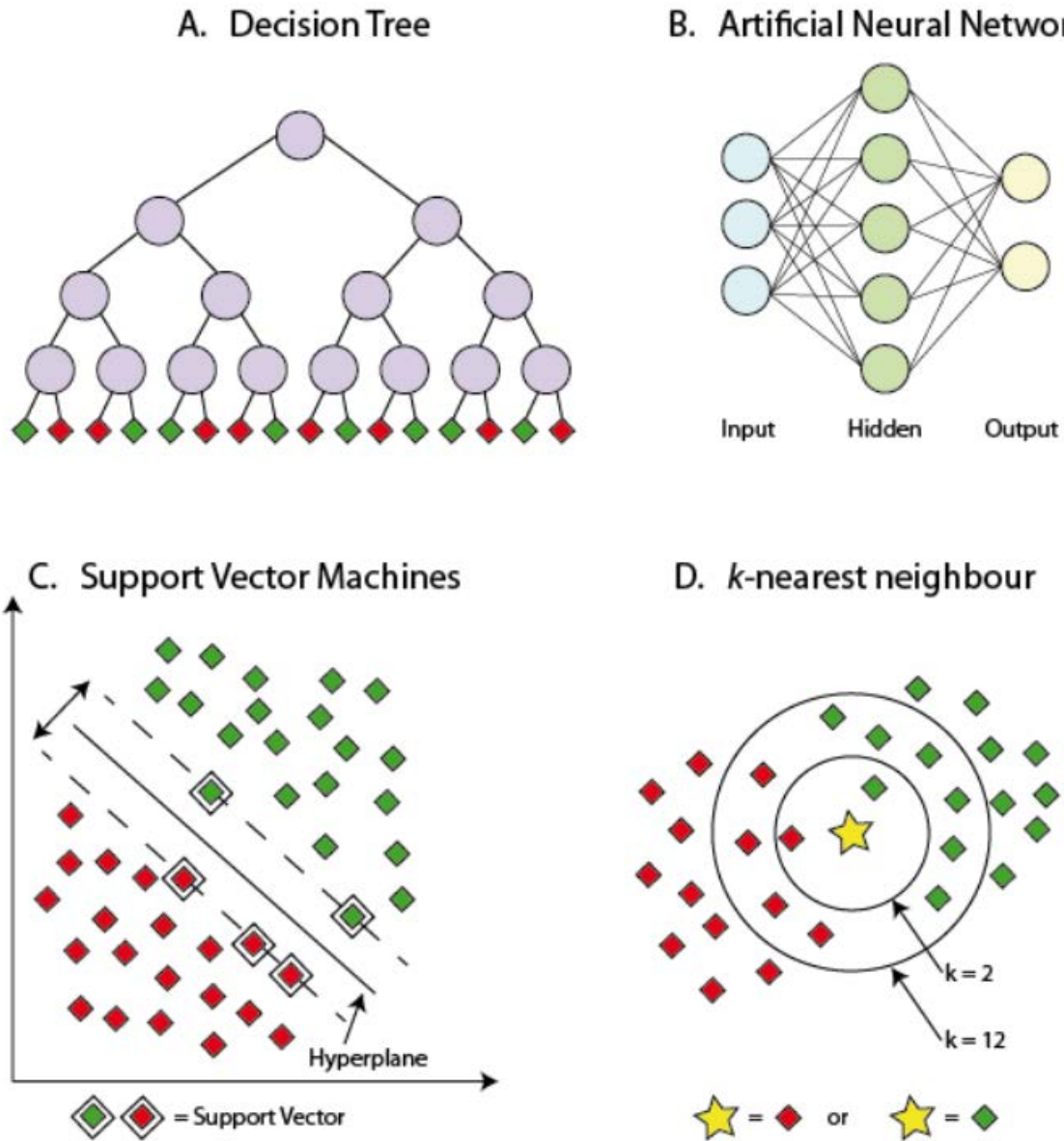


Figure 3: Principles of the different machine learning techniques.
 Red and green diamonds symbolize instances of the training dataset belong to different classes.
 Decision tree: each violet circle stands for a binary decision that has to be made.

1.2.2.1 Decision trees and Random Forest

Decision trees (Figure 3.A) are among the most popular algorithms for machine learning. One main advantage is the easy interpretability of the outcome, as the most important descriptors used for the prediction can be extracted from the model. These descriptors may shed a light on the biological process. For example, Newby et al. (2015) revealed the influence of permeability and solubility on intestinal absorption using decision trees.

However, one huge disadvantage of decision trees is the tendency of overfitting to the training dataset. Besides the reduction of descriptors, the main approach to reduce overfitting in decision trees is called *pruning*, which means the restriction of the model to generate a tree with higher number of branches (*pre-pruning*) or removing parts of an already generated tree (*post-pruning*) (Bramer 2013). Another approach is the generation of a forest of decision trees, where every tree is only trained on a random sample of the training dataset. This approach is called *Random Forest*. The probability for a specific outcome is calculated based on the votes from every single tree.

1.2.2.2 Artificial Neural Networks and Deep Learning

Artificial Neural Networks (aNN) (Figure 3.B) are, as the name suggests, a brain-inspired algorithm, intended to replicate the way humans learn. An aNN consist at least of input and output layers, and in most cases also one or more hidden layer(s). Each unit is called artificial neuron. The input neurons are the input interface for the network and have therefore no predecessor. The output neurons are the output interface of the network and have no successor. According to the input, the artificial neurons (input, hidden, and output) change their internal state (*activation*), and produce output depending on the input and activation method, which is then forwarded to the connected artificial neurons, if present.

A further development in aNN is called *Deep Learning Network (DL)*. They have a greater depth of layers compared to aNN, which is defined by having at least two hidden layer (making a total of at least four layer including the input and output layer). Each layer of the deep learning network trains on a distinct set of features, which is based on the output of the preceding layer. Thus, with each successive layer, the network is able to identify more and more complex features.

1.2.2.3 Support Vector Machines

Support Vector Machines (SVM) (Figure 3.C) try to find a hyperplane (lower dimensional separation²) that best divides the dataset into two classes for classification purpose. The best hyperplane results in the largest separation of the classes, with the largest distance to the data points nearest to the hyperplane. The data points that are nearest to the hyperplane are called support vectors. Removal of these points would alter the position of the dividing hyperplane. Because of this, support vectors are critical elements of the dataset. New substances are classified according to their position in relation to the hyperplane.

1.2.2.4 *k*-nearest neighbour

k-nearest neighbour (kNN) (Figure 3.D) algorithm assigns test substances to the most common class in its neighbourhood, with the neighbours being substances from the training dataset. The variable *k* defines the number of neighbours that shall be taken into account, e.g. if *k*=1, only the single nearest neighbour is considered (and consequentially the test substance is assigned to the same class).

² in a two-dimensional room, the hyperplane is a line, in a three-dimensional room a plane (surface)

In comparison to the other machine learning algorithms, no explicit training of the model is required, as only the ‘neighbourhood’ of the testing substance is considered, but the method does not learn rules, based on which the outcome is predicted.

1.2.2.5 Comparison of the different machine learning models

It is not possible to use the same machine learning method for all problems, as each has different strength and weaknesses. A tabulated comparison of the different models is provided in the table below.

Table 2: Strength and weaknesses of different machine learning approaches (modified from (Blower & Cross 2006))

Characteristic	Decision trees	Artificial Neural Network	Support Vector Machines	<i>k</i> -nearest neighbour
Natural handling of data of mixed type	+	-	-	-
Handling of missing values	+	-	-	+
Robustness to outliers in input space	+	-	-	+
Insensitive to monotone transformations of inputs	+	-	-	-
Computational scalability (large N)	+	+	-	-
Ability to deal with irrelevant inputs	+	-	-	-
Ability to extract linear combinations of features	-	+	+	o
Interpretability	+	-	-	-
Predictive power	-	+	+	+

+ = good o = fair - = poor

Generally, it is a good approach to train at least two predictive models with different machine learning methods for a specific problem and compare the outcome. A comparable prediction from different models increases the confidence in the results.

1.2.3 Validation

Validation of the predictive model is an important step to assess how accurately the model will performance on new/ unseen data from the testing dataset. One approach is called internal

cross-validation. During cross-validation, the training dataset, for which the outcome is known, is randomly separated, usually in a split containing 90% of the substances (training data) and a split containing the remaining 10% of the substances (validation data). The bigger split is used to train the model. This model is then used to predict the outcome of the smaller split. If a feature selection method was applied that considered the outcome, this procedure needs to be included in the validation (that means that the split of the dataset needs to be made before that step). To assess the performance of the model, the actual outcome of the smaller split is compared with the predicted outcome of the model. This procedure is often repeated multiple times with different, random splits. For example, if this procedure is repeated 10-times, a 10-fold cross-validation was performed. For the assessment of the predictive power of the model, the results are displayed in a confusion matrix (see Figure 4.A).

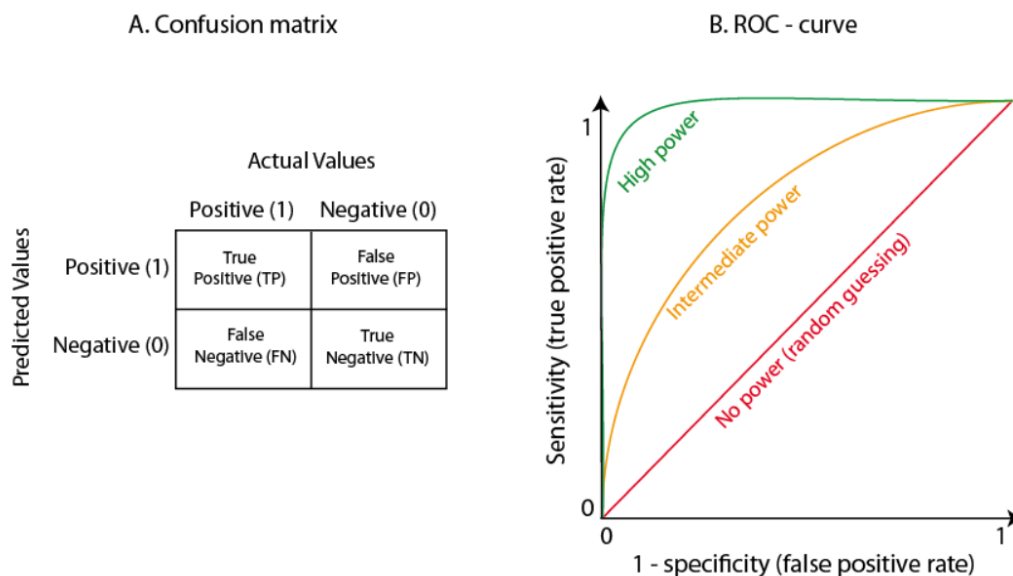


Figure 4: Confusion matrix and ROC (*Receiver Operating Characteristics*)-curve. The area under the ROC-curve (AUC) is a measure of the predictive power. Higher AUC of the ROC-curve indicates higher predictive power

Based on these amounts of *True Positives* (TP), *False Positives* (FP), *False Negatives* (FN) and *True Negatives* (TN), different parameters of the model are calculated, e.g.:

- $Accuracy = \frac{TP+TN}{total}$
- $Sensitivity = \frac{TP}{actual\ positives}$
- $Selectivity = 1 - \frac{FP}{actual\ negatives}$

The predictive power of the model can be graphically displayed as *Receiver Operating Characteristics* (ROC)-curve (see Figure 4.B). In this case, the *Area Under the Curve* (AUC) can also be used to assess the power of the model, the closer the value is to '1' or '0' (for inverse prediction), the better is the power, and therefore the predictivity of the model.

After cross-validation, the final model is generated on basis of the complete training dataset.

Another validation approach is called *y-randomisation*. This validation aims to exclude chance relationships between the outcome and any of the numerous descriptors. For this purpose, the outcome values (sometimes also referred to as *y-variable*) of the training dataset is randomly permuted while the rest of the training dataset is unchanged. Then the whole process of model generation, including feature selection, is performed. The predictive power of this model is assessed e.g. using internal cross-validation. The *y-randomisation* is successful when the accuracy of the randomised model drops to around 50%, which means that the prediction is only by chance. This is because no relationship could be established between the randomised outcomes and the descriptors. From this, it is concluded that a real relationship between the actual outcome and the descriptors is present.

1.3 Applicability domain

The physicochemical, structural or biological descriptors of the training dataset create a highly dimensional virtual space, where each descriptor represents one dimension. This space is called

applicability domain of the predictive model. Within this space, the model is applicable and able to predict the outcome new substances. Substances, which are not within the applicability domain of the model, cannot be predicted correctly. For example, if a model is trained to predict the toxicity of small molecules, it is not able to correctly predict the toxicity of proteins. The model will still predict something, but the prediction is not reliable. Therefore, for each testing dataset, applicability of the model has to be confirmed.

For this purpose, the closeness of the training and the testing dataset can be analysed using statistic approaches, e.g. the *Jaccard distance* or the *Tanimoto coefficient*. The resulting values range from '0' to '1', and indicate the similarity of the datasets. Lower values of the Jaccard distance stand for similarity, whereas higher values indicate diversity. Only if the testing dataset is close to the training dataset, it can be assumed that the former is within the applicability domain of the latter.

1.4 Limitations of machine learning

While machine learning is a useful tool, especially in drug development and toxicology, the limitations need to be kept in mind.

The accuracy of a predictive model is in general below 100%. Usually, model with a correct classification rate of 65% or above in the validation are published in literature (Hammann et al. 2018). However, when using these models on compounds with unknown properties, misclassification is still common and needs to be considered. Therefore, especially in drug development, machine learning is mainly useful of large scaled screening of potential drug substances, discarding those with unwanted properties and identify potential candidates for further development. Still, further *in vitro* and *in vivo* testing of potential candidates cannot be

omitted based on *in silico* results, but selection of promising drug candidates for time- and cost-expensive *in vitro* and *in vivo* testing can be facilitated.

A further issue, which needs to be kept in mind, is the specificity of the model. Models are trained on one or more endpoints and are only able to predict the outcome of these endpoints. The data for the training dataset have to be rather homogenous considering the outcome. It is known that different animals react differently to toxins. If the training dataset consists of toxicological data from different animal species, some outcomes may be equivocal. Furthermore, the machine learning algorithm might encounter problems to clearly separate toxic substances from non-toxic substances. The same is true for toxins, which require metabolic activation and those, which do not. The quality of the training dataset has therefore a significant influence on the performance of the model.

1.5 Summary and aim of conducted studies

The aim of this work was the development of predictive machine learning models for the estimation of risk of hepatotoxicity and genotoxicity. These models were then applied on two different substance groups and the outcome was compared to available literature data.

In the first study (see chapter 2), which was conducted under the lead of F. Hammann, four different machine-learning models, Decision Trees, *k*-nearest neighbour, Support Vector Machines, and artificial Neural Networks, were trained to predict clinically relevant acute hepatotoxicity /drug-induced liver injury (DILI). The training dataset was taken from an expert-committee reviewed DILI dataset. The corrected classification rates of the models were up to 89%. Additionally, the association of drug's interaction with carriers, enzymes, and transporters, and the relationship of defined daily doses (DDD) with hepatotoxicity was investigated. The results presented here are useful as a screening tool both in a clinical setting

in the assessment of DILI as well as in early stages of drug development to rule out potentially hepatotoxic candidates.

Based on these results, it was decided to use the training dataset of this study to assess the acute hepatotoxic potential of over 600 different pyrrolizidine alkaloids (PAs) (see chapter 3). For this purpose, the training dataset was used to train two different models, using the methods Random Forest and artificial Neural Networks. The correct classification rates of these models were 89.0 and 76.2%, respectively. The predicted qualitative hepatotoxicity of both models was highly correlated. Furthermore, specific structural motives showed different hepatotoxic potential. Overall, the obtained results fitted well with already published *in vitro* and *in vivo* data on the acute hepatotoxic properties of PAs.

As the main safety problem with PAs is not the acute hepatotoxicity, but the genotoxic/mutagenic potential, this issue was addressed in a further study (see chapter 4). Different machine learning methods were used to train models for the prediction of the mutagenic potential, LAZAR (*Lazy Structure-Activity Relationships*, which works in principle like *k*-nearest neighbour by direct comparison of the PA structure to other structures with known mutagenic potential), Support Vector Machines, Random Forest and Deep Learning Networks. The PA dataset was partly outside the applicability domain of LAZAR. Training of the other four models, Random Forest, Support Vector Machines, and Deep Learning (using two different approaches), did result in significant predictions, however, the models achieved only low to moderate accuracy rates between 59 and 68%

In a further study, the models for the prediction of acute hepatotoxicity, which were already established during the study concerning the acute hepatotoxicity of PAs, were used on a dataset of 165 protein kinase inhibitors (PKIs) (see chapter 5). The models confirmed clinical observations that PKIs have in general a high probability for inducing hepatotoxicity. However,

interestingly, there seemed to be a target specific difference, with inhibitors of Janus kinases having the lowest hepatotoxic probability of 60-67%.

To confirm the *in silico* results on the hepatotoxic potential of PAs *in vitro*, it was decided to compare the toxicity of commercially available PAs in different hepatic cell lines. Therefore, an *in vitro* screening method to compare the toxic potentials of PAs was developed (see chapter 6). K. Forsch was mainly responsible for the experimental design and conduction of lab work and was supported by the author of this work during the analysis and interpretation of the results and preparation of the manuscript.

2 Prediction of clinically relevant drug-induced liver injury from structure using machine learning

Authors

Felix Hammann, Verena Schöning, and Jürgen Drewe

Published in³:

Journal of Applied Toxicology. 2018; 1–8, ISI Impact factor 2.91

Corresponding author:

Dr. Dr. Felix Hammann

2.1 Abstract

Drug induced liver injury (DILI) is the most common cause of acute liver failure and often responsible for drug withdrawals from market. Clinical manifestations vary, and toxicity may or may not appear dose-dependent.

We present several machine-learning models (decision tree induction, *k*-nearest neighbour, support vector machines, artificial neural networks) for the prediction of clinically relevant DILI based solely on drug structure, with data taken from published DILI cases. Our models achieved corrected classification rates of up to 89%. We also studied the association of a drug's interaction with carriers, enzymes, and transporters, and the relationship of defined daily doses

³ This is a pre-copyedited, author-produced version of an article accepted for publication in Journal of Applied Toxicology following peer review. The version of record 'Hammann F, Schöning V, Drewe J. 2018. Prediction of clinically relevant drug-induced liver injury from structure using machine learning. J Appl Toxicol. 2018 Oct 16 is available online at: <https://doi.org/10.1002/jat.3741>. In course of harmonisations for this manuscript, the numbering and sometimes also the allocations of figures, annexes, and supplementary material was amended. Furthermore, terms were harmonised. No other changes were made.

with hepatotoxicity. The results presented here are useful as a screening tool both in a clinical setting in the assessment of DILI as well as in early stages of drug development to rule out potentially hepatotoxic candidates.

2.2 Introduction

Drug induced liver injury (DILI) is a diagnosis of exclusion for hepatotoxicity causally linked to a xenobiotic (synthetic drugs, herbal preparations, dietary supplements) when all other explanations have been ruled out. It is the most common cause of acute liver failure in developed countries, and a major reason for withdrawal of approved drugs from the US market (Lasser et al. 2002; Reuben et al. 2010). The manifestations range from asymptomatic elevation of liver enzymes to outright acute liver failure. The two main clinical pictures are hepatocellular damage and cholestasis, with many intermediate presentations, as well as changes as liver damage progresses and resolves (Benichou et al. 1993; Danan & Benichou 1993). This heterogeneity is reflected by the various forms of pathophysiological mechanisms implicated, which include disruption of mitochondrial metabolism, changes in transport protein function, immunological processes and hypersensitivity, and direct hepatocellular damage (Kock et al. 2014). Antibiotics are a common source of DILI, with amoxicillin / clavulanic acid posing the greatest risk.

Risk factors are bio-activation by metabolic enzymes (Boelsterli & Lee 2014; Thompson et al. 2016), higher lipophilicity ($\log P \geq 3$), and dose (daily dose ≥ 50 mg) (Chalhoub et al. 2014; Chen et al. 2013; Yu et al. 2014b). Also, DILI has been observed after low-dose medications (Lammert et al. 2008) in patients with a predisposition due to genetic polymorphisms or other ADMET particularities that have gone unrecognized until now, resulting in a false labelling of an adverse event as idiosyncratic.

A well-described risk factor for causing cholestatic injury is the inhibition of the canalicular bile salt export pump (BSEP). Hepatocytes are thought to be flooded with bile salts, eventually leading to apoptosis (Morgan et al. 2010). The basolateral ATP-dependent efflux pumps MRP3 (ABCC3) and MRP4 (ABCC4) can be recruited to shift bile salts into the sinusoidal veins (they are in fact upregulated in cholestasis), and inhibition can among other factors (Aleo et al. 2014; Guo et al. 2015) contribute to cholestatic DILI (Chai et al. 2012; Gradhand et al. 2008). Immunological processes have also been shown to play a role in flucloxacillin cholestatic DILI, wherein hepatic biliary cells are destroyed preferentially in HLA-B*5701-positive patients (Daly et al. 2009).

Previous research, showing higher risk for certain DILIs in specific countries or ethnicities, supports the existence of a genetic component (Ibanez et al. 2002). Also, females appear to be more susceptible to DILI than males (Parkinson et al. 2004).

While risk factors can predispose an individual to develop DILI, these risk factors are often not known, and such cases are then often labelled as 'idiosyncratic'. However, in order to assess DILI clinically, drug-related risk factors also need to be taken into account, e.g. certain structural motifs or other physicochemical properties. To our knowledge, there is still a lack of a predictive model for clinically manifest DILI, a tool, which could be a valuable adjunct in evaluating hepatic dysfunction in a given patient.

A drug's defined daily dose (DDD) is a standardized measure of drug consumption. Interestingly, it appears that high daily doses are predictive of DILI, especially when administered with cytochrome P450 inhibitors (Chen et al. 2013; Yu et al. 2014b). The respective authors believed this to be the result of an increased exposure to mother substances of a drug both through higher dose and decreased detoxification. Another possibility is that

more complex (that is, heavier) drugs have greater hepatotoxic potential. An analysis of molar DDDs could answer this question.

2.3 Methods

2.3.1 Data collection and preparation

2.3.1.1 Data acquisition and structure analysis

The datasets of DILI-positive compounds were taken from different sources, consisting of 311 drugs, that were withdrawn from the market in the USA (Ekins et al. 2010) or European countries due to hepatotoxicity, not marketed there, have received a black box warning because of hepatotoxicity, or are well-known hepatotoxic agents. Other sources were literature-based databases (Ekins et al. 2010; Greene et al. 2010; Stine & Lewis 2011), and 319 drugs from the three Western DILI registries (USA, Sweden and Spain) (Stine & Lewis 2011). We found a total of 627 individual substances in the literature. From these, we removed ambiguous identifiers (for example, ‘oestrogens’).

We also removed proteins and peptides as well as metallic or inorganic compounds (e.g. arsenic trioxide, iron sulphate). This restricted our dataset to one of small molecule substances chemically similar to what is used in most areas of pharmacotherapy today. Furthermore, the structural and physicochemical parameters calculated in this study are largely applicable only to smaller molecules with a unique structure. We used the PubChem Substance and Compounds databases (<http://pubchem.ncbi.nlm.nih.gov/>) to find the associated two-dimensional structures in simplified molecular-input line-entry system (SMILES; isomeric if available, canonical otherwise). Finally, we stripped the molecules of associated salts under the

assumption that they are pharmacologically inert. This process ultimately gave us a list of 588 compounds labelled either 'hepatotoxic' or 'non-hepatotoxic'.

Initial physicochemical calculations were performed with the PaDEL-Descriptor package (version 2.21). We computed the entire range of available 1D and 2D descriptors (n=1381) for all compounds. As some descriptors cannot be calculated for all molecules for technical reasons, this resulted in 526 complete cases (i.e. molecules with complete sets of descriptors). Most incomplete cases were due to only a select few descriptors. We therefore excluded all descriptors that failed in 5% of molecules, which brought the number of complete cases to n=575. The descriptors removed (n=63) included several eigenvalues of the Burden matrix (BCUT) (Burden 1989), simple and valence chi chain descriptors (SCH, VCH), valence and average valence path descriptors (VP, AVP), and a van-der-Waals volume descriptor (VABC).

The remaining incomplete cases were gallium nitrate, trichloroethylene, bromoethanamine, sodium bicarbonate, carbon tetrachloride, chloroform, cadmium chloride, thioacetamide, probucol, dichloroethylene, hydrazine, nitrosamine, and ferrous sulphate. We removed them as they are not representative of small molecular drugs. Afterwards, we removed low-variance descriptors, which were mostly counts of substructural motifs. The final set consisted of 575 compounds and 1'001 descriptors.

For metabolic information, we turned to DrugBank Version 4.3 (<https://www.drugbank.ca>). DrugBank is a freely available resource maintained by the University of Alberta, Canada, which, amongst other things, provides curated information on drug targets and metabolic pathways. We downloaded the entire database and constructed the network of drugs to bio-entities (BE; an umbrella term comprising metabolic enzymes, transporters, carriers, and targets). From this network we removed all substances not in our dataset as well as bio-entities that had no association with the remaining substances (i.e. if an enzyme did not interact with

any of the compounds in our dataset, it was deleted from the network). A total of 417 substances (70.9 %) were listed in DrugBank. Because some of the interactions are asymmetrical (drug to target) and some are not (a drug can be metabolized by and / or induce / inhibit an enzyme's activity) we chose an undirected network architecture. The network is also bipartite since no drugs were assumed to interact directly with each other, and the same assumption was made for BEs. We then constructed unipartite projections so that drugs are removed from the network, and edges (connections) were inserted where two BEs interact with the same drug. For example, the lipid lowering drug simvastatin is a substrate of both cytochrome P450 CYP3A4 and CYP2D6. This would correspond to a connection (edge) between the two isoforms when simvastatin is removed. We performed these steps separately for hepatotoxic and non-hepatotoxic compounds, leaving us with two different networks that can help understand differences in metabolism in DILI and non-DILI situations. The complete architecture is given in Annex 4.

2.3.1.2 Structural similarity

The structural heterogeneity of a collection of molecules can be quantified by considering individual molecules as points in a high-dimensional space wherein each axis corresponds to a descriptor. Similar compounds will then lie closer together, and a set of compounds is considered homogenous if it is tightly packed. The Tanimoto coefficient is a widely adopted method, where the similarity between compounds i and j is calculated from a set of k descriptors as

$$sim(i, j) = \frac{\sum_{d=1}^k X_{di} X_{dj}}{\sum_{d=1}^k (X_{di})^2 + \sum_{d=1}^k (X_{dj})^2 - \sum_{d=1}^k X_{di} X_{dj}}$$

The values of the coefficient range from 0 to 1, with low values indicating diversity and high values similarity.

2.3.1.3 Model learning process

All models discussed here were learned with 10-fold cross-validation to avoid overfitting. Overfitting arises when models with high degrees of complexity and a high accuracy are created that are not generalizable, i.e. perform much worse on unseen data. Additionally, we repeated the cross-validated learning runs ten times with different random seeds to detect any variations in model quality. The final reported models were chosen from these ten runs.

We judged model performance based on their corrected classification rate (CCR), given as

$$CCR = \frac{1}{2} \left(\frac{T_N}{N_0} + \frac{T_P}{N_1} \right)$$

for the two-class case. T_N and T_P represent the number of true negative and positive predictions, respectively, and N_0 and N_1 the total number of negative and positive observations in the model. This measure is more appropriate for skewed datasets such as the one presented here where one class (hepatotoxic compounds) outnumbers the other (non-hepatotoxic compounds).

We surveyed several commonly used machine learning paradigms: decision tree induction (DTI), k-nearest neighbour classification (kNN), support vector machines (SVM), and artificial neural networks (aNN). We implemented these models in GNU R Version 3.3.3.

Decision tree induction is not considered to require feature selection as the number of attributes included in the models is limited by the learning parameters (e.g. maximal tree depth, minimum number of instances per split, minimum number of instances per node). For other paradigms (kNN, SVM, aNN), we performed separate feature selection (dimensionality reduction) with

two commonly used methods: recursive feature elimination (RFE) and correlation-based feature subset selection (CFSS). We provide a full list of the descriptors selected by each method in the supplementary material S5.

As a last step, we repeated the model building processes with y-randomization (Rücker et al. 2007). Here, the observed activities were replaced with random activities with the same proportions of classes as the original data. This is useful to ensure models detect true relationships between attributes and outcomes in situations where the number of attributes and the dimensionality of the paradigms (which can equal infinity in SVM setups) are very large.

2.3.1.4 Defined daily doses (DDD)

The WHO Collaborating Centre for Drug Statistics Methodology maintains a list of drugs and their DDDs (https://www.whocc.no/atc_ddd_index). We manually checked the 588 substances in our original dataset against this database and noted the maximum DDD. No DDD was recorded when the mode of application was topical or local (creams, inhalers, etc.), assuming that no systemic exposure (and, consequently, hepatotoxicity) occurs with their use. We found 245 (41.6 %) drugs for which we recorded the dose in mg/d and the millimolar dose (mmol/d; conversion made with molecular mass as per PaDEL calculations).

2.4 Results

2.4.1 Dataset

The final set for the creation of the machine learning models contains 384 (66.8%) DILI-positive drugs and 191 (33.2%) DILI-negative drugs (total n=575), and is reproduced in the supplementary material S3. The overall Tanimoto similarity index value was fairly low at 0.24, indicating a heterogeneous dataset based on the descriptors employed.

2.4.1.1 Decision Tree Induction

We performed decision tree analysis with an implementation of the CART algorithm in GNU R ('r-part'). The minimum number of cases per split was set to 10, and the minimum number of instances per node was set to 5. Models were learned from the original data with ten-fold cross-validation. The final model performed with a CCR of 0.89 and is reproduced in Figure 5. Y-randomized runs had a maximum CCR of 0.53. There was no increase in performance by balancing datasets during the learning process (maximum CCR 0.88).

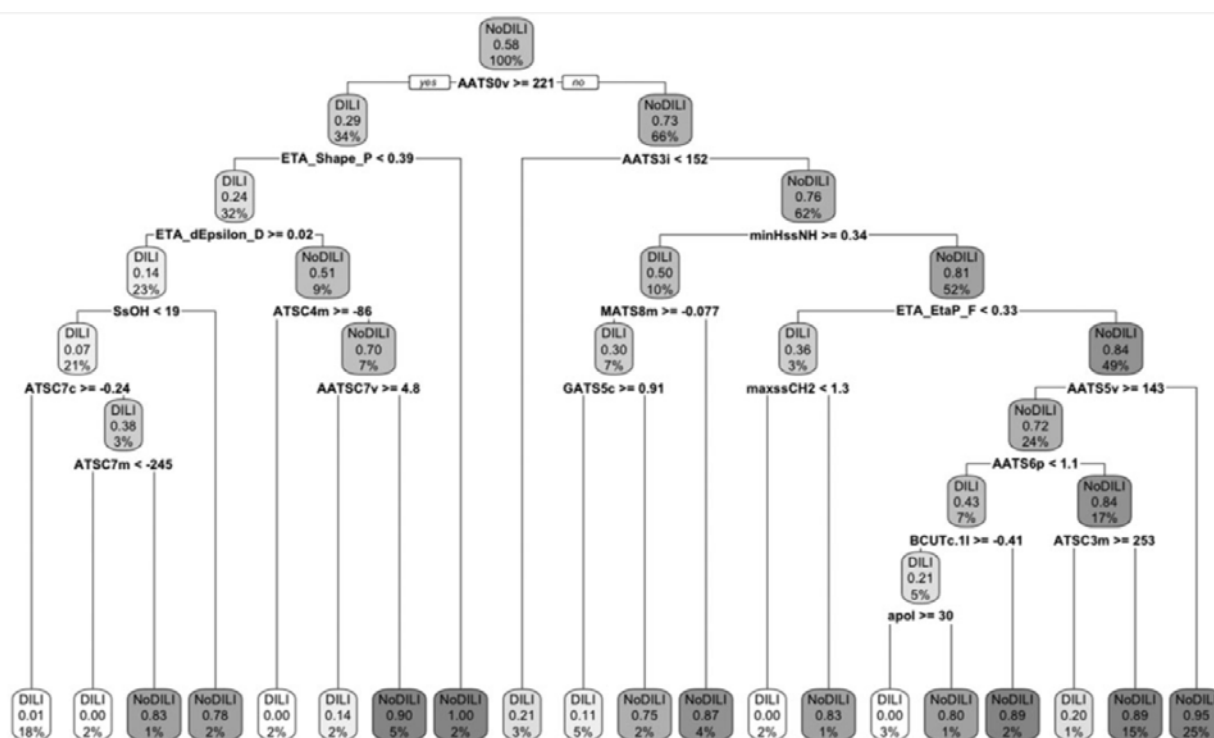


Figure 5: Decision tree model for hepatotoxic ('DILI') and non-hepatotoxic ('NoDILI') compounds.

The more intense the shading, the more of one class is present at each node.

The descriptors selected in this model were mostly topological and include autocorrelation descriptors (AATS2e, AATS2m, AATS4p, AATS5m, AATSC1c, ATS2e, ATSC0e, ATSC3e, ATSC3v, ATSC4e, ATSC4s, ATSC4v, ATSC6v, MATS1e, MATS3c), atom type electrotopological state descriptors (hmin, maxaasC, maxsNH2, SHBa) (Gramatica et al. 2000;

Hall & Kier 1995), structural information content (an index of neighbourhood symmetry of the third order, SIC3) (Basak et al.), the topological distance matrix (SpMin1_Bhs, SpMax1_Bhi), Barysz matrix (VE1_D, VE1_Dzs) (Barysz et al. 1983) and molecular polarizability (Mp). All of these descriptors serve to characterize different molecular shapes, branching, and distributions of charge.

The most readily interpretable attributes were an estimator of logP (ALogP) with a cut-off of -0.72, where higher values are more likely to be predicted as hepatotoxic, and the number of hydrogens (nH). The latter appears very late (i.e. the decision influences few compounds), with > 20 hydrogens being associated with hepatotoxicity.

2.4.1.2 k-nearest Neighbours

We screened several values of k (5 to 20) and found the best performance for k=11. The CCR was 0.73, although little difference was seen between different k values (minimal CCR=0.71). The descriptor set used here (n=27) was the one selected in the decision tree induction model. Other feature subset selection methods (RFE, CFSS) were markedly less successful (maximum CCR=0.65).

We were able to increase the predictive performance on the original dataset to a CCR of 0.83 (maximum CCR in y-randomized runs was 0.56) by using SMOTE balanced internal training sets during cross-validation. Again, k=11 produced the best model.

2.4.1.3 Support Vector Machines

The CCR of the best performing SVM model was 0.74 (CCR = 0.54 in y-randomization) for the decision tree feature subset, while RFE and CFSS subsets were less successful (maximum CCR = 0.66). Using balanced datasets markedly increased the CCR to 0.98, with specificity

and sensitivity at equally high values (0.98). The CCR in the y-randomized runs was 0.89, however, which is why we chose to discard the models learned using balanced training data.

2.4.1.4 Artificial Neural Networks

We trained feed-forward neural networks with a single hidden layer with all feature sets. Best performance was seen with the decision tree feature set (CCR=0.86, CCR in y-randomization = 0.49), while RFE and CFSS both achieved CCRs of 0.74. Balancing the training data did not improve predictivity.

2.4.2 Interactions with bio-entities

Our survey of DrugBank listed interactions with carriers, transporters, and metabolizing enzymes showed (Figure 6) that the largest share of interactions was with CYP3A4, CYP2C9, MRP1, CYP2D6, and CYP2C19. Of statistical significance were CYP2C9, CYP2C8, CYP3A5/7, SLC22A6, ABCC2, serum albumin, and prostaglandin G/H synthase 1. It is of particular interest, that there is not only a statistical difference between individual bio-entities but that the network of interactions (see supplementary material S4) is more complex for hepatotoxic compounds compared to non-toxic compounds.

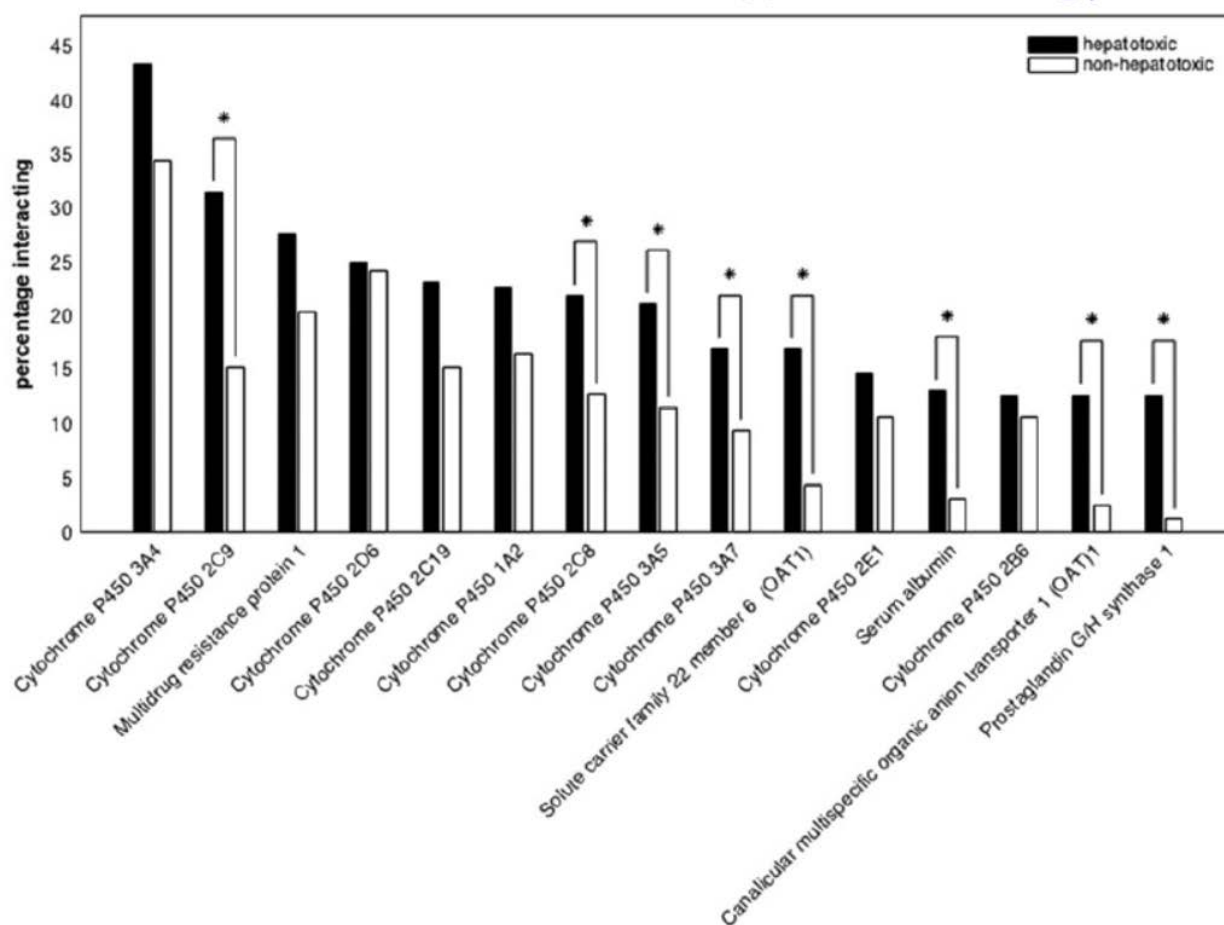


Figure 6: Fraction of drugs interacting with the 15 most common enzymes, carriers, transporters, and targets, grouped by hepatotoxicity. Significance brackets (*) at p -value < 0.05 (Fisher's exact test).

2.4.3 Defined daily doses (DDD)

Based on previous analysis (Chalhoub et al. 2014; Chen et al. 2013; Yu et al. 2014b), the distribution of compounds' DDDs with regard to the threshold of 50 mg between hepatotoxic and non-hepatotoxic groups showed (Figure 7) that for compounds with $DDD < 50$ mg the same number of compounds have been observed to be hepatotoxic and non-hepatotoxic ($n=48$ each). For compounds with $DDD \geq 50$ mg, more than twice as many compounds were hepatotoxic ($n=167$) than non-hepatotoxic ($n=68$; $p < 0.001$). A ROC analysis showed that the criterion of $DDD \geq 50$ mg alone was, however, only a moderate predictor of hepatotoxicity

with a sensitivity of 77.2% and specificity of 22.4%. The same sensitivity was observed for micromolar DDD (threshold 144 μ moles) with a specificity of 40.5% and a significant ($P = 0.001$) distribution of DDD between the treatment groups.

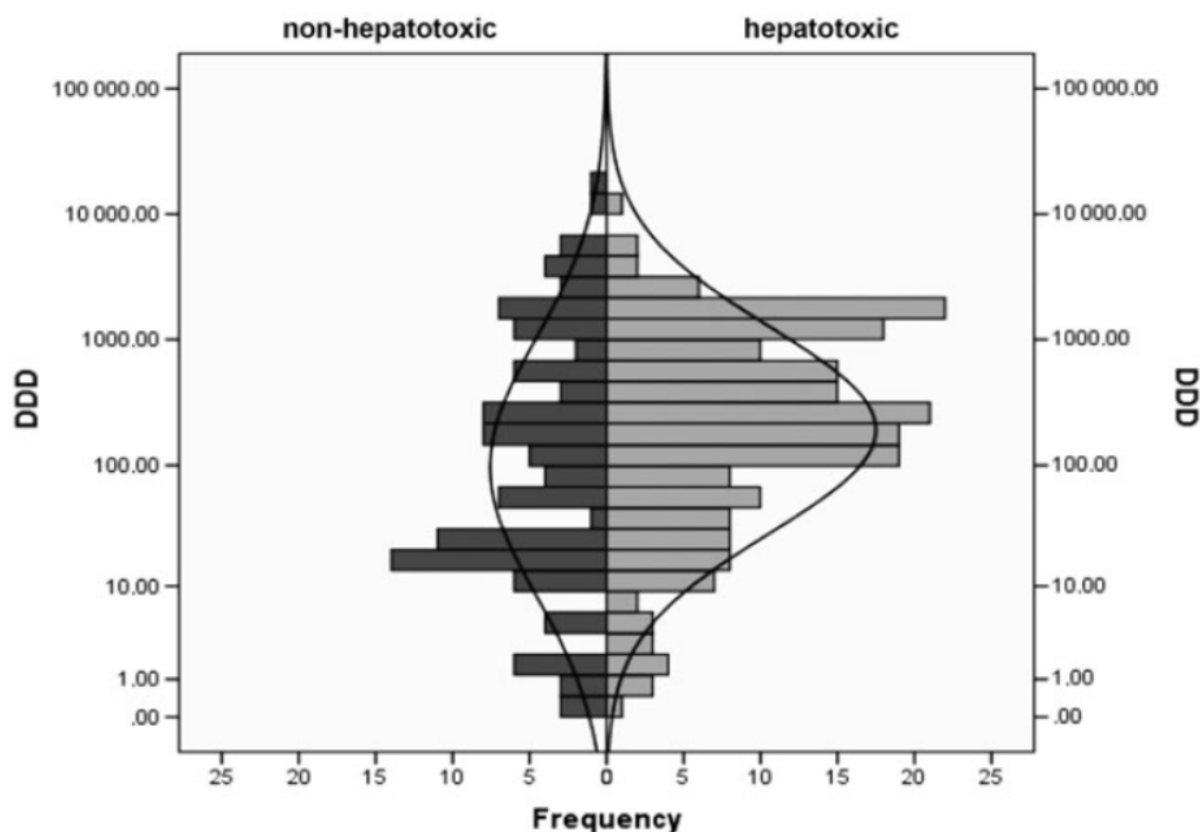


Figure 7: Distribution of defined daily doses (DDD) is different for hepatotoxic and non-hepatotoxic compounds ($P < 0.001$)

2.5 Discussion

2.5.1 Predictive models

Our survey of machine learning paradigms for the prediction of clinically relevant hepatotoxicity points to decision trees as the most useful method. Its precision is matched by artificial neural networks and, if combined with the more involved cross-validation process of balancing the training data, also by the k-nearest neighbour method. Decision tree induction

also provided the most informative feature selection for paradigms that require dimensionality reduction (such as SVM, aNN, or kNN). Because hepatotoxicity can occur for a variety of reasons, the capability of DTI to separate a problem space (DILI) into several subspaces (different pathomechanisms) could be grounds for its effectiveness in this setting.

To put our results into context, we evaluated comparable machine learning approaches to modelling drug-induced liver injury in literature published from 2011 to 2017 (full reference list and overview given as supplementary material S5). Out of the 15 studies, some were performed solely on animal data. Of those based on human data, only five used clinically validated outcomes (such as the FDA's Liver Toxicity Knowledge Base (LTKB)) that can be held to a similarly rigorous standard. Our DTI models are highly predictive, and are on par with – if not superior to – the other published efforts.

Not only can hepatotoxicity arise through a multitude of mechanisms but it can also be precipitated by risk factors. For example, age and gender can greatly affect liver function and toxicity of compounds by changes in cytochrome activity, reduction in hepatic blood flow, decreased drug binding, and malnutrition. (Hunt et al. 1992) Valproate and erythromycin, for instance, show greater hepatotoxicity in children compared to adults. Ethanol consumption increases toxicity of acetaminophen through induction of CYP2E1 (chronic intake) and formation of *N*-acetyl-*p*-benzoquinone imine (NAPQI) as well as reduction of glutathione stores necessary for NAPQI elimination (Stine & Chalasani 2017). Isoniazid toxicity through slow acetylation is genetically predetermined and racial predisposition has been extensively researched (Walker et al. 2009). Risk also increases with concomitant diseases (e.g. pre-existing liver disease, diabetes mellitus, renal failure, HIV/AIDS, obesity) or drug-drug interactions. These individual host factors are not represented in structure-activity relationship models. This is of course complicated by the fact that the same drug can induce multiple forms

of injury, in certain cases at different time points in therapy (early hepatocellular damage progressing to mixed patterns). Therefore, it is possible that even better predictivity could be achieved by integrating individual patient characteristics into future models, e.g. disease state, genotype, or concomitant medications.

2.5.2 Interactions with bio-entities

From our analysis of DrugBank, we found that the network of interactions for hepatotoxic compounds appears more interconnected for hepatotoxic compounds. There were also several statistically significant interaction differences for both classes. It is important to note that DrugBank's data is the result of extensive literature searches and not systematic in vitro testing of the interactions given. As such, there are bound to be biases. For instance, routine regulatory and industry endpoints such as CYP3A4 interactions will be more consistently determined than other endpoints that may stem from academic research.

2.5.2.1 CYP3A subfamily

Members of the CYP3A subfamily are involved in the phase I metabolism of an estimated 50% of small molecule drugs on the market. They catalyse a variety of reactions such as dealkylations, hydroxylations, aromatic oxidations, dehydrogenations, and epoxidations (Rendic 2002). The last process specifically produces exquisitely toxic compounds: the highly reactive electrophilic epoxides (Niederer et al. 2004). Like members of the CYP2C family, CYP3A4 has epoxygenase activity. This has been shown for arachidonic acid and its epoxygeneration to carcinogenous compounds (Bishop-Bailey et al. 2014). There are also specific examples of how CYP3A4 metabolism confers toxicity. The *N*-demethylation of cocaine is greater with CYP3A4 induction and so is cocaine's hepatotoxicity (Pellinen et al. 1994). While there seems to be a trend towards CYP3A4 interactions for hepatotoxic

compounds, there were only statistical significances for CYP3A5 and CYP3A7. All of these enzymes have interindividual variability in activity and, in the case of CYP3A4 and CYP3A7, they show developmental patterns, with CYP3A4 increasing in activity through infancy and CYP3A7 losing activity with age (de Wildt et al. 1999). Possibly, this variability explains why CYP3A4 did not reach significance. Another explanation could be the substrate overlap for CYP2C8 and CYP3A4.

2.5.2.2 CYP2C subfamily

The isoforms 2C8, 2C9, 2C18, and 2C19 share > 80% amino acid identity, although there is only little overlap in substrates. The subfamily accounts for roughly 20% of cytochrome activity in hepatic microsomes. Whereas CYP2C8 metabolizes weakly acidic large molecules, CYP2C9 recognizes weak acids with a hydrogen acceptor, and CYP2C19 basic molecules or amides with two hydrogen acceptors (Zanger et al. 2008). Many have narrow therapeutic indices (coumarines, phenytoin). The CYPs 2C8, 2C9, 2C18, and 2C19 have epoxygenase activity and can generate superoxide (O_2^-), cytotoxic reactive oxygen species (Fleming 2014; Miners & Birkett 1998).

2.5.2.3 CYP2D6

The CYP2D6 enzyme is polymorphic, i.e. there are several interindividual and ethnic differences in activity, and individuals can be grouped into different phenotypes (ultra-rapid, extensive, intermediate, and poor metabolizers). There are inhibitors, for example fluoxetine and quinidine, but no known inducers of CYP2D6 (Teh & Bertilsson 2012). This variation has drug safety consequences, as many antipsychotics and antidepressants, but also oncologicals like tamoxifen, are metabolized at least partly by CYP2D6. As a consequence, there may be symptoms of overdose where the dosage is not matched with individual metabolic capacity, or

where there are toxic metabolites. This has been documented *in vitro* and *in vivo* for psychiatric drugs (e.g. quetiapine, venlafaxine, and trazodone) (Jornil et al. 2013; Li & Cameron 2012; Najibi et al. 2016). Other examples include primaquine (Ganesan et al. 2009) and acetaminophen (Dong et al. 2000). It is therefore not surprising that CYP2D6 interactions are a risk factor for drug toxicity.

2.5.2.4 Serum albumin

Binding to albumin, alpha-1 acid glycoprotein, and lipoproteins can play an important part in drug distribution. Hypoproteinaemia will lead to higher free (= unbound) concentrations for these drugs, and, by consequence, stronger effects and possibly toxicity. Clinically, this is well recognized for antiepileptics such as valproic acid (Ahmed & Siddiqi 2006), anti-inflammatory drugs such as salicylates (Gitlin 1980), or anti-infectives (Makhlouf et al. 2008). Many disease states can influence protein binding, ranging from malnutrition and malignancies to hepatotoxicity itself, as hepatic insufficiency involves altered protein synthesis.

2.5.2.5 Cellular Transporters

The basolaterally expressed MRP1 (ABCC1) and the apically expressed MRP2 (ABCC2) are members of the ATP-binding cassette family, found throughout the body (also hepatically), and involved in the transport of a wide range of compounds, both charged and uncharged. Their clinical importance can be seen in the response to and toxicity of methotrexate depending on MRP1 activity (Lima et al. 2015), or statin disposition in relation to MRP1 and MRP2 activity (Rodrigues 2010).

2.5.3 Defined daily doses (DDD)

We confirmed previous observations (Chalhoub et al. 2014; Chen et al. 2013; Yu et al. 2014b) that DDD is a predictor of hepatotoxicity. The trend towards better specificity of high DDDs in micromolar units compared to mg is indication that the sheer number of circulating molecules is more important than the dose amount in the system. However, DDDs alone were both only moderate predictors of hepatotoxicity with sensitivity 72.2% and low specificity.

2.6 Conclusions

We present a study of the structural and metabolic features associated with hepatotoxicity. There are only few instances where drug induced liver injury is not considered idiosyncratic. Our study indicates that this is not the case, and that the vast majority of hepatotoxicity seems to be predictable from a drug's structure – a potentially very useful tool in clinical pharmacological practice as well for avoiding costly attritions in drug development.

Despite the predictive power of our models, they could be markedly improved by incorporating these susceptibility factors into more comprehensive systems. These could help in individual therapy (personalized medicine) and in regulatory questions, e.g. for judging the toxic potential in special populations such as children, the elderly, or pregnant / lactating women. Similarly, given a sufficiently large dataset, subgroup analyses by injury pattern (hepatocellular *vs.* cholestatic *vs.* mixed) would be informative.

We also show that different metabolic pathways are active in hepatotoxicity, and these may be influenced by predisposing factors (age, gender, or ethnicity), concomitant medication, or disease states. The major limitation here is that the drug interactions evaluated are likely heavily biased by regulatory requirements. Furthermore, we were able to confirm that higher

DDDs, esp. in micromolar units, are a risk factor for the development of acute hepatotoxic effects.

3 Identification of any structure-specific hepatotoxic potential of different pyrrolizidine alkaloids using Random Forest and artificial Neural Network

Authors

Verena Schöning, Felix Hammann, Mark Peinl, Jürgen Drewe

Published in⁴:

Toxicological Sciences, 160(2), 2017, 361–370, ISI Impact factor 4.081

(The paper was accompanied with an editorial: Toxicological Sciences, 160(2), 2017, 191–192, <https://doi.org/10.1093/toxsci/kfx242>)

Corresponding author:

Prof. Dr. Jürgen Drewe, MSc

3.1 Abstract

Pyrrolizidine alkaloids (PAs) are characteristic metabolites of some plant families and form a powerful defence mechanism against herbivores. More than 600 different PAs are known. PAs are ester alkaloids composed of a necine base and a necic acid, which can be used to divide PAs in different structural subcategories. The main target organs for PA metabolism and

⁴ This is a pre-copyedited, author-produced version of an article accepted for publication in Toxicological Science following peer review. The version of record 'Schöning V, Hammann F, Peinl M, Drewe J. 2017. Editor's Highlight: Identification of Any Structure-Specific Hepatotoxic Potential of Different Pyrrolizidine Alkaloids Using Random Forests and Artificial Neural Networks. Toxicol Sci 160:361-70' is available online at: <https://doi.org/10.1093/toxsci/kfx187>. In course of harmonisations for this manuscript, the numbering and sometimes also the allocations of figures, annexes, and supplementary material was amended. Furthermore, terms were harmonised. No other changes were made.

toxicity are liver and lungs. Additionally, PAs are potentially genotoxic, carcinogenic and exhibit developmental toxicity. Only for very few PAs, *in vitro* and *in vivo* investigations have characterised their toxic potential. However, these investigations suggest that structural differences have an influence on the toxicity of single PAs. To investigate this structural relationship for a large number of PAs, a quantitative structural-activity relationship (QSAR) analysis for hepatotoxicity of over 600 different PAs was performed, using Random Forest- and artificial Neural Networks-algorithms. These models were trained with a recently established dataset specific for acute hepatotoxicity in humans. Using this dataset, a set of molecular predictors was identified to predict the hepatotoxic potential of each compound in validated QSAR models. Based on these models, the hepatotoxic potential of the 602 PAs was predicted and the following hepatotoxic rank order in three main categories defined: (i) for necine base: otonecine > retronecine > platynecine; (ii) for necine base modification: dehydropyrrolizidine >> tertiary PA = *N*-oxide and (iii) for necic acid: macrocyclic diester \geq open-ring diester > monoester. A further analysis with combined structural features revealed that necic acid has a higher influence on the acute hepatotoxicity than the necine base.

3.2 Introduction

Pyrrolizidine alkaloids (PAs) are characteristic metabolites of some plant families, with more than 95% of the PA-containing species belonging to the following four families: *Asteraceae*, *Boraginaceae*, *Fabaceae* and *Orchidaceae* (Hartmann & Witte 1995; Langel et al. 2011). More than 600 natural occurring PAs have been identified from approximately 6000 angiosperm species (Chen et al. 2010). They form a powerful defence mechanism against herbivores (insects, mammals).

PAs are heterocyclic ester alkaloids composed of a necine base (two fused five-membered rings joined by a single nitrogen atom) and a necic acid (one or two carboxylic ester arms), occurring principally in two forms, tertiary base PAs and PA *N*-oxides.

The necine base may have different structures, which divide PAs into several types, e.g. otonecine, platynecine, and retronecine. Furthermore, a classification based on the necic acid is possible (Langel et al. 2011). A coarse classification of the necic acid would be macrocyclic diester, open-ring diester and monoester (see Figure 8).

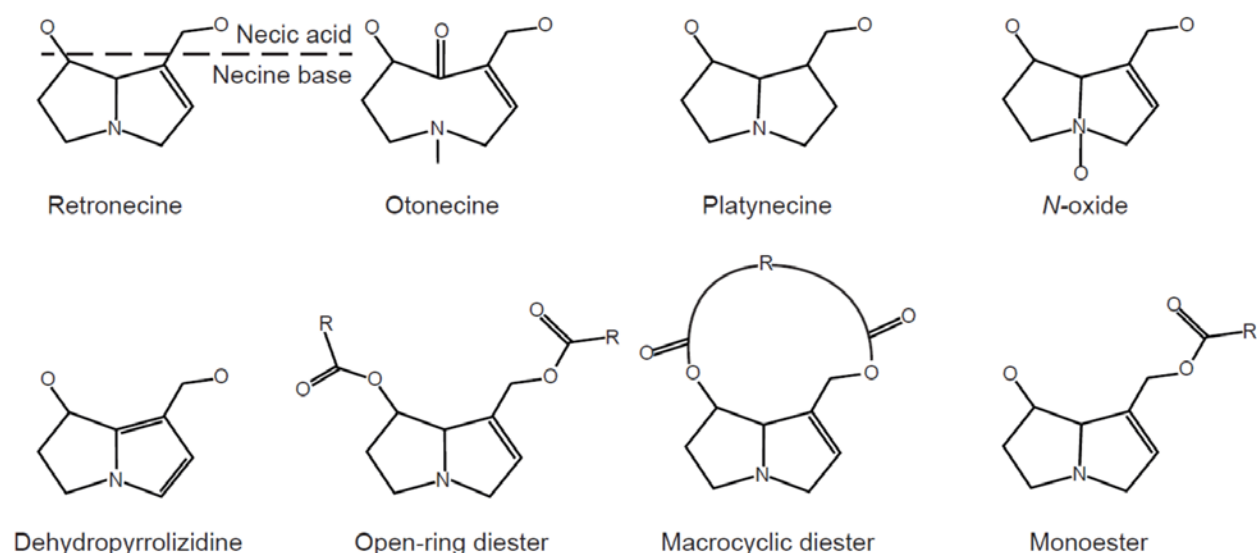


Figure 8: Common structural features of PAs.

Plants synthesise and translocate PAs as hydrophilic *N*-oxides, but may be store as either lipophilic tertiary base or hydrophilic *N*-oxide (Hartmann et al. 1989). Upon ingestion of plants by herbivores, the *N*-oxides are reduced in the gut to its tertiary alkaloids-form and then passively absorbed (Lindigkeit et al. 1997). PA metabolism occurs mainly in the liver, which is also the main target organ of toxicity (Bull & Dick 1959; Bull et al. 1958; Butler et al. 1970; DeLeve et al. 1996; Jago 1971; Li et al. 2011; Neumann et al. 2015). There are three principal metabolic pathways for 1,2-unsaturated PAs (Chen et al. 2010): (i) Detoxification by

hydrolysis of the ester bond on positions C7 and C9 by non-specific esterases to release necine base and necic acid, which are then subjected to further phase II-conjugation and excretion. (ii) Detoxification by *N*-oxidation of the necine base (only possible for retronecine-type PAs) to form PA *N*-oxides, which can be conjugated by phase II enzymes e.g. glutathione and then excreted. PA *N*-oxides may be converted back into the corresponding parent PA (Wang et al. 2005). (iii) Metabolic activation or toxification of PAs by oxidation (for retronecine-type PAs) or oxidative *N*-demethylation (for otonecine-type PAs (Lin 1998)). This pathway, which is mainly catalysed by cytochrome P450 isoforms CYP2B and 3A (Ruan et al. 2014b), results in the formation of dehydropyrrolizidine (DHP, also known as pyrrolic ester or reactive pyrroles). DHPs cause damage in the cells where they are formed, usually hepatocytes, but can pass from the hepatocytes into the adjacent sinusoids and damage the endothelial lining cells (Gao et al. 2015) predominantly by reaction with protein, lipids and DNA. There is even evidence, that conjugation of DHP to glutathione, which would generally be considered a detoxification step, could result in reactive metabolites, which might also lead to DNA adduct formation (Xia et al. 2015). Due to the ability to form DNA adducts, DNA crosslinks and DNA breaks 1,2-unsaturated PAs are generally considered genotoxic and carcinogenic (Chen et al. 2010; EFSA 2011; Fu et al. 2004; Li et al. 2011; Takanashi et al. 1980; Yan et al. 2008; Zhao et al. 2012). However, there is no evidence yet that PAs are carcinogenic in humans (ANZFA 2001; EMA 2016). After acute intoxication of humans, the most common lesions in the liver are haemorrhagic necrosis, lesions in the central and sublobular veins of the liver, and acute venoocclusive disease (DeLeve et al. 2003; EFSA 2011).

There is evidence that the oral bioavailability (Hessel et al. 2014) and the specific toxicity of single PAs depends on structural features of the necic acid and the necine base. Considering the necine base, only 1,2-unsaturated PAs (retronecine- and otonecine-type PAs) can be

metabolically activated in the liver to DHPs. Saturated PAs (platynecine-type PAs) are also metabolised by cytochromes, but the metabolites are water-soluble and readily excreted (Ruan et al. 2014a; Ruan et al. 2014b). No formation of DNA-adducts could be shown for saturated PAs (Xia et al. 2013). Therefore, saturated PAs may be regarded as less/non-toxic. Also, differences in the toxicity of 1,2-unsaturated PAs were observed, with otonecine-type PAs being more toxic than retronecine-type PAs (Li et al. 2013). Furthermore, from experimental experience, PAs with macrocyclic diesters are considered more toxic than those with an opening diester or monoester (EFSA 2011; Fu et al. 2004; Ruan et al. 2014b).

However, a drawback of these *in vitro* and *in vivo* studies is – due to limited availability of pure substances - the limited number of PAs investigated with regards to their structure-specific toxicity. To overcome this bottleneck, the structure-specific hepatotoxic potential of over 600 different PAs was predicted using two QSAR models, implementing either Random Forest (RF) or an artificial Neural Network (aNN), and which were trained specifically for acute human drug-induced liver injuries (DILI).

3.3 Materials and methods

3.3.1 Compilation of the PA dataset

The PA dataset was created from five independent, necine base substructure searches in PubChem (Supplementary material S1). The resulting standard data files (sdf-files) were scanned with Bioclipse (v2.6) (Spjuth et al. 2009; Spjuth et al. 2007). The downloaded structures were compared to the PAs listed in the EFSA publication (EFSA 2011) and the book by Mattocks (Mattocks 1986), using the CAS-number and the synonyms, to ensure, that all major PAs were included. PAs mentioned in these publications which were not found in the downloaded substances were searched individually in PubChem and, if available, downloaded

separately. Non-PA substances, duplicates, and isomers were removed from the files by hand. Artificial PAs, even if unlikely to occur in nature, were included in the analysis. As result, the final PA dataset comprised a total of 602 different PAs. For each PA molecular 1D and 2D descriptors were calculated using PaDEL-Descriptors (version 2.21) (Yap 2011; 2014). The process of standardization involved removing any salts from SMILES structures, for instance chlorides or lysinate residues. Additionally, we removed explicit hydrogens.

The PAs in the dataset were classified according to structural features. A total of 9 different structural features were assigned to the necine base, modifications of the necine base and to the necic acid (see Figure 8).

For the necine base, the following structural features were chosen:

- Retronecine-type (1,2-unsaturated necine base)
- Otonecine-type (1,2-unsaturated necine base)
- Platynecine-type (1,2-saturated necine base)

For the modifications of the necine base, the following structural features were chosen:

- *N*-oxide-type
- Tertiary-type (PAs which were neither from the *N*-oxide- nor DHP-type)
- DHP-type (pyrrolic ester)

For the necic acid, the following structural features were chosen:

- Monoester-type
- Open-ring diester-type

- Macrocyclic diester-type

Then, to assess the combined influence of the necine base and the necic acid on hepatotoxicity, the above-mentioned features were combined. This resulted in the following 15 groups:

- Retronecine with a monoester (80 compounds), open-ring diester (80 compounds), or macrocyclic diester (139 compounds)
- Retronecine *N*-oxide with a monoester (25 compounds), open-ring diester (24 compounds), or macrocyclic diester (21 compounds)
- Otonecine with a monoester (1 compounds), open-ring diester (1 compounds), or macrocyclic diester (41 compounds)
- Platynecine with a monoester (45 compounds), open-ring diester (43 compounds), or macrocyclic diester (38 compounds)
- Platynecine *N*-oxide with a monoester (3 compounds), open-ring diester (6 compounds), or macrocyclic diester (2 compounds)

Otonecine *N*-oxides do not exist, since the carboxyl-group at the nitrogen prevents *N*-oxidation.

3.3.2 Data pre-processing and feature selection

A flowchart of the development of the prediction models, including validations, is provided in Figure 9.

The DILI dataset, which was used to train the QSAR-models, was established by Chen et al. (2016) and was built up from different sources: marketed drugs approved by the FDA, which are a) withdrawn or labelled in boxed warning or warnings and precautions with severe DILI

indication (most-DILI-concern), b) DILI labelling in warnings and precautions with mild DILI indication or adverse reactions (less-DILI-concern), and c) no DILI indicated in the labelling (no-DILI-concern). Verification of DILI-concerns was made with reference to public resources (i.e. the NIH LiverTox database), and cases from major DILI registries (Spanish DILI Registry, Swedish Adverse Drug Reactions Advisory Committee Database, and the Drug-Induced Liver Injury Network (DILIN) in the USA).

Substances which were validated classified as being of less-DILI-concern and of most-DILI-concern were regarded as hepatotoxic, whereas substances classified as no-DILI-concern were regarded as non-hepatotoxic. Substances with ambiguous-DILI-concern and antibodies were removed from the dataset. The final dataset consisted of 721 substances, containing 453 hepatotoxic and 268 non-hepatotoxic substances. For each substance 1444 molecular descriptors were calculated using PaDEL-descriptors (version 2.21) (Yap 2011; 2014), analogously to the PA dataset.

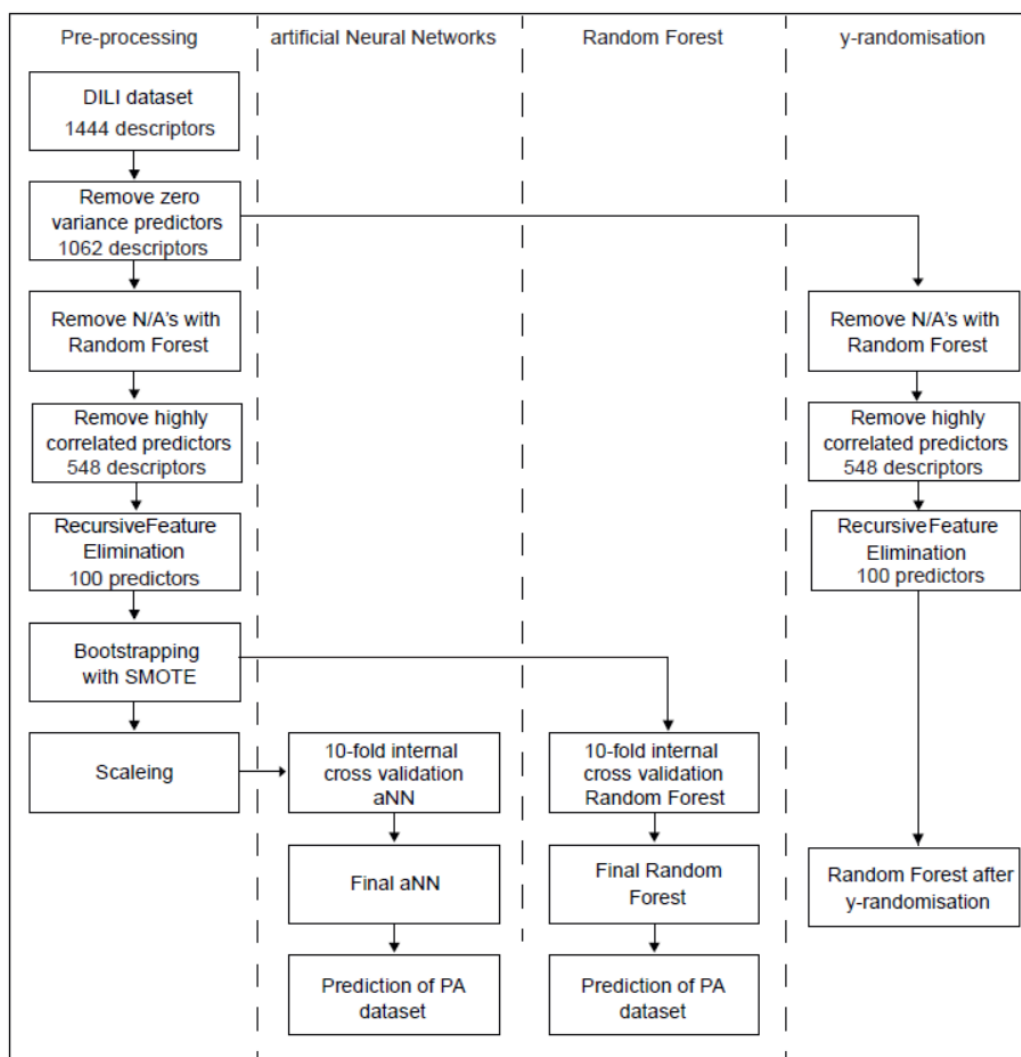


Figure 9: Flowchart of the creation and validation of the Random Forest and the artificial Neural Network (aNN) models.

In the course of data cleaning for import, two substance had to be removed from the dataset, as many descriptors could not be computed. Furthermore, values in the dataset, which were smaller than 1×10^{-10} were set to zero. Then the dataset was imported into R (R Project for Statistical Computing, <https://www.r-project.org/>; version 3.3.1) and all further steps were performed using additional R packages (packages are identified for each step in the description below).

The second step after data cleaning was variable selection to identify the descriptors, which are actually related to the outcome. First of all, descriptor variables with a near zero variance were identified and removed using the 'NearZeroVar'-function (package 'caret'). A descriptor was classified as near zero variance if the percentage of unique values was less than 10% or when the ratio of the frequency of the most common value to the frequency of the second most common value was greater than 95:5 (e.g. 95 instances of the most common value and only 5 or less instances of the second most common value). A total of 1062 descriptors were left after this step. The DILI dataset contained 2.38% of missing values. These missing values were imputed using the 'rfimpute'-function (package 'randomForest'). The use of imputation was driven by the need for complete cases in learning RF models. As the training dataset is by its very nature homogeneous (mostly small molecule drugs), imputation of missing values is justifiable. Furthermore, it was not necessary to impute any descriptors for the prediction of PA dataset.

Then, highly correlated descriptors were removed using the 'findCorrelation'-function (package 'caret') with a cut-off of 0.9 yielding 548 descriptors. A Recursive Feature Elimination (RFE) method with Random Forest (Zhu et al. 2015) was then used to identify the most important descriptors (the final predictors) to describe the outcome. For this model it was aimed to use approximately 100 predictors to avoid overfitting. Therefore, different numbers of predictors (1, 10, 50, 75, 100, 200, 548) were tested and the accuracy of the predicted outcome was compared. As optimal accuracy was achieved with 100 descriptors, these descriptors were chosen as predictor for modelling.

Unbalanced datasets can adversely affect the training of the QSAR model. A dataset is considered unbalanced if certain classes are overrepresented. Different approaches are possible, e.g. artificially balancing the dataset, assigning penalties to the model for

misprediction of the minority class, or giving the minority class a higher weight. In this study, it was decided to use the ‘**Synthetic Minority Oversampling Technique**’ (Chawla et al. 2002), function ‘ubSMOTE’ (package ‘unbalanced’) to balance the dataset. To verify the suitability of the SMOTE-function, a total of 50 balanced dataset were created and the performance compared in a cross-validation approach. The mean accuracy of the 50 forests was 89% (range: 83-94%), indicating that the creation of artificial instances with the SMOTE algorithm does not introduce systematic bias. The final balanced DILI dataset consisted of 458 hepatotoxic and 455 non-hepatotoxic observations.

3.3.3 Random Forest model (RF)

Based on the 100 most important predictors and the balanced DILI dataset, a RF model (Breimann 2001) was trained using the ‘randomForest’-function (package ‘randomForest’). A forest with 1000 decision trees was grown, where 75 variables were randomly sampled as candidates at each split.

3.3.4 Artificial Neural Network model (aNN)

For the aNN model, an additional pre-processing step was necessary. The DILI dataset was normalised by calculating the standard deviation for each predictor and then divide each value by that standard deviation (‘preProcess’-function, package ‘caret’). The same scaling used for the DILI dataset was applied to the PA dataset.

The aNN model consisted of a multilayer perceptron which was created by using the ‘mlp’-function (package ‘RSNNS’) (Bergmeir & Benítez 2012). It consisted of three layers, an input layer with 100 units, a hidden layer 75 units, and an output layer with one unit. A logistic activation function was used.

3.3.5 Prediction model and assessment of outcome

The RF and the aNN models were used to predict the probability of hepatotoxicity of the PA dataset. Therefore, the models indicated the probability for each substance to be a hepatotoxin. A higher percentage probability value does not mean that the substance is more toxic than a substance with a low value, but rather indicates that the chances are higher for these substances to be actually hepatotoxic (Breimann 2003).

The probability results were binned into probability classes in increments of 10% (e.g. 70-80% probability for hepatotoxicity) and these probability classes were compared to the structural features assigned to the PAs. Statistical significance was tested using an unpaired student's t-test ('t.test'-function, package 'stats').

3.3.6 Validation of prediction model

The following methods were used for the validation of the prediction model in this study (Mitchell 2014; Nantasenamat et al. 2009):

Confirmation of applicability domain

The suitability of a prediction model for a specific dataset depends on the applicability domain of the training and the test dataset. This means that the range of the predictor values of the training dataset have to match with the test dataset. A test compound is unlikely to be correctly predicted if there is no similar compound in the training set. To confirm the applicability domain of the DILI and the PA dataset, a principal component analysis (PCA) was performed, using the identified, relevant 100 predictors and the first four principal components. Furthermore, the distance between the DILI dataset and the PA dataset was calculated using the Jaccard distance measure.

Cross-validation

Due to the relatively small number of observations in the DILI dataset, no external cross-validation was performed. It was assumed, that a 10-15% reduction of the training dataset might adversely affect the applicability domain of the total model. Instead, a 10-fold, internal cross-validation was conducted.

The accuracy of predictions is given as the ratio of hits to total number of compounds. This measure may grossly overestimate the actual quality in skewed datasets, i.e. where the members of one class greatly outnumber those of other ones. Here, we report the predictive power of each model as *correct classification rate* (CCR):

$$CCR = \frac{1}{2} \left(\frac{T_N}{N_0} + \frac{T_P}{N_1} \right)$$

where T_N and T_P represent the number of *true negative* and *positive predictions*, respectively, and N_0 and N_1 the total number of negative and positive compounds in the model. Also, the sensitivity and specificity of the models were calculated.

Y-randomisation

To exclude chance correlation of the descriptors and the outcome a y-randomisation (Rücker et al. 2007) was performed. The real model is compared with an alternative model, where the outcome (y-variable) is randomly permuted and the model, including feature selection, is built on basis of these randomised outcomes.

This validation was only performed using a RF model. As the permuted outcome variables were already balanced, the bootstrapping step of the data pre-processing was omitted. Also, no 10-fold cross-validation was performed. The quality of the permuted model was only evaluated

based on the *ROC* (Receiver Operating Characteristics)-curve, the corresponding AUC (Area Under the Curve) and the *confusion matrix*.

3.4 Results

3.4.1 Validation

The compliance of the applicability domains of the DILI and the PA dataset was tested using a PCA. The PCA, considering the first four principal components (PC1 to PC4), showed that in principal, the PA dataset was within the range of the DILI dataset (see Annex 1). The former result was also confirmed by the calculation of the Jaccard distance, which showed an average distance below 0.2 for all PAs relative to the training dataset. Therefore, it can be assumed that the DILI dataset is suitable to build predictive models for the PA dataset.

A 10-fold internal cross-validation was conducted to test the performance of the models. The RF model had a CCR of 89.0%, a sensitivity of 88.8%, a specificity of 89.3%, and a ROC-AUC of 0.96. The performance of the aNN model was slightly inferior, with a CCR of 76.2%, a sensitivity of 77.5%, a specificity of 74.9%, and a ROC-AUC of 0.84.

After y-randomisation of the outcome, the RF model had only a CCR of 52.2%, a sensitivity of 46.0%, a specificity of 58.5%, and a ROC-AUC of 0.53. These results indicate that the predictions were by chance, and no correlation between predictors and outcome can be established. Therefore, the predictors of the DILI dataset were actually related to the outcome and a by chance correlation can be excluded.

The results of the four validation approaches show that prediction models based on the DILI dataset are valid and suitable to predict the acute hepatotoxic potential of the PA dataset.

3.4.2 Prediction of the PA dataset

From the 602 PA analysed, a total of 105 and 496 PAs were predicted as hepatotoxic (probability of at least 50%) by the RF and the aNN model, respectively.

The prediction of single PAs was highly correlated between both models ($R=0.977$, $p<0.0001$, see Figure 10). RF generally predicted a lower probability of hepatotoxicity than aNN. However, this analysis showed that the aNN prediction were on average higher than the predictions with the RF model (intercept -12.7%, slope 0.80).

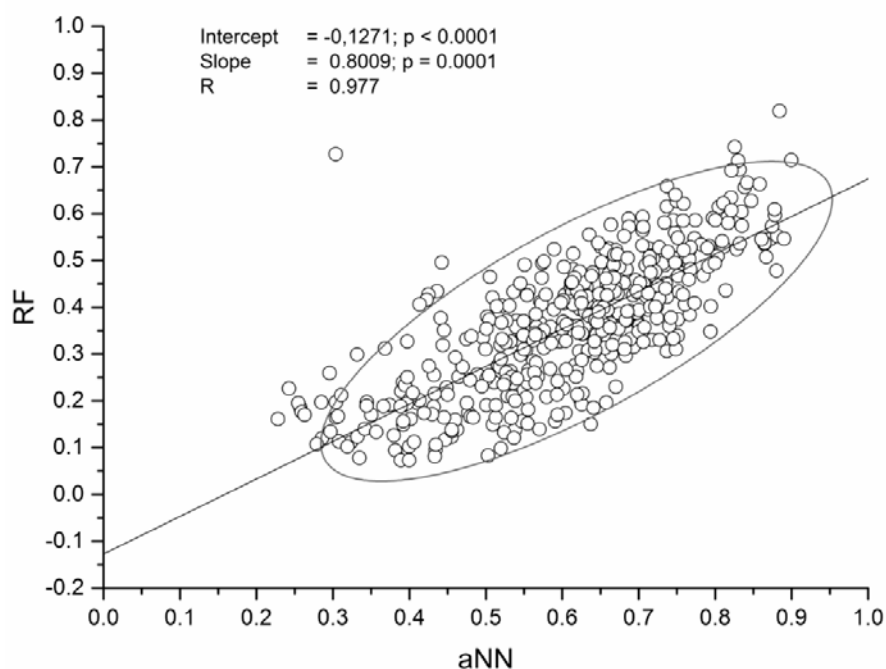


Figure 10: Correlation of the hepatotoxic potential of single PAs as predicted by the RF and the aNN model.

Intercept = -0.1271 ($p < 0.0001$), slope = 0.8009 ($p = 0.0001$), $R = 0.977$

For selected single PAs the prediction of our models was compared to the reported *in vivo* hepatotoxic potential in literature. Monocrotaline (DeLeve et al. 1996; Yang et al. 2017; Zhang et al. 2017; Zheng et al. 2016), riddelliine (NTP 2003; Schoental & Head 1957) and lasiocarpine (NTP 1978) are known hepatotoxic PAs, whereas retronecine and lycopsamine did not show hepatotoxic potential *in vivo* (Xia et al. 2013). Accordingly, in both models, the

former three, hepatotoxic PAs had much higher probabilities of being hepatotoxic (RF model: 47%, 47%, and 48%; aNN model: 76%, 72%, and 67%, respectively) than the latter two, non-hepatotoxic PAs (RF model: 16% and 16%; aNN model: 40% and 48%, respectively).

To closer investigate the distribution of the probabilities within the single groups the cumulative percentage of PAs was plotted against the probability of hepatotoxicity (see Figure 11). In general, a curve that is more on the left side of the plotting area, indicates that the group has a lower overall probability to be hepatotoxic than a curve that is shifted more to the right.

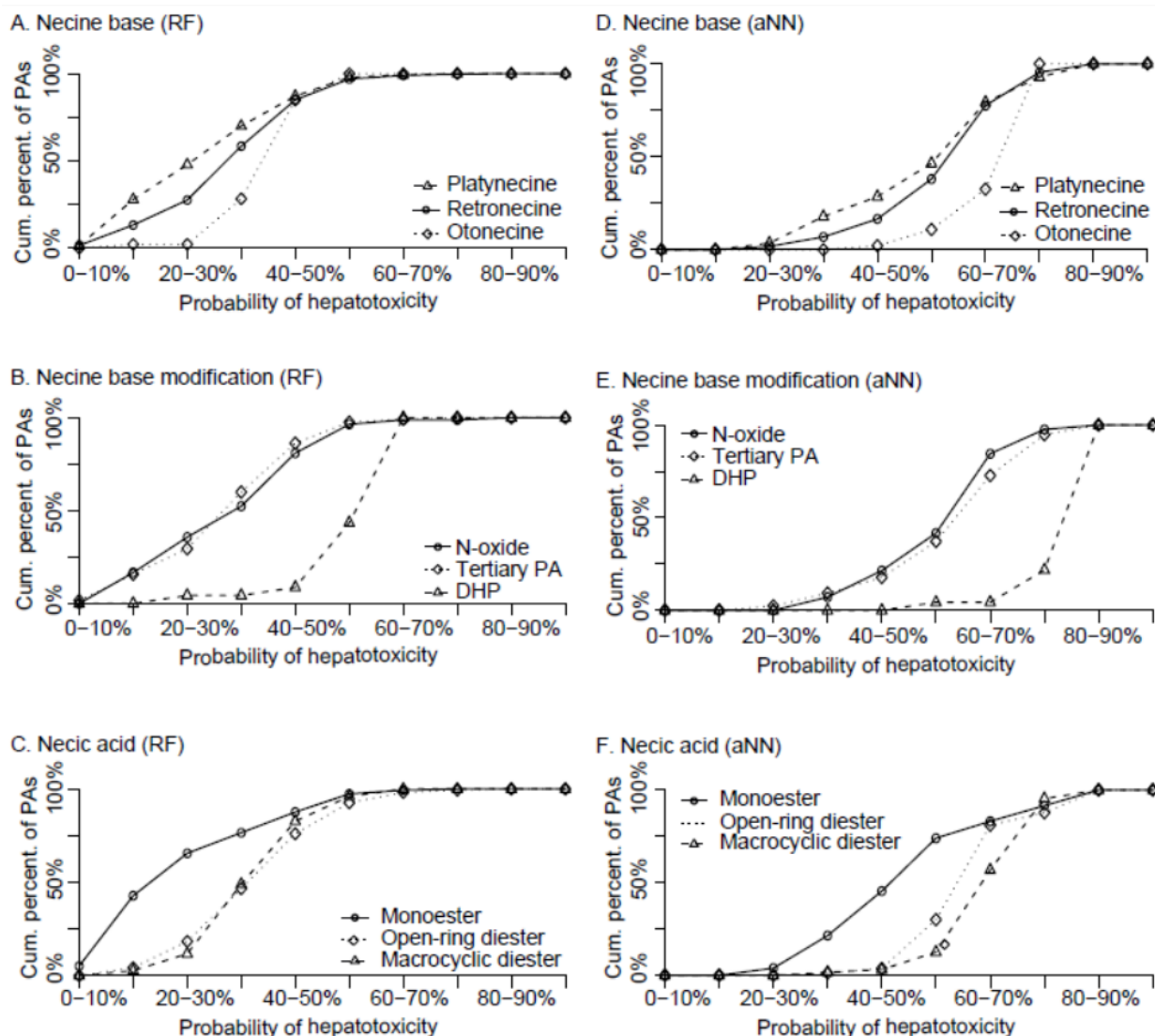


Figure 11: Cumulative number of PA (in percent) in structural feature groups versus the probability of hepatotoxicity.

DHP: dehydropyrrolizidine, RF: Random Forest, aNN: artificial Neural Network. A shift of the curve to the right indicates a higher probability of hepatotoxicity, a shift to the left a lower probability.

- A: All groups are significantly different from each other ($p < 0.001$)
- B: DHP are significantly different from the other two groups ($p < 0.001$)
- C: Monoester are significantly different from the other two groups ($p < 0.001$)
- D: All groups are significantly different from each other ($p < 0.05$)
- E: DHP are significantly different from the other two groups ($p < 0.001$)
- F: All groups are significantly different from each other ($p < 0.001$)

Considering the group of the necine base, otonecine-type PAs had in the both models significantly ($p < 0.001$) higher potential for hepatotoxic potential compared to the retronecine-type PAs. Platynecine had a significantly ($p < 0.001$ in the RF model and $p < 0.05$ in the aNN

model) lower hepatotoxic potential than retronecine. Therefore, the rank order for the necine base for their hepatotoxic potential can be assumed as: otonecine > retronecine > platynecine.

Modifications of the necine base seem to have a significant influence on the prediction of hepatotoxicity. Not only is the majority of PAs from the DHP-type predicted as hepatotoxic, but also is the difference to the other two groups highly significant ($p < 0.001$) for both models. The cumulative plots show, that very few DHPs have a lower hepatotoxic potential and the curve is far more right than those from the other two groups. The difference between *N*-oxides and tertiary PAs is not significant in either model. Therefore, the rank order for the necine base modification is: DHP >> tertiary PA = *N*-oxide.

The structural features of the necic acid also determine the prediction of hepatotoxicity by the QSAR models. PAs from the macrocyclic diester-type had a significantly ($p < 0.001$) higher probability in the aNN model to be hepatotoxic compared to PAs from the other two groups. In the RF model, the difference is only significant between macrocyclic diester and monoester-type PAs. The difference between open-ring diester- and monoester-type PAs is significant ($p < 0.001$) in both models. PAs with a monoester as necic acid have the lowest probability to be predicted as hepatotoxic. The rank order for the necic acid is therefore: macrocyclic diester \geq open-ring diester > monoester.

To better characterise the influence of the necine base and the necic acid on the hepatotoxic potential, the combination of structural features was investigated. The boxplots of the results are presented in Figure 12. Unfortunately, the number of substances in some groups was very low (indicated by a dollar sign); therefore, the otonecine- and the platynecine-*N*-oxide group could only partly or not at all be included in the evaluation. However, a clear trend is observable in both models. The hepatotoxic probabilities of PAs with the same necine base (retronecine, retronecine-*N*-oxide, and platynecine) but different necic acids are almost always significantly

($p < 0.05$) different (except for platynecine open-ring diester and platynecine macrocyclic diester in the aNN model), with the same rank order as in the evaluation of the single PA features. In contrast, despite different necine bases, PAs with the same necic acids seemed to have comparable hepatotoxic probabilities.

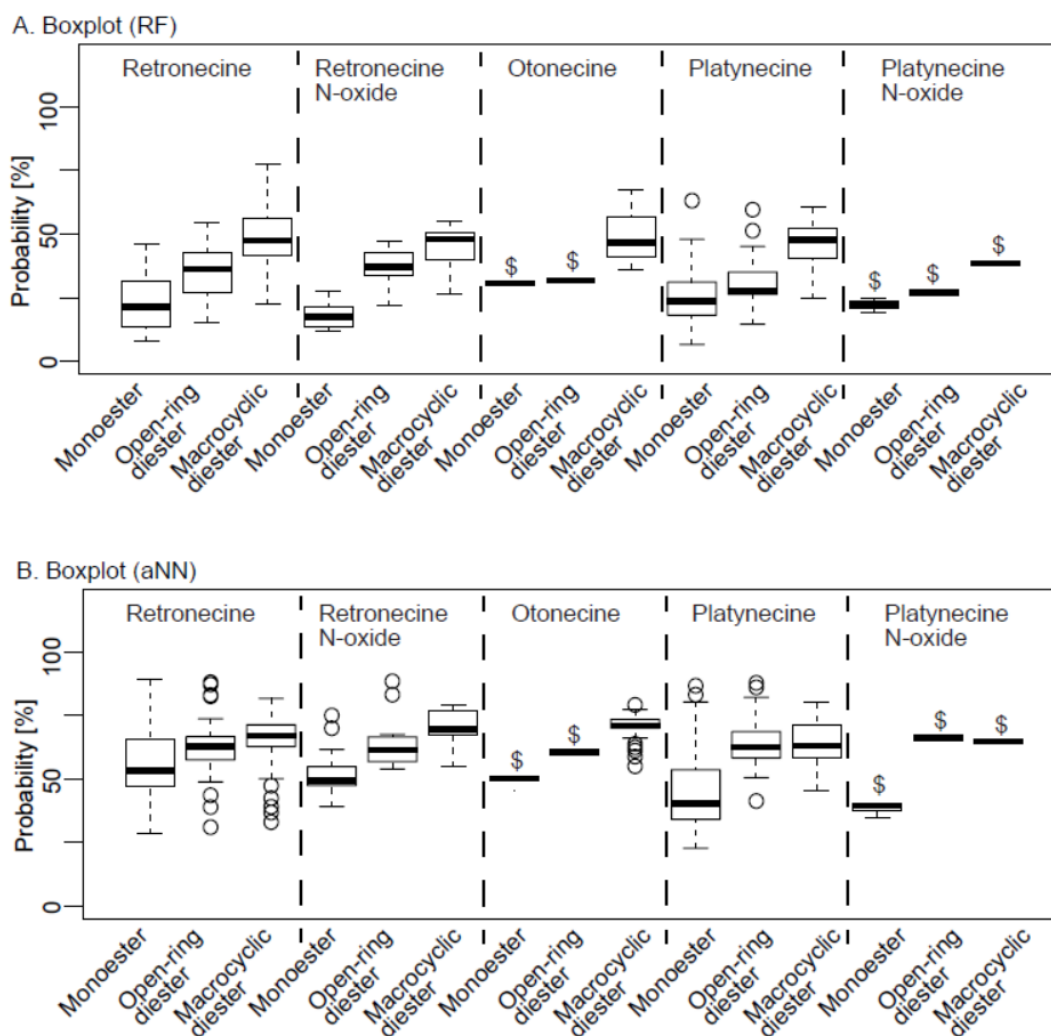


Figure 12: Boxplots of the combined PA-structures, the necine base is indicated above the boxplot, the necic acid below.

RF: Random Forest, aNN: artificial Neural Network, \$ denotes groups comprising of less than 10 PAs. In the boxplot, the median is indicated by a horizontal line, the bottom and top of the box are the 25th (P25%) and 75th (P75%) percentile, the whiskers are the P75% or P25% plus or minus 1.5*IQR, respectively. Outliers are indicated as open circles.

The investigation on combined structural features clearly suggests, that the necic acid has a higher influence on the hepatotoxicity probability of PAs than the necine base.

3.5 Discussion

Relatively early during the investigation of the toxicity of PAs, a relationship between hepatotoxicity and structure was assumed (Mattocks 1986). This relationship was repeatedly confirmed in different *in vitro* studies with different toxicological endpoints (Fu et al. 2004; Kim et al. 1993; Li et al. 2013; Ruan et al. 2014a; Ruan et al. 2014b; Xia et al. 2013). Factors contributing to the structure-toxicity relationship of PAs are e.g. different modes of action (direct cytotoxicity *vs.* genotoxicity), different pathways and rates of metabolic activation, leading to different amounts of DHP, and different pathways and rates of detoxification.

A drawback of *in vitro* and *in vivo* studies is that the number of different PAs tested is usually limited and dependent on the about 35 different, commercially available PAs. Therefore, more or less the same PAs are tested and compared over and over again.

Other *in silico* studies, which were already performed with PAs, can be considered as further evidence that the structure of pyrrolizidine alkaloids has an influence on the bioactivation and toxicity. Srinivas et al. (2014) modelled different structural alterations of monocrotaline and tested them for toxicity reduction in different *in silico* models. Some structural alterations showed a significant reduction in toxicity and bio-availability accompanied by drug-likeness properties. Fashe et al. (2015) used three different *in silico* analyses (ligand-based Fukui electrophilic Fukui function, hydrogen bond dissociation energies, and structure-based molecular docking) to identify the site of oxidation by CYP 3A4 in the toxification pathway leading to the DHPs for two PAs from the retronecine-type and one from the otonecine-type. Interestingly, the sites of oxidation were different for the two different necine base-types studied. However, the *in silico* studies also focused on very few PAs.

The present study analysed a comprehensive number of 602 different PAs with human DILI outcome data with two different machine learning techniques. Even though PAs are structurally a quite homogenous substance class, both models were able to assign different hepatotoxic potential to structural features and thereby, were able to confirm a structure-toxicity relationship.

Even though, the RF model had a better performance in the validation (correct classification 89% vs 76%), the separation of the structural features in both models is comparable.

The predicted hepatotoxic probability of single PAs (monocrotaline, riddelliine, retronecine, lasiocarpine, lycopsamine) by the two models was qualitatively comparable to the hepatotoxic potential reported in literature.

However, there are also noteworthy differences between the RF model and published literature data. Even though monocrotaline, riddelliine and lasiocarpine are considered as hepatotoxic in *in vitro* (Field et al. 2015; Ruan et al. 2014b) and *in vivo* experiments (Xia et al. 2013), the probability in RF model were only 47%, 47%, and 48%, respectively. In terms of binary classification (cut off 50%), these PAs would have been classified as not hepatotoxic by the RF model. However, considering the percentage value, other conclusion should be drawn. In general, values around 50% indicate a low confidence of the prediction and are therefore difficult to interpret (Breimann 2003). Therefore, the values for these PAs do not mean, that these substances can be considered as not hepatotoxic, but that the prediction lacks confidence. Furthermore, it has to be taken into consideration, that the DILI dataset is based on experience with drugs in humans. However, the data for these three PAs are derived from *in vitro* (in cells of different origin) and animal experiments with different experimental designs. As the main purpose of the present study is to perform a qualitative analysis of PAs, relating structural features to the probability of toxicity, low confidence predictions (with a probability of around

50%) do not principally limit the overall conclusion, but may indicate that these should be interpreted with caution.

The rank orders of the different structural features of both models are generally comparable to each other. Furthermore, the identified ranking fits to the toxification and detoxification pathways of PAs. The most indicative structural feature for hepatotoxicity is DHP. DHP is the reactive pyrrolic ester of the toxification pathway and the actual toxic principle of PAs. Both models identified this feature as most reliable predictor for hepatotoxicity. This is also in compliance with the observations by Kim et al. (1993), who compared the cytotoxicity of DHP with their parent compound.

In contrast, PAs with an *N*-oxide structural motive or tertiary PAs were less likely to be predicted as hepatotoxic. However, the difference between these two groups was not significant in both models. PA *N*-oxides are generally regarded as detoxification products as the metabolites can be conjugated for excretion (Chen et al. 2010). Accordingly, *N*-oxides are more easily eliminated from the body (Chen et al. 2010). As *N*-oxides can be easily transformed back to the corresponding tertiary PA (Wang et al. 2005) it may be questioned, whether *N*-oxides themselves are generally less toxic than the corresponding tertiary PAs or rather whether reduced toxicity may result only from the reduced pool of retained *N*-oxides only.

Within the necine base group, otonecine-type PAs have the highest probability to be hepatotoxic in both models. This might be due to the methylated nitrogen in the necine base, which disables it for direct *N*-oxidation. This would be in concordance with observations by Li et al. (2013), but not with the study from Ruan et al. (2014b), who found retronecine-type PA to be more toxic than otonecine-type PAs. The saturated platynecine-type PAs had the lowest hepatotoxic probability in both models. This is in agreement with the general view of the

platynecine-type PAs, which are considered as less/non-toxic than 1,2-unsaturated PAs (Fu et al. 2004; Ruan et al. 2014a).

In addition, the analysis of combined structural features revealed, that the necic acid was more strongly correlated with the toxic potential of a PA than the necine base.

Especially as the necic acids are sometimes quite large structures, steric hindrance might be involved with enzymes along the toxification and detoxification pathway. Several experimental observations led various authors to the conclusion that macrocyclic diesters are more toxic than open-ring diesters and monoesters (EFSA 2011; Fu et al. 2010; Ruan et al. 2014b). Furthermore, open-ring diesters were shown to be more toxic than monoesters (Ruan et al. 2014b; Tamta et al. 2012). These observations are in agreement with the results of the aNN model. However, in the RF model, the difference between open-ring diester and macrocyclic diester is not significant.

The fact, that open-ring diesters are more likely to be hepatotoxic than monoester might be explained by the hydrolysis detoxification pathway. In this pathway, the necine base and the necic acid are separated. For open-ring diesters, this would include two steps (one for each ester arm), for monoesters only one.

In contrast to earlier experiments and the present study, the experiments by Ruan et al. (2014b) indicated, that open-ring diesters had a higher metabolic activation rate than macrocyclic diesters, resulting in a higher efficiency of adduct formation. Interestingly, the PAs used in this study all had the same necine base.

In the last few years, PAs, especially in herbal medicinal products, became a widely discussed issue, with the European Medicinal Agency striving for a reduction of PAs in herbal medicinal products (EMA 2014; 2016). The limits are set for all PAs on the basis of toxicological animal

studies with only one PA (lasiocarpine). Considering the evident structural-toxicity relationship it is recommended to establish rather a rank order of known PAs, calculated in lasiocarpine-equivalents.

In a next step, additional outcomes (e.g. chronic toxicity) should be modelled *in silico* (genotoxic/ carcinogenic potential of PAs). Also, further *in silico* investigations addressing the influence of the various structural moieties of PAs on the activity of the enzymes involved in PA metabolism (cytochrome P450, carboxyl esterase, UDP-glucuronosyltransferase) could shed further light not only on the structure-toxicity relationship, but also on the pronounced differences in sensitivity between species for hepatotoxicity effects of PAs (partly due to different expression levels of metabolic enzymes) (EFSA 2011).

4 Prediction of the mutagenic potential of different pyrrolizidine alkaloids using LAZAR, Random Forest, Support Vector Machines, and Deep Learning

Authors

Verena Schöning, Christoph Helma, Philipp Boss, Jürgen Drewe

Manuscript in preparation.

Corresponding author:

Prof. Dr. Jürgen Drewe, MSc

4.1 Abstract

Pyrrolizidine alkaloids (PAs) are secondary plant metabolites of some plant families, which protect against predators and generally considered as genotoxic and mutagenic. This mutagenicity is also the point of concern in regulatory risk assessment of this substance group (EFSA 2011; EMA 2014; 2016). Several investigations already showed that the mutagenic potential of PAs is different, and largely depends on the structure.

Since only very few of over 600 known PAs are available for *in vitro* or *in vivo* experiments, the mutagenicity of PAs in this study was estimated using four different machine learning techniques LAZAR and Deep Learning, Random Forest and Support Vector Machines. However, all models were not optimal for predicting the genotoxic potential of PAs either due to problems with the applicability domain or due to low performance. Therefore, no estimation regarding the genotoxic potential of single PAs could be made. An analysis of the genotoxic

potential of different structural groups, showed promising results. For necine base and necic acid, the results fitted well with literature for three models. However, the prediction of the toxic principle of PAs, dehydropyrrolizidine was only within expectation in one model (TensorFlow-generated Deep Learning model), but not in the other four models. This study shows convincingly the need to critically review and assess the predictions obtained from machine learning approaches by internal cross-validation, but also by external validation through comparison with literature.

4.2 Introduction

Pyrrolizidine alkaloids (PAs) are secondary plant ingredients found in many plant species as protection against predators (Hartmann & Witte 1995; Langel et al. 2011). PAs are ester alkaloids, which are composed of a necine base (two fused five-membered rings joined by a nitrogen atom) and one or two necic acid (carboxylic ester arms). The necine base can have different structures and thereby divides PAs into several structural groups, e.g. otonecine, platynecine, and retronecine. The structural groups of the necic acid are macrocyclic diester, open-ring diester and monoester (Langel et al. 2011).

PA are mainly metabolised in the liver, which is at the same time the main target organ of toxicity (Bull & Dick 1959; Bull et al. 1958; Butler et al. 1970; DeLeve et al. 1996; Jago 1971; Li et al. 2011; Neumann et al. 2015). There are three principal metabolic pathways for 1,2-unsaturated PAs (Chen et al. 2010): (i) Detoxification by hydrolysis: the ester bond on positions C7 and C9 are hydrolysed by non-specific esterases to release necine base and necic acid, which are then subjected to further phase II-conjugation and excretion. (ii) Detoxification by *N*-oxidation of the necine base (only possible for retronecine-type PAs): the nitrogen is oxidised to form a PA *N*-oxides, which can be conjugated by phase II enzymes e.g. glutathione and then

excreted. PA *N*-oxides can be converted back into the corresponding parent PA (Wang et al. 2005). (iii) Metabolic activation or toxification: PAs are metabolic activated/ toxified by oxidation (for retronecine-type PAs) or oxidative *N*-demethylation (for otonecine-type PAs (Lin 1998)). This pathway is mainly catalysed by cytochrome P450 isoforms CYP2B and 3A (Ruan et al. 2014b), and results in the formation of dehydropyrrolizidines (DHP, also known as pyrrolic ester or reactive pyrroles). DHPs are highly reactive and cause damage in the cells where they are formed, usually hepatocytes. However, they can also pass from the hepatocytes into the adjacent sinusoids and damage the endothelial lining cells (Gao et al. 2015) predominantly by reaction with protein, lipids and DNA. There is even evidence, that conjugation of DHP to glutathione, which would generally be considered a detoxification step, could result in reactive metabolites, which might also lead to DNA adduct formation (Xia et al. 2015). Due to the ability to form DNA adducts, DNA crosslinks and DNA breaks 1,2-unsaturated PAs are generally considered genotoxic and carcinogenic (Chen et al. 2010; EFSA 2011; Fu et al. 2004; Li et al. 2011; Takanashi et al. 1980; Yan et al. 2008; Zhao et al. 2012). Still, there is no evidence yet that PAs are carcinogenic in humans (ANZFA 2001; EMA 2016). One general limitation of studies with PAs is the number of different PAs investigated. Around 30 PAs are currently commercially available, therefore all studies focus on these PAs. This is also true for *in vitro* and *in vivo* tests on mutagenicity and genotoxicity. To gain a wider perspective, in this study over 600 different PAs were assessed on their mutagenic potential using four different machine learning techniques.

4.3 Materials and Methods

4.3.1 Training dataset

For all methods, the same validated training dataset was used. The training dataset was compiled from the following sources:

- Kazius/Bursi Dataset (4337 compounds, (Kazius et al. 2005)):
http://cheminformatics.org/datasets/bursi/cas_4337.zip
- Hansen Dataset (6513 compounds, (Hansen et al. 2009)):
http://doc.ml.tu-berlin.de/toxbenchmark/Mutagenicity_N6512.csv
- EFSA Dataset (695 compounds, (EFSA 2011)):
<https://data.europa.eu/euodp/data/storage/f/2017-0719T142131/GENOTOX%20data%20and%20dictionary.xls>

Mutagenicity classifications from Kazius and Hansen datasets were used without further processing. To achieve consistency between these datasets, EFSA compounds were classified as mutagenic, if at least one positive result was found for TA98 or T100 Salmonella strains.

Dataset merges were based on unique SMILES (*Simplified Molecular Input Line Entry Specification*) strings of the compound structures. Duplicated experimental data with the same outcome was merged into a single value, because it is likely that it originated from the same experiment. Contradictory results were kept as multiple measurements in the database. The combined training dataset contains 8281 unique structures.

Source code for all data download, extraction and merge operations is publicly available from the git repository <https://git.in-silico.ch/pyrrolizidine> under a GPL3 License.

4.3.2 Testing dataset

The testing dataset consisted of 602 different PAs. The compilation of the PA dataset is described in detail in Schöning et al. (2017). The PAs were assigned to groups according to structural features of the necine base and necic acid.

For the necine base, following groups were assigned:

- Retronecine-type (1,2-unsaturated necine base)
- Otonecine-type (1,2-unsaturated necine base)
- Platynecine-type (1,2-saturated necine base)

For the modification of necine base, following groups were assigned:

- *N*-oxide-type
- Tertiary-type (PAs which were neither from the *N*-oxide- nor DHP-type)
- DHP-type (dehydropyrrolizidine, pyrrolic ester)

For the necic acid, following groups were assigned:

- Monoester-type
- Open-ring diester-type
- Macrocyclic diester-type

For the Random Forest (RF), Support Vector Machines (SVM), and Deep Learning (DL) models, molecular descriptors of the PAs were calculated using the program PaDEL-Descriptors (version 2.21) (Yap 2011; 2014). From these descriptors were chosen, which were actually used for the generation of the DL model.

4.3.3 LAZAR

LAZAR (*lazy structure activity relationships*) is a modular framework for read-across model development and validation. It follows the following basic workflow: For a given chemical structure LAZAR:

- searches in a database for similar structures (neighbours) with experimental data,
- builds a local QSAR model with these neighbours and
- uses this model to predict the unknown activity of the query compound.

This procedure resembles an automated version of read across predictions in toxicology, in machine learning terms it would be classified as a k-nearest-neighbour algorithm.

Apart from this basic workflow, LAZAR is completely modular and allows the researcher to use any algorithm for similarity searches and local QSAR (*Quantitative structure–activity relationship*) modelling. Algorithms used within this study are described in the following sections.

4.3.3.1 Neighbour identification

Similarity calculations were based on MolPrint2D fingerprints (Bender et al. 2004) from the OpenBabel cheminformatics library (O'Boyle et al. 2011). The MolPrint2D fingerprint uses atom environments as molecular representation, which resembles basically the chemical

concept of functional groups. For each atom in a molecule, it represents the chemical environment using the atom types of connected atoms.

MolPrint2D fingerprints are generated dynamically from chemical structures and do not rely on predefined lists of fragments (such as OpenBabel FP3, FP4 or MACCS fingerprints or lists of toxicophores/toxicophobes). This has the advantage that they may capture substructures of toxicological relevance that are not included in other fingerprints.

From MolPrint2D fingerprints a feature vector with all atom environments of a compound can be constructed that can be used to calculate chemical similarities.

The chemical similarity between two compounds a and b is expressed as the proportion between atom environments common in both structures $A \cap B$ and the total number of atom environments $A \cup B$ (Jaccard/Tanimoto index).

$$sim = \frac{|A \cap B|}{|A \cup B|}$$

Threshold selection is a trade-off between prediction accuracy (high threshold) and the number of predictable compounds (low threshold). As it is in many practical cases desirable to make predictions even in the absence of closely related neighbours, we follow a tiered approach:

- First a similarity threshold of 0.5 is used to collect neighbours, to create a local QSAR model and to make a prediction for the query compound.
- If any of these steps fails, the procedure is repeated with a similarity threshold of 0.2 and the prediction is flagged with a warning that it might be out of the applicability domain of the training data.
- Similarity thresholds of 0.5 and 0.2 are the default values chosen by the software developers and remained unchanged during the course of these experiments.

Compounds with the same structure as the query structure are automatically eliminated from neighbours to obtain unbiased predictions in the presence of duplicates.

4.3.3.2 Local QSAR models and predictions

Only similar compounds (neighbours) above the threshold are used for local QSAR models. In this investigation, we are using a weighted majority vote from the neighbour's experimental data for mutagenicity classifications. Probabilities for both classes (mutagenic/non-mutagenic) are calculated according to the following formula and the class with the higher probability is used as prediction outcome.

$$p_c = \frac{\sum sim_{n,c}}{\sum sim_n}$$

p_c	Probability of class c (e.g. mutagenic or non-mutagenic)
$\sum sim_{n,c}$	Sum of similarities of neighbours with class c
$\sum sim_n$	Sum of all neighbours

4.3.3.3 Applicability domain

The applicability domain (AD) of LAZAR models is determined by the structural diversity of the training data. If no similar compounds are found in the training data no predictions will be generated. Warnings are issued if the similarity threshold had to be lowered from 0.5 to 0.2 in order to enable predictions. Predictions without warnings can be considered as close to the applicability domain and predictions with warnings as more distant from the applicability domain. Quantitative applicability domain information can be obtained from the similarities of individual neighbours.

4.3.3.4 Availability

- LAZAR experiments for this manuscript: <https://git.in-silico.ch/pyrrolizidine> (source code, GPL3)
- LAZAR framework: <https://git.in-silico.ch/lazar> (source code, GPL3)
- LAZAR GUI: <https://git.in-silico.ch/lazar-gui> (source code, GPL3)
- Public web interface: <https://lazar.in-silico.ch>

4.3.4 Random Forest, Support Vector Machines, and Deep Learning in R-project

In comparison to LAZAR, three other models (Random Forest (RF), Support Vector Machines (SVM), and Deep Learning (DL)) were evaluated.

For the generation of these models, molecular 1D and 2D descriptors of the training dataset were calculated using PaDEL-Descriptors (version 2.21) (Yap 2011; 2014).

As the training dataset contained over 8280 instances, it was decided to delete instances with missing values during data pre-processing. Furthermore, substances with equivocal outcome were removed. The final training dataset contained 8080 instances with known mutagenic potential. The RF, SVM, and DL models were generated using the R software (R-project for Statistical Computing, <https://www.r-project.org/>; version 3.3.1), specific R packages used are identified for each step in the description below. During feature selection, descriptor with near zero variance were removed using ‘*NearZeroVar*’-function (package ‘*caret*’). If the percentage of the most common value was more than 90% or when the frequency ratio of the most common value to the second most common value was greater than 95:5 (e.g. 95 instances of the most common value and only 5 or less instances of the second most common value), a descriptor

was classified as having a near zero variance. After that, highly correlated descriptors were removed using the *'findCorrelation'*-function (package *'caret'*) with a cut-off of 0.9. This resulted in a training dataset with 516 descriptors. These descriptors were scaled to be in the range between 0 and 1 using the *'preProcess'*-function (package *'caret'*). The scaling routine was saved in order to apply the same scaling on the testing dataset. As these three steps did not consider the outcome, it was decided that they do not need to be included in the cross-validation of the model. To further reduce the number of features, a LASSO (*least absolute shrinkage and selection operator*) regression was performed using the *'glmnet'*-function (package *'glmnet'*). The reduced dataset was used for the generation of the pre-trained models.

For the RF model, the *'randomForest'*-function (package *'randomForest'*) was used. A forest with 1000 trees with maximal terminal nodes of 200 was grown for the prediction.

The *'svm'*-function (package *'e1071'*) with a *radial basis function kernel* was used for the SVM model.

The DL model was generated using the *'h2o.deeplearning'*-function (package *'h2o'*). The DL contained four hidden layer with 70, 50, 50, and 10 neurons, respectively. Other hyperparameter were set as follows: $l1=1.0E-7$, $l2=1.0E-11$, $\epsilon = 1.0E-10$, $\rho = 0.8$, and $\text{quantile_alpha} = 0.5$. For all other hyperparameter, the default values were used. Weights and biases were in a first step determined with an unsupervised DL model. These values were then used for the actual, supervised DL model.

To validate these models, an internal cross-validation approach was chosen (see Figure 10). The training dataset was randomly split in training data, which contained 95% of the data, and validation data, which contain 5% of the data. A feature selection with LASSO on the training data was performed, reducing the number of descriptors to approximately 100. This step was

repeated five times. Based on each of the five different training data, the predictive models were trained and the performance tested with the validation data. This step was repeated 10 times. Furthermore, a y-randomisation using the RF model was performed. During y-randomisation, the outcome (y-variable) is randomly permuted. The theory is that after randomisation of the outcome, the model should not be able to correlate the outcome to the properties (descriptor values) of the substances. The performance of the model should therefore indicate a by chance prediction with an accuracy of about 50%. If this is true, it can be concluded that correlation between actual outcome and properties of the substances is real and not by chance (Rücker et al. 2007).

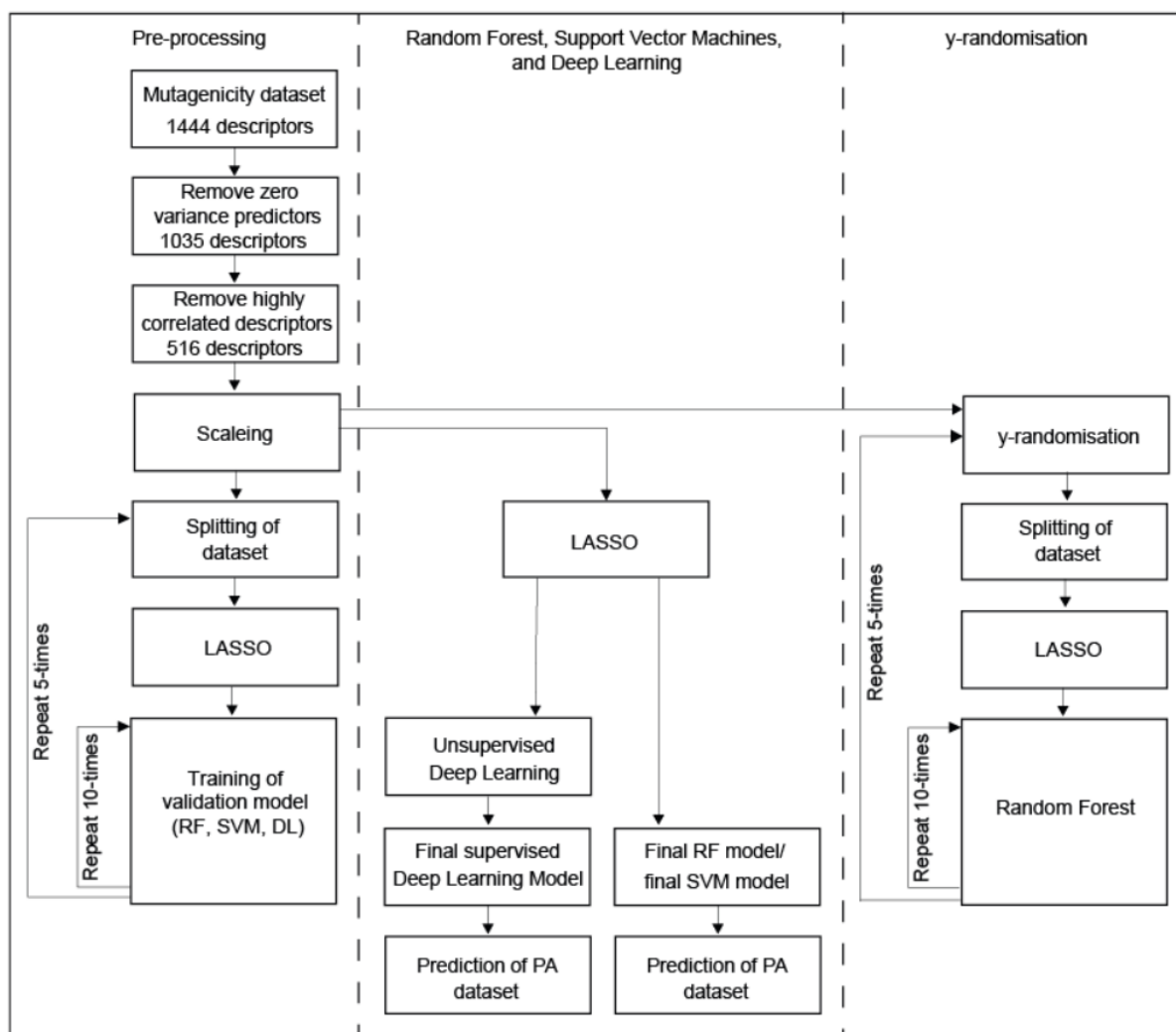


Figure 13: Flowchart of the generation and validation of the models generated in R-project

4.3.5 Deep Learning in TensorFlow

Alternatively, a DL model was established with Python-based TensorFlow program (<https://www.tensorflow.org/>) using the high-level API Keras (<https://www.tensorflow.org/guide/keras>) to build the models.

Data pre-processing was done by rank transformation using the ‘*QuantileTransformer*’ procedure. A sequential model has been used. Four layers have been used: input layer, two hidden layers (with 12, 8 and 8 nodes, respectively) and one output layer. For the output layer, a sigmoidal activation function and for all other layers the ReLU (‘*Rectified Linear Unit*’)

activation function was used. Additionally, a L^2 -penalty of 0.001 was used for the input layer. For training of the model, the ADAM algorithm was used to minimise the cross-entropy loss using the default parameters of Keras. Training was performed for 100 epochs with a batch size of 64. The model was implemented with Python 3.6 and Keras. For training of the model, a 6-fold cross-validation was used. Accuracy was estimated by ROC-AUC and confusion matrix.

4.4 Results

4.4.1 LAZAR

For 46 PAs, no prediction could be made. 26 PAs had no neighbours and 20 PAs had only one neighbour. For additional 396 PAs, the similarity threshold had to be reduced from 0.5 to 0.2 to obtain enough neighbours for a prediction. This means that these substances might not be within the applicability domain (AD). Therefore, only 160 of 602 PAs were well within the stricter AD with the similarity threshold of 0.5 and 556 PAs in the AD with the similarity threshold of 0.2.

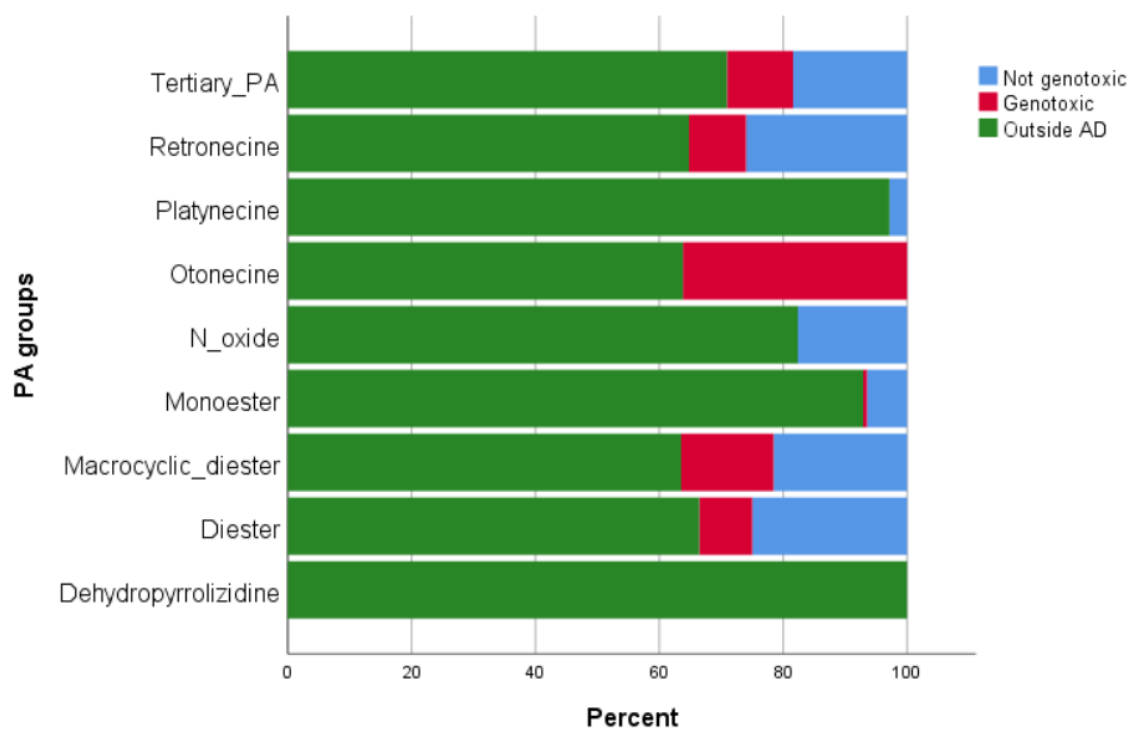


Figure 14: Genotoxic potential of the different PA groups as predicted by LAZAR, using the **similarity threshold of 0.5**.

Genotoxic: percentage number of compounds per group, which were predicted to be genotoxic.

Not genotoxic: percentage number of compounds per group, which were predicted to be not genotoxic

Outside AD: percentage number of compounds per group, which were outside the applicability domain (AD).

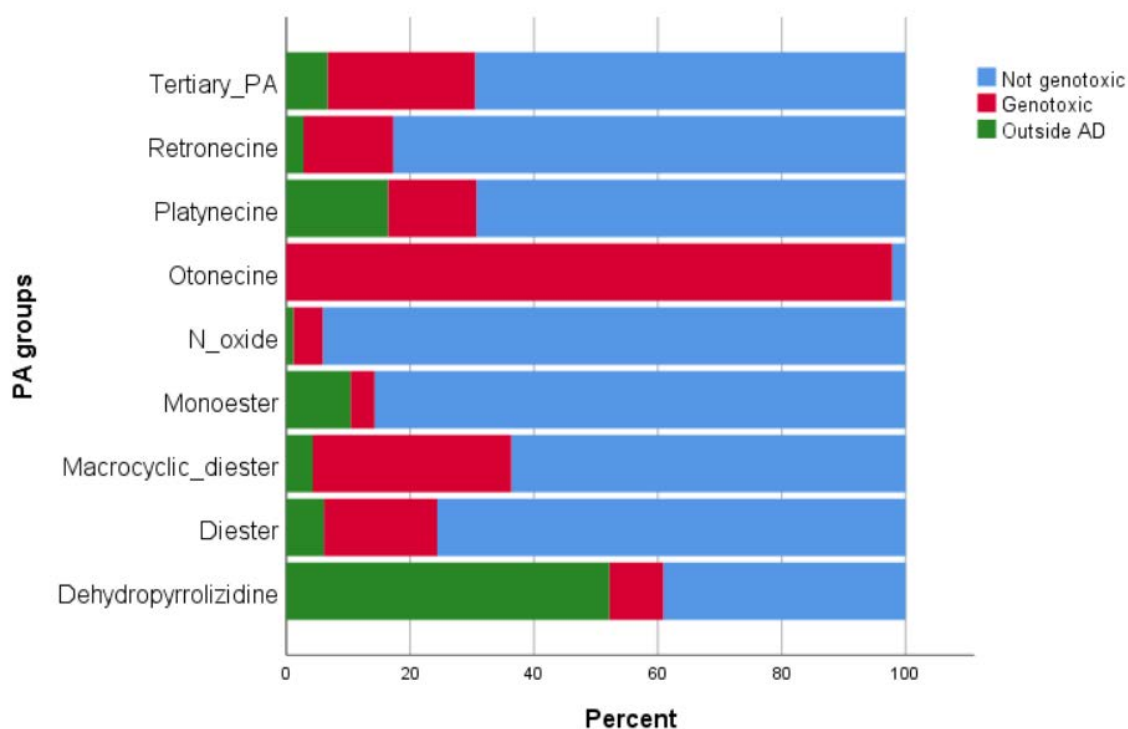


Figure 15: Genotoxic potential of the different PA groups as predicted by LAZAR, using the **similarity threshold of 0.2**

Genotoxic: percentage number of compounds per group, which were predicted to be genotoxic.

Not genotoxic: percentage number of compounds per group, which were predicted to be not genotoxic

Outside AD: percentage number of compounds per group, which were outside the applicability domain (AD).

Interestingly, using both similarity thresholds (e.g. 0.2 and 0.5), the majority of PAs in all groups except otonecine, were predicted to be not genotoxic.

The following rank order for genotoxicity probability can be deduced from the results of both similarity thresholds:

- Necine base: platynecine \leq retronecine \ll otonecine
- Necic acid: monoester < diester < macrocyclic diester
- Modification of necine base: *N*-oxide < DHP < tertiary PA

4.4.2 Random Forest, Support Vector Machines, and Deep Learning

Applicability domain

The AD of the training dataset and the PA dataset was evaluated using the Jaccard distance. A Jaccard distance of '0' indicates that the substances are similar, whereas a value of '1' shows that the substances are different. The Jaccard distance was below 0.2 for all PAs relative to the training dataset. Therefore, PA dataset is within the AD of the training dataset and the models can be used to predict the genotoxic potential of the PA dataset.

y-randomisation

After y-randomisation of the outcome, the accuracy and CCR are around 50%, indicating a chance in the distribution of the results. This shows, that the outcome is actually related to the predictors and not by chance.

Random Forest

The validation showed that the RF model has an accuracy of 64%, a sensitivity of 66% and a specificity of 63%. The confusion matrix of the model, calculated for 8080 instances, is provided in Table 3.

Table 3: Confusion matrix of the RF model

	Predicted genotoxicity			<i>Total</i>
		<i>PP</i>	<i>PN</i>	
Measured genotoxicity	<i>TP</i>	2274	1163	3437
	<i>TN</i>	1736	2907	4643
<i>Total</i>	4010	4070	8080	

PP: Predicted positive; PN: Predicted negative, TP: True positive, TN: True negative

In general, the majority of PAs were considered to be not genotoxic by the RF model (Figure 16).

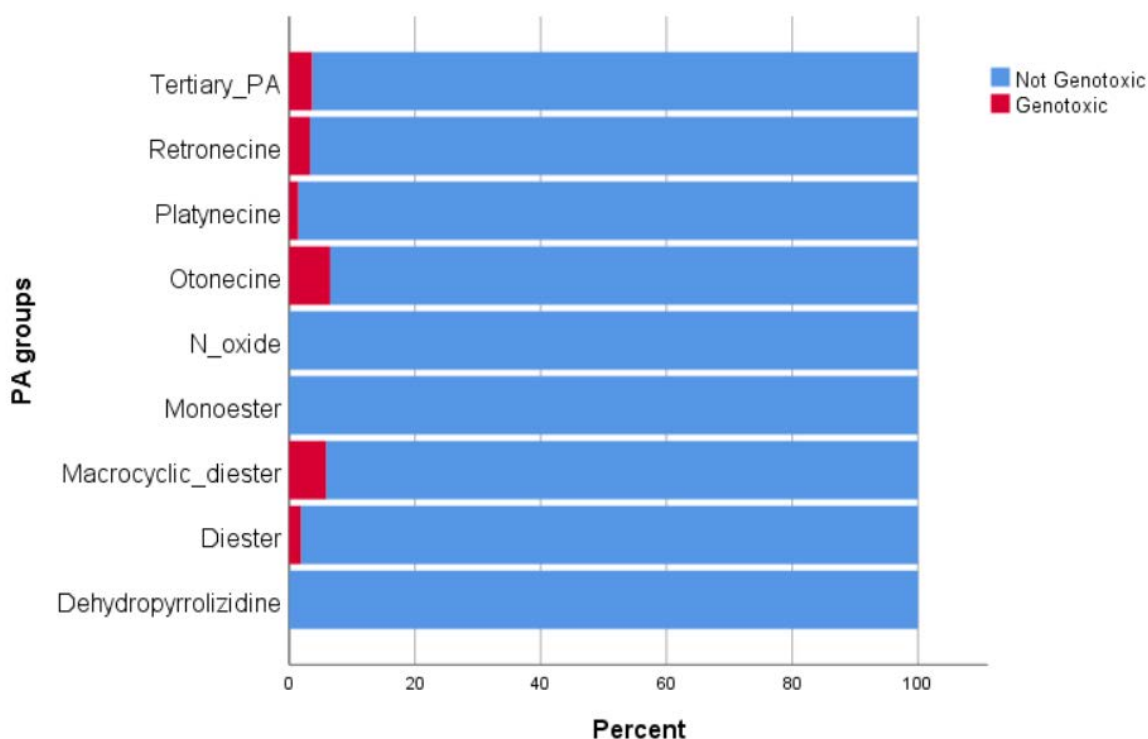


Figure 16: Genotoxic potential of the different PA groups as predicted by **RF model**
Genotoxic: percentage number of compounds per group, which was predicted to be genotoxic.
Not genotoxic: percentage number of compounds per group, which was predicted to be not genotoxic.

From the results, the following rank orders of genotoxic potential could be deduced:

- Necine base: platynecine < retronecine < otonecine
- Necic acid: monoester (= 0%) < diester < macrocyclic diester
- Modification of necine base: *N*-oxide = dehydropyrrolizidine (0%) < tertiary PA

Support Vector Machines

The validation showed that the SVM model has an accuracy of 62%, a sensitivity of 65% and a specificity of 60%. The confusion matrix of SVM model, calculated for 8080 instances, is provided in Table 4.

Table 4: Confusion matrix of the SVM model

Measured genotoxicity	Predicted genotoxicity			Total
		<i>PP</i>	<i>PN</i>	
<i>TP</i>		2057	1107	3164
<i>TN</i>		1953	2963	4916
Total		4010	4070	8080

PP: Predicted positive; PN: Predicted negative, TP: True positive, TN: True negative

In the SVM model, also the majority of PAs were considered to be not genotoxic (Figure 17).

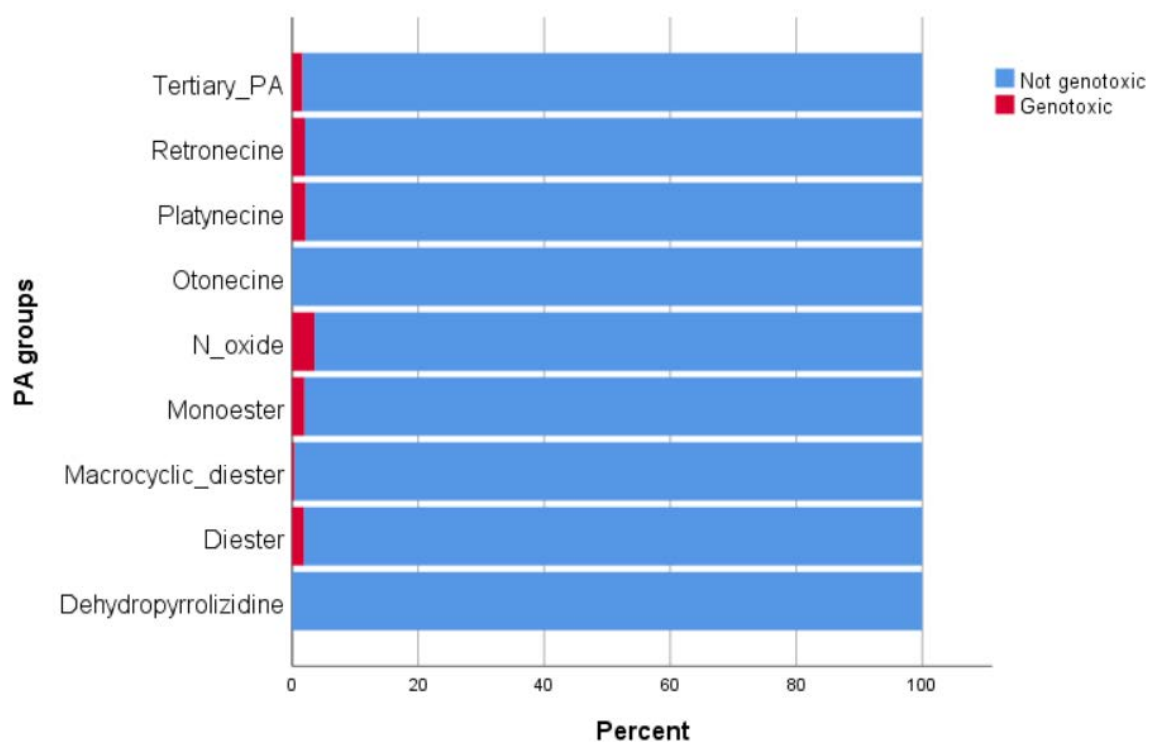


Figure 17: Genotoxic potential of the different PA groups as predicted by **SVM model**

Genotoxic: percentage number of compounds per group, which was predicted to be genotoxic.

Not genotoxic: percentage number of compounds per group, which was predicted to be not genotoxic

From the results, the following rank orders of genotoxic potential could be deduced:

- Necine base: otonecine < platynecine = retronecine
- Necic acid: macrocyclic diester < monoester = diester
- Modification of necine base: dehydropyrrolizidine < tertiary PA < N-oxide

Deep Learning (R-project)

The validation showed that the DL model generated in R has an accuracy of 59%, a sensitivity of 89% and a specificity of 30%. The confusion matrix of the model, normalised to 8080 instances, is provided in Table 5.

Table 5: Confusion matrix of the DL model (R-project)

Measured genotoxicity	Predicted genotoxicity			Total
		<i>PP</i>	<i>PN</i>	
<i>TP</i>		3575	435	4010
<i>TN</i>		2853	1217	4070
<i>Total</i>		6428	1652	8080

PP: Predicted positive; PN: Predicted negative, TP: True positive, TN: True negative

In contrast, the majority of PAs were considered to be genotoxic by the DL model in R (Figure 18).

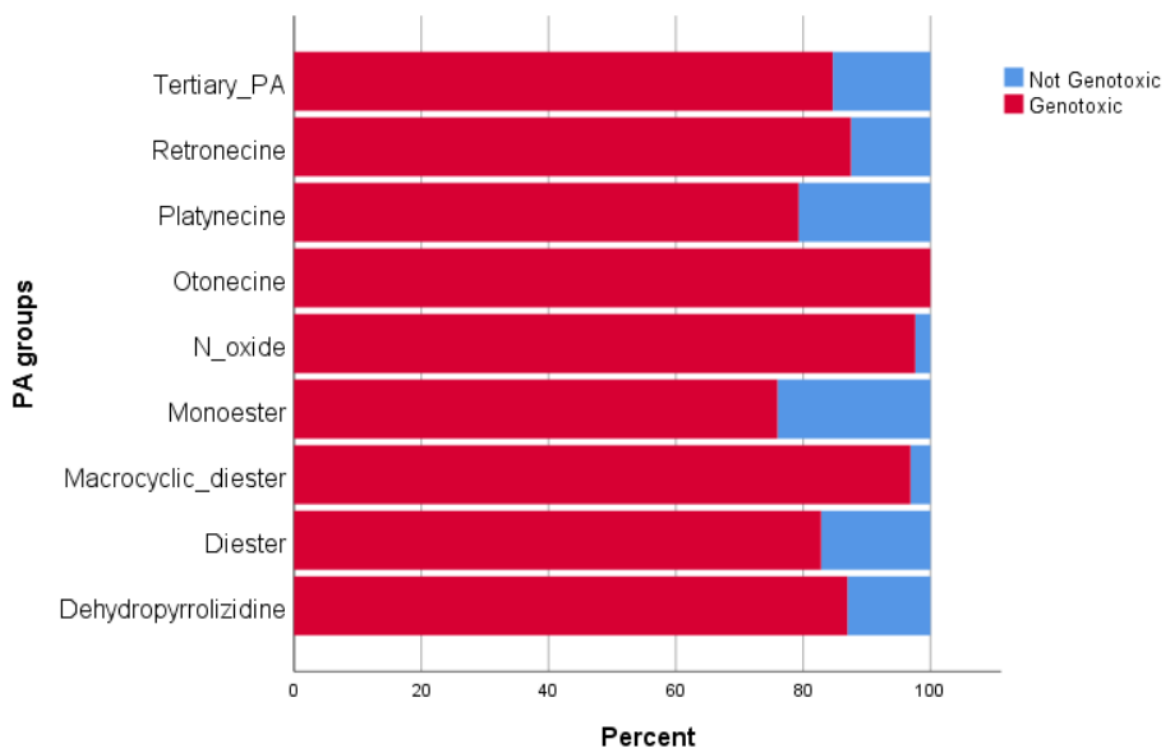


Figure 18: Genotoxic potential of the different PA groups as predicted by **DL model (R-project)**

Genotoxic: percentage number of compounds per group, which was predicted to be genotoxic.
Not genotoxic: percentage number of compounds per group, which was predicted to be not genotoxic

From the results, the following rank orders of genotoxic potential could be proposed:

- Necine base: platynecine < retronecine < otonecine
- Necic acid: monoester < diester < macrocyclic diester
- Modification of necine base: tertiary PA = dehydropyrrolizidine < *N*-oxide.

DL model (TensorFlow)

The validation showed that the DL model generated in TensorFlow has an accuracy of 68%, a sensitivity of 70% and a specificity of 46%. The confusion matrix of the model, normalised to 8080 instances, is provided in Table 6.

Table 6: Confusion matrix of the DL model (TensorFlow)

		Predicted genotoxicity		
		<i>PP</i>	<i>PN</i>	
Measured genotoxicity	<i>TP</i>	2851	1227	4078
	<i>TN</i>	1825	2177	4002
<i>Total</i>		4676	3404	8080

PP: Predicted positive; PN: Predicted negative, TP: True positive, TN: True negative

The ROC curves from the 6-fold validation are shown in Figure 19.

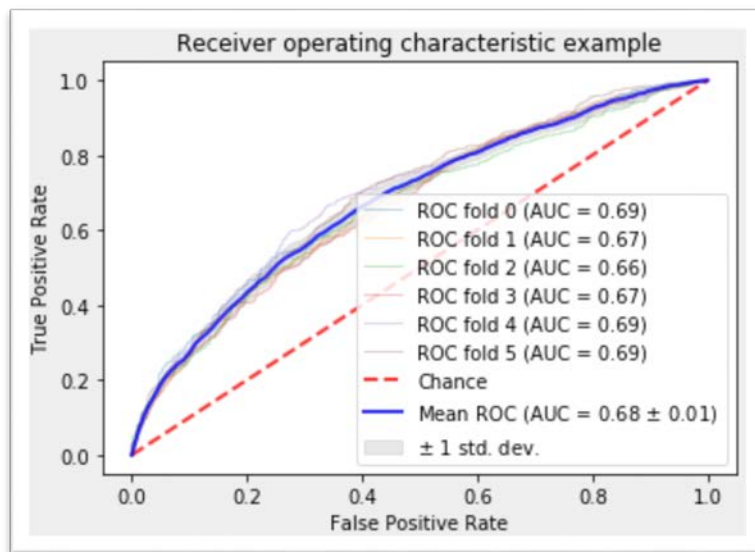


Figure 19: Six-fold cross-validation of TensorFlow DL model show an average area under the ROC-curve (ROC-AUC; measure of accuracy) of 68%.

In contrast to the DL generated in R, the DL model generated in TensorFlow predicted the majority of PAs as not genotoxic.

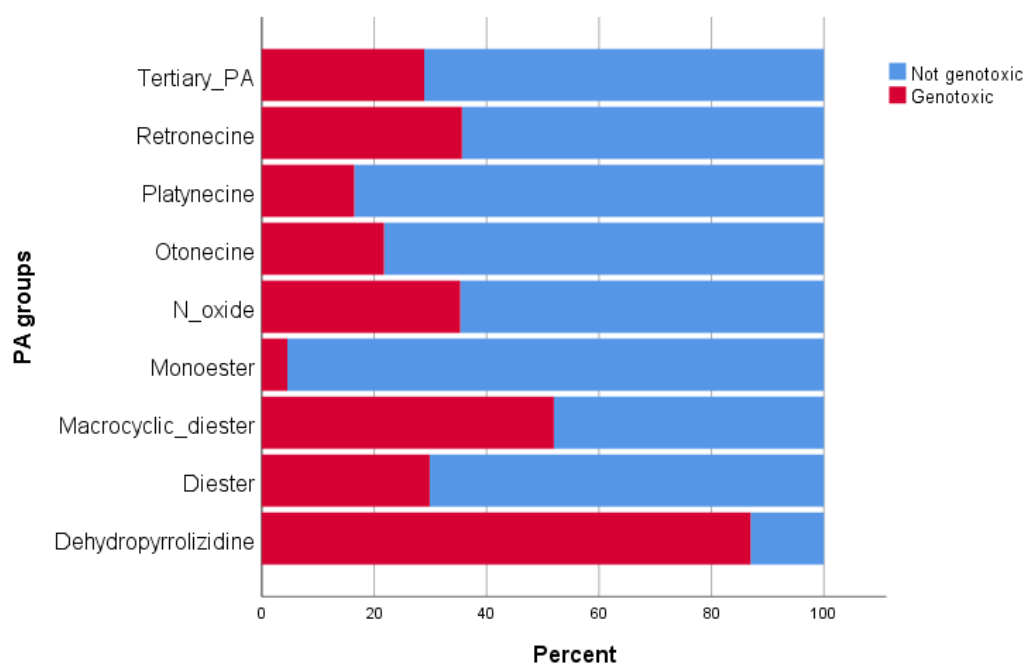


Figure 20: Genotoxic potential of the different PA groups as predicted by **DL model (TensorFlow)**

Genotoxic: percentage number of compounds per group, which was predicted to be genotoxic.

Not genotoxic: percentage number of compounds per group, which was predicted to be not genotoxic

The following rank orders of genotoxic potential could be proposed based on the results:

- Necine base: platynecine < otonecine < retronecine
- Necic acid: monoester < diester < macrocyclic diester
- Modification of necine base: tertiary PA < *N*-oxide << dehydropyrrolizidine.

In summary, the validation results of the four methods are presented in the following table.

Table 7 Results of the cross-validation of the four models and after y-randomisation

	Accuracy	CCR	Sensitivity	Specificity
RF model	64.1%	64.4%	66.2%	62.6%
SVM model	62.1%	62.6%	65.0%	60.3%
DL model (R-project)	59.3%	59.5%	89.2%	29.9%
DL model (TensorFlow)	68%	62.2%	69.9%	45.6%
y-randomisation	50.5%	50.4%	50.3%	50.6%

CCR (correct classification rate)

4.5 Discussion

General model performance

Based on the results of the cross-validation for all models, LAZAR, RF, SVM, DL (R-project) and DL (TensorFlow) it can be state that the prediction results are not optimal due to different reasons. The accuracy as measured during cross-validation of the four models (RF, SVM, DL (R-project and TensorFlow)) was partly low with CCR values between 59.3 and 68%, with the R-generated DL model and the TensorFlow-generated DL model showing the worst and the best performance, respectively. The validation of the R-generated DL model revealed a high sensitivity (89.2%) but an unacceptably low specificity of 29.9% indicating a high number of false positive estimates. The TensorFlow-generated DL model, however, showed an acceptable but not optimal accuracy of 68%, a sensitivity of 69.9% and a specificity of 45.6%. The low specificity indicates that both DL models tends to predict too many instances as positive (genotoxic), and therefore have a high false positive rate. This allows at least with the TensorFlow generated DL model to make group statements, but the confidence for estimations of single PAs appears to be insufficiently low.

Several factors have likely contributed to the low to moderate performance of the used methods as shown during the cross-validation:

1. The outcome in the training dataset was based on the results of AMES tests for genotoxicity (ICH 2011), an *in vitro* test in different strains of the bacteria *Salmonella typhimurium*. In this test, mutagenicity is evaluated with and without prior metabolic activation of the test substance. Metabolic activation could result in the formation of genotoxic metabolites from non-genotoxic parent compounds. However, no distinction was made in the training dataset between substances that needed metabolic activation before being mutagenic and those that were mutagenic without metabolic activation. LAZAR is able to handle this ‘inaccuracy’ in the training dataset well due to the way the algorithm works: LAZAR predicts the genotoxic potential based on the neighbours of substances with comparable structural features, considering mutagenic and not mutagenic neighbours. Based on the structural similarity, a probability for mutagenicity and no mutagenicity is calculated independently from each other (meaning that the sum of probabilities does not necessarily adds up to 100%). The class with the higher outcome is then the overall outcome for the substance.

In contrast, the other models need to be trained first to recognise the structural features that are responsible for genotoxicity. Therefore, the mixture of substances being mutagenic with and without metabolic activation in the training dataset may have adversely affected the ability to separate the dataset in two distinct classes and thus explains the relatively low performance of these models.

2. Machine learning algorithms try to find an optimized solution in a high-dimensional (one dimension per each predictor) space. Sometimes these methods do not find the global optimum of estimates but only local (not optimal) solutions. Strategies to find the global solutions are systematic variation (grid search) of the hyperparameters of the methods, which may be very time consuming in particular in large datasets.

Mutagenicity of PAs

Due to the low to moderate predictivity of all models, quantitative statement on the genotoxicity of single PAs cannot be made with sufficient confidence.

The predictions of the SVM model did not fit with the other models or literature, and are therefore not further considered in the discussion.

Necic acid

The rank order of the necic acid is comparable in the four models considered (LAZAR, RF and DL (R-project and TensorFlow)). PAs from the monoester type had the lowest genotoxic potential, followed by PAs from the open-ring diester type. PAs with macrocyclic diesters had the highest genotoxic potential. The result fit well with current state of knowledge: in general, PAs, which have a macrocyclic diesters as necic acid, are considered more toxic than those with an open-ring diester or monoester (EFSA 2011; Fu et al. 2004; Ruan et al. 2014b).

Necine base

The rank order of necine base is comparable in LAZAR, RF, and DL (R-project) models: with platynecine being less or as genotoxic as retronecine, and otonecine being the most genotoxic. In the TensorFlow-generate DL model, platynecine also has the lowest genotoxic probability, but are then followed by the otonecines and last by retronecine. These results partly correspond to earlier published studies. Saturated PAs of the platynecine-type are generally accepted to be less or non-toxic and have been shown in *in vitro* experiments to form no DNA-adducts (Xia et al. 2013). Therefore, it is striking, that 1,2-unsaturated PAs of the retronecine-type should have an almost comparable genotoxic potential in the LAZAR and DL (R-project) model. In literature, otonecine-type PAs were shown to be more toxic than those of the retronecine-type (Li et al. 2013).

Modifications of necine base

The group-specific results of the TensorFlow-generated DL model appear to reflect the expected relationship between the groups: the low genotoxic potential of *N*-oxides and the highest potential of dehydropyrrolizidines (Chen et al. 2010).

In the LAZAR model, the genotoxic potential of dehydropyrrolizidines (DHP) (using the extended AD) is comparable to that of tertiary PAs. Since, DHP is regarded as the toxic principle in the metabolism of PAs, and known to produce protein- and DNA-adducts (Chen et al. 2010), the LAZAR model did not meet this expectation it predicted the majority of DHP as being not genotoxic. However, the following issues need to be considered. On the one hand, all DHP were outside of the stricter AD of 0.5. This indicates that in general, there might be a problem with the AD. In addition, DHP has two unsaturated double bonds in its necine base, making it highly reactive. DHP and other comparable molecules have a very short lifespan, and usually cannot be used in *in vitro* experiments. This might explain the absence of suitable neighbours in LAZAR.

Furthermore, the probabilities for this substance groups needs to be considered, and not only the consolidated prediction. In the LAZAR model, all DHPs had probabilities for both outcomes (genotoxic and not genotoxic) mainly below 30%. Additionally, the probabilities for both outcomes were close together, often within 10% of each other. The fact that for both outcomes, the probabilities were low and close together, indicates a lower confidence in the prediction of the model for DHPs.

In the DL (R-project) and RF model, *N*-oxides have a by far more genotoxic potential than tertiary PAs or dehydropyrrolizidines. As PA *N*-oxides are easily conjugated for excretion, they are generally considered as detoxification products, which are *in vivo* quickly renally

eliminated (Chen et al. 2010). On the other hand, *N*-oxides can be also back-transformed to the corresponding tertiary PA (Wang et al. 2005). Therefore, it may be questioned, whether *N*-oxides themselves are generally less genotoxic than the corresponding tertiary PAs. However, in the groups of modification of the necine base, dehydropyrrolizidine, the toxic principle of PAs, should have had the highest genotoxic potential. Taken together, the predictions of the modifications of the necine base from the LAZAR, RF and R-generated DL model cannot – in contrast to the TensorFlow DL model - be considered as reliable.

Overall, when comparing the prediction results of the PAs to current published knowledge, it can be concluded that the performance of most models was low to moderate. This might be contributed to the following issues:

1. In the LAZAR model, only 26.6% PAs were within the stricter AD. With the extended AD, 92.3% of the PAs could be included in the prediction. Even though the Jaccard distance between the training dataset and the PA dataset for the RF, SVM, and DL (R-project and TensorFlow) models was small, suggesting a high similarity, the LAZAR indicated that PAs have only few local neighbours, which might adversely affect the prediction of the mutagenic potential of PAs.
2. All above-mentioned models were used to predict the mutagenicity of PAs. PAs are generally considered to be genotoxic, and the mode of action is also known. Therefore, the fact that some models predict the majority of PAs as not genotoxic seems contradictory. To understand this result, the basis, the training dataset, has to be considered. The mutagenicity of in the training dataset are based on data of mutagenicity in bacteria. There are some studies, which show mutagenicity of PAs in the AMES test (Chen et al. 2010). Also, Rubiolo et al. (1992) examined several different PAs and several different extracts of PA-containing plants in the AMES test.

They found that the AMES test was indeed able to detect mutagenicity of PAs, but in general, appeared to have a low sensitivity. The pre-incubation phase for metabolic activation of PAs by microsomal enzymes was the sensitivity-limiting step. This could very well mean that this is also reflected in the QSAR models.

4.6 Conclusions

In this study, an attempt was made to predict the genotoxic potential of PAs using five different machine learning techniques (LAZAR, RF, SVM, DL (R-project and TensorFlow)). The results of all models fitted only partly to the findings in literature, with best results obtained with the TensorFlow DL model. Therefore, modelling allows statements on the relative risks of genotoxicity of the different PA groups. Individual predictions for selective PAs appear, however, not reliable on the current basis of the used training dataset.

This study emphasises the importance of critical assessment of predictions by QSAR models. This includes not only extensive literature research to assess the plausibility of the predictions, but also a good knowledge of the metabolism of the test substances and understanding for possible mechanisms of toxicity.

In further studies, additional machine learning techniques or a modified (extended) training dataset should be used for an additional attempt to predict the genotoxic potential of PAs.

5 The hepatotoxic potential of protein kinase inhibitors predicted with Random Forest and Artificial Neural Networks

Authors

Verena Schöning, Stephan Krähenbühl and Jürgen Drewe

Published in⁵:

Toxicology Letters 299 (2018) 145–148, ISI Impact factor 3.858

Corresponding author:

Prof. Dr. Jürgen Drewe, MSc

5.1 Abstract

Protein kinases (PKs) play a role in many pivotal aspects of cellular function. Dysregulation and mutations of protein kinases are involved in the development of different diseases, which might be treated by inhibition of the corresponding kinase. Protein kinase inhibitors (PKIs) are generally well tolerated, but unexpected and serious adverse events on the heart, lung, kidney and liver were observed clinically. In this study, the structure-activity relationship of PKIs in relation to hepatotoxicity was investigated. A dataset of 165 PKIs was compiled and the probability of human hepatotoxicity with two different machine learning algorithms (Random

⁵ This is a pre-copyedited, author-produced version of an article accepted for publication in Toxicology Letters following peer review. The version of record ‘Schöning V, Krähenbühl S, Drewe J. 2018. The hepatotoxic potential of protein kinase inhibitors predicted with Random Forest and Artificial Neural Networks. Tox Let 299, 145–148’ is available online at: <https://doi.org/10.1016/j.toxlet.2018.10.009>. In course of harmonisations for this manuscript, the numbering and sometimes also the allocations of figures, annexes, and supplementary material was amended. Furthermore, terms were harmonised. No other changes were made.

Forest and Artificial Neural Networks) was analysed. The estimated probability of hepatotoxicity was generally high for single PKIs. However, depending on the target kinase of the PKI, a difference in hepatotoxic potential could be observed. The similarity of the PKIs to each other is caused by the conserved site of action of the protein kinases. Hepatotoxicity may therefore always be an issue in PKIs.

5.2 Introduction

Protein kinases (PKs) play a role in many pivotal aspects of cellular function. PKs catalyse the transfer of phosphate groups from ATP (adenosine triphosphate) to specific proteins. Protein phosphorylation can modify the conformation and function of a protein and thus serves as an important regulator in signalling pathways, which impacts gene transcription and protein synthesis, cell metabolism, division and movement as well as programmed cell death. Therefore, dysregulation of protein kinases is associated with the development of different diseases, making this family of enzymes one of the most important drug targets over the past two decades (Roskoski 2015). Protein kinase inhibitors (PKIs) can be classified based on their molecular targets on kinases. Type I and type II inhibitors bind reversibly to the ATP-binding pocket of the protein kinases and exhibit competitive inhibition with respect to ATP. The difference is that the former binds to the active enzyme conformation, whereas the latter binds to the inactive conformation. Type III inhibitors bind to an allosteric pocket adjacent to the ATP binding site but without direct interaction with the ATP-binding pocket (Dar & Shokat 2011). Most of the FDA (Food and Drug Administration, the US health authority)-approved PKIs act as competitive inhibitors at the ATP binding site (type I and II) (Jeon et al. 2017; Roskoski 2015; Yu et al. 2014a), leading to a structurally homogenous group of substances.

Even though PKIs are generally well tolerated, unexpected and serious adverse events on the heart, lung, kidney and liver were observed clinically (Shah et al. 2013). The mechanisms of PKI-mediated hepatotoxicity are only partially elucidated. For some PKIs, an extensive metabolism and bioactivation by cytochrome P450 enzymes is known, resulting in the production of reactive metabolites (Teo et al. 2013). *In vitro* investigations proved that some, but not all, PKIs exhibit a strong mitochondrial toxicity and inhibit glycolysis at clinically relevant concentrations (Mingard et al. 2018; Paech et al. 2017).

Since PKIs have different hepatotoxic potencies despite structural similarities and since the mechanisms of hepatotoxicity are varying, the hepatotoxic potential of individual PKIs is difficult to predict. The aim of the current study was to identify a relationship between the structure of individual PKIs and their hepatotoxic potential using machine learning algorithms. For this reason, a dataset of 165 PKIs (independent from their regulatory status) was compiled and the probability of human hepatotoxicity was analysed with two different machine learning algorithms. The probability of hepatotoxicity was compared with clinical findings for individual PKIs and also matched with the PKI target. In addition, the similarity of the PKIs to each other was investigated.

5.3 Materials and methods

5.3.1 PKI dataset

A dataset of 165 protein kinase inhibitors (PKIs), mainly tyrosine kinase inhibitors (TKIs) and few structurally related compounds (such as proteasome inhibitors), was compiled (see supplementary material S2). A total of 21 specific targets were assigned to the PKIs from DrugBank (www.drugbank.ca, last accessed in February 2018) and Selleckchem (http://www.selleckchem.com/pharmacological_receptor-tyrosine-kinase.html, last accessed

in February 2018). Chemical structures were coded by ‘simplified molecular-input line-entry system’ (SMILES) that were obtained from PubChem (<https://pubchem.ncbi.nlm.nih.gov/>, last accessed in February 2018). For each PKI, molecular 1D and 2D descriptors were calculated using PaDEL-Descriptors (version 2.21) (Yap 2011; 2014). The process of standardization involved removing any salts from SMILES structures, for instance chlorides or lysinate residues. Additionally, explicit hydrogens were removed. This PKI dataset was used as a test set in the study and the probability of hepatotoxicity for each substance in this dataset was predicted using the RF and aNN model described in section 5.3.2.

The probability of the whole dataset was evaluated, as well as the probability with relation to the target of the PKIs. Furthermore, the similarity of the PKIs with each other was calculated using the Jaccard distance.

5.3.2 DILI dataset and model training

All computation steps were performed in R (R Project for Statistical Computing, <https://www.r-project.org/>; version 3.3.1; last accessed September 9, 2017) using additional R packages (packages are identified in Schöning et al. (2017) or in the description below).

Two different QSAR models were used for the calculation of the hepatotoxic probability, Random Forest and Artificial Neural Network (aNN). With minor variations, both models were mainly generated as described in Schöning et al. (2017). In short, based on the human DILI-dataset from Chen et al. (2016), a training dataset was established, containing 453 hepatotoxic and 268 non-hepatotoxic substances. In contrast to Schöning et al. (2017), nine hepatotoxic-classified substances (bortezomib, dasatinib, erlotinib, gefitinib, imatinib, lapatinib, pazopanib, sorafenib, and sunitinib) were excluded from the training dataset, as these substances were also part of the PKI test dataset. This step was done to avoid any bias for these nine and further

structurally related substances in the QSAR models. Therefore, the DILI training dataset consisted of 444 hepatotoxic and 268 non-hepatotoxic substances. PaDEL-Descriptors (version 2.21) (Yap 2011; 2014) was used to calculate 1444 molecular descriptors for each substance. Removal of zero variance descriptors, missing values and highly correlated descriptors were performed analogous to Schöning et al. (2017). Using a *Recursive Feature Elimination* (Zhu et al. 2015), the 100 most important descriptors, the final predictors for the models, were identified. After that, the training dataset was balanced by artificial over-sampling of the non-hepatotoxic class to obtain an almost equal number of substances in the two classes (hepatotoxic and non-hepatotoxic) (Chawla et al. 2002). After this procedure, the final training dataset consisted of 458 hepatotoxic and 455 non-hepatotoxic substances. The RF model, which based on this dataset, used the 100 most important predictors as identified by the *Recursive Feature Elimination*. The forest contained 1000 trees and 75 variables that were randomly sampled at each split. For the aNN model, the training dataset was normalized by calculating the standard deviation for each predictor and then each value was divided by that standard deviation. The aNN model consisted of 3 layers, the input layer with 100 units, the hidden layer with 75 units and a single-unit output layer.

5.3.3 Model validation

Compared to the original model in Schöning et al. (2017), the training dataset was reduced by deleting the nine substances present in the test dataset (see section 2.2). This is equal to a reduction of the training dataset by 1.2%. To confirm the validity of the altered models, it was decided to repeat the 10-fold internal cross-validation of both QSAR-models (Mitchell 2014; Nantasenamat et al. 2009). Based on that, the correct classification rate (CCR) for both models was calculated to measure the predictive power:

$$CCR = \frac{1}{2} \left(\frac{T_N}{N_0} + \frac{T_P}{N_1} \right)$$

T_N and T_P represent the number of true negative and positive predictions, respectively, and N_0 and N_1 the total number of negative and positive compounds in the model, respectively. In addition, the sensitivity, specificity, and area under the receiver characteristic curve (ROC-AUC) were determined.

Additionally, it was confirmed, that the compounds of the training dataset are within the applicability domain, which is defined by the compound of the test dataset. For this purpose, on the one hand, a *principal component analysis* (PCA) was performed (R package ‘stats’, function ‘prcomp’), using the identified, most relevant 100 predictors and the first 4 principal components (PC). On the other hand, the distance, using the Jaccard distance measure, between the DILI dataset and the PKI dataset was calculated.

5.4 Results and discussion

5.4.1 Model validation and predictor importance

For the RF model, the CCR was 90.3%, the sensitivity 90.3%, the specificity 90.2%, and the ROC-AUC 0.96. For the aNN model, the CCR was 79.7%, the sensitivity 78.4%, the specificity 80.9%, and the ROC-AUC 0.87. Even though the training dataset was slightly reduced in comparison to Schöning et al. (2017), the main parameters of the models (CCR, sensitivity, specificity and ROC-AUC) are comparable to the original models.

The applicability domain between the DILI and the PKI dataset was confirmed using a PCA, considering the first 4 PC (PC1-PC4). This analysis shows that the PKI dataset falls within the range of the DILI dataset (Annex 2). Additionally, the average distance for all PKIs to the DILI dataset was calculated as the Jaccard distance. The Jaccard distance is a statistic value used for

comparing the dissimilarity of sample sets, which can vary between 0 (similar) and 1 (not similar). As the Jaccard distance between the PKI and DILI dataset was < 0.13 , both datasets can be regarded as 'close'. Therefore, the PKI dataset falls within the applicability domain of the DILI dataset.

Overall, it can be concluded that both machine learning models, based on the DILI dataset, are valid and may be used for the prediction of the PKI dataset.

Additionally, the 30 most important predictors, as determined by the RF model, were evaluated (see Annex 3). Except for the autocorrelation predictors, which lack intuitive understanding, the probability for hepatotoxicity was mainly determined by the reactivity of the molecule (e.g. number of atoms in the largest pi system) and its lipophilicity (e.g. ALogP).

5.4.2 Overall acute hepatotoxic probability of PKIs

The vast majority of PKIs (93% and 95% in the RF and aNN model, respectively) had a probability of 50% or above to be acutely hepatotoxic. A probability of more than 75% was seen in more than half of the PKIs analysed (57% and 63% in the RF and aNN model, respectively). Values well over 50% indicate a high confidence of the prediction (Breimann 2003). Therefore, the hepatotoxic potential of PKIs in general can be considered as high.

The prediction of hepatotoxic probability of single PKIs was highly correlated between both models ($R^2 = 0.64$, $P < .001$, slope = 0.996, see Figure 21). In addition to the above mentioned validation of the two models (internal cross-validation) and the confirmation of the applicability domain, this correlation provided further evidence for the validity of the analysis.

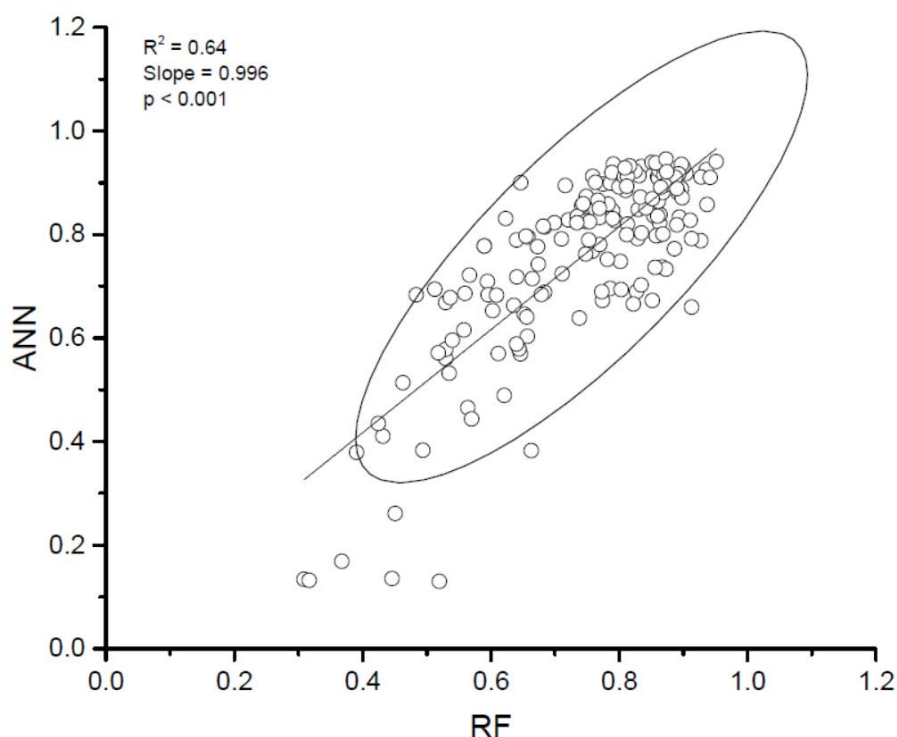


Figure 21: Correlation of the hepatotoxic potential of single PKIs as predicted by the RF and the aNN model
 $R^2 = 0.64$, $P < .001$, slope = 0.996.

Literature research corroborated the high hepatotoxic potential of PKIs. 25-35% of the patients in pre-approval clinical trials of TKIs experienced low-grade increases in ALT (alanine transaminase) and/or AST (aspartate transaminase) ($\leq 5x$ upper limit of the normal range (National Cancer Institute 2006)), whereas high-grade increase in serum transaminase level were observed in 2% of the patients (Shah et al. 2013). Most PKIs currently approved by the FDA are metabolised by cytochrome P450 enzymes and undergo hepatic excretion. Ruxolitinib, a JAK (Janus kinase) inhibitor, is an exception, undergoing mainly renal excretion (Jeong et al. 2013). Even though ruxolitinib is considered as hepatotoxic by both models, the probability is notably lower than for other PKIs (0.53 and 0.56 in the RF and aNN model, respectively). It is known that being a cytochrome P450 substrate increases the hepatotoxic potential of a substance (Yu et al. 2014a). For some PKIs (e.g. dasatinib (RF: 0.87, aNN: 0.73), erlotinib (RF: 0.65, aNN: 0.57), gefitinib (0.83, aNN: 0.69), imatinib (RF: 0.75, aNN: 0.83),

lapatinib (RF/aNN: 0.85), nilotinib (RF: 0.84, aNN: 0.93), pazopanib (RF: 0.81, aNN: 0.91) and sunitinib (RF: 0.59, aNN: 0.78)) bioactivation through CYP P450 enzymes and formation of reactive metabolites was shown (Teo et al. 2015).

The authors believe that the calculated probabilities can be used to assess the hepatotoxic potential of a given substance in the real world and are therefore meaningful. At least, the probabilities can be used to rank the substances regarding the hepatotoxic potential, which is clinically useful.

We compared the probability calculated and the clinical hepatotoxicity of some PKI drugs, for which extensive clinical data are available:

For instance, the prevalence of hepatotoxicity for sunitinib is <1% (Medscape, <https://search.medscape.com>). Pazopanib shows a high prevalence of hepatotoxic adverse effects (ALT level raised (all grades, 53%; grade 3, 10%; grade 4, 2%), Medscape). The corresponding probabilities in our models were estimated to be 0.59/0.78 and 0.80/0.91 for sunitinib and pazopanib, respectively.

The prevalence of gefitinib-related hepatotoxicity of grade ≥ 3 was significantly higher than erlotinib-related hepatotoxicity in a pooled safety analysis of EGFR mutation-positive non-small cell lung cancer trials (Takeda et al. 2015). The probabilities for hepatotoxicity in our models were estimated to be 0.83/0.69 and 0.55/0.59, respectively, showing the same rank order.

5.4.3 Target-specific hepatotoxic probability of PKIs

Even though the overall probability of hepatotoxicity was very high in both models, some differences could be observed with relation to the targets of the PKIs (Figure 22). The median

probability for PKIs for Janus kinases JAK1, JAK2, JAK3, and Tyk2 were in both models between 0.60-0.67, whereas the probability of PKIs inhibiting other targets was over 0.7. The protein kinase families with the highest median probability were EGFR (ErbB1/2; 0.79-0.82), BCR/ABL (0.81-0.85) and VEGFR (VEGFR1/2/3; 0.78-0.84, see Figure 22) inhibitors. This is an interesting observation, which may be related to the physico-chemical properties of a drug necessary to interact with these targets. Whereas the current study is able to predict the hepatotoxic potential of a drug or of a group of drugs, it cannot answer the question why such associations exist.

5.4.4 Similarity of PKIs

The average distance between the PKIs investigated (calculated as Jaccard distance) was below 0.08, except for two compounds. The two PKIs with a higher distance were FLLL32 and daphnetin (Jaccard distance of 0.13 and 0.11, respectively). Considering these results, PKIs can be considered as a quite homogenous group of substances, which can be explained by the similarity of the targets. In general, protein kinases have a small amino-terminal lobe (N-lobe) and a large carboxy-terminal lobe (C-lobe). Between the two lobes, a cleft is formed that serves as docking site for ATP (Meharena et al. 2013). As almost all PKIs interact with the ATP-binding site with either the active (type I inhibitors) or inactive (type II inhibitors) form of the protein kinases (Wu et al. 2015), possible chemical structures of PKIs are restricted. This explains the observed high similarity of PKIs.

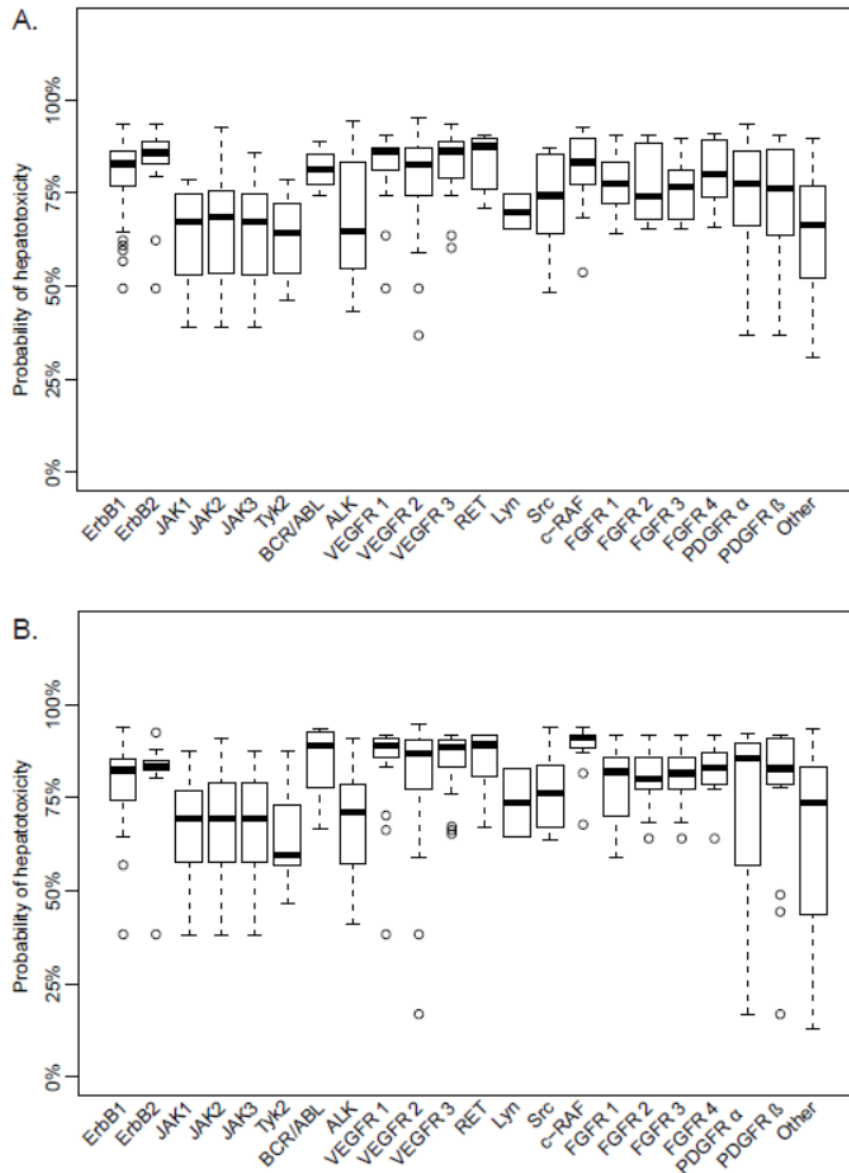


Figure 22: Hepatotoxic probability of PKIs in relation to their target.

A. Prediction results of RF model.

B. Prediction results of aNN model.

ErbB1: Epidermal Growth Factor Receptor 1, ErbB2: Epidermal Growth Factor Receptor 1, JAK1: Janus Kinase 1, JAK2: Janus Kinase 2, JAK3: Janus Kinase 3, Tyk2: Tyrosine Kinase 2, BCR/ABL: Bcr-Abl tyrosine-kinase, ALK: Anaplastic lymphoma kinase, VEGFR 1: Vascular Endothelial Growth Factor Receptor 1, VEGFR 2: Vascular Endothelial Growth Factor Receptor 2, VEGFR 3: Vascular Endothelial Growth Factor Receptor 3, RET: RET proto-oncogene, Lyn: Tyrosine-protein kinase Lyn, Src: Tyrosine-protein kinase Src, c-RAF: RAF proto-oncogene serine/threonine-protein kinase, FGFR 1: Fibroblast Growth Factor Receptor 1, FGFR 2: Fibroblast Growth Factor Receptor 2, FGFR 3: Fibroblast Growth Factor Receptor 3, FGFR 4: Fibroblast Growth Factor Receptor 4, PDGFR α : Platelet-Derived Growth Factor Receptor α , PDGFR β : Platelet-Derived Growth Factor Receptor β , Other: other targets, not belonging to any of the afore mentioned. In the boxplot, the median is indicated by a horizontal line, the bottom and top of the box are the 25th (P25%) and 75th (P75%) percentile, the whiskers are the P75% or P25% plus or minus 1.5*Interquartile Range (IQR) respectively. Outliers are indicated as open circles.

5.4.5 Limitation of the study

PKIs are usually used in patients that have high co-morbidity, e.g. patients with non-small cell lung cancer, renal cell carcinoma, and chronic myeloid leukaemia (Jeon et al. 2017), which also includes liver disease. Therefore, a correlation of the predicted hepatotoxicity with clinically observed hepatotoxic events in patients is difficult. The hepatotoxic mode of action for PKIs is still under investigation. For some PKIs, mitochondrial toxicity and inhibition of glycolysis are discussed as possible pathways (Mingard et al. 2018; Paech et al. 2017). The current study is not able to provide a mechanism of hepatotoxicity for these compounds.

5.5 Conclusion

Almost all of the known PKIs today have a high hepatotoxic probability in both prediction models. The clinicians should be aware of potential hepatotoxic effects of PKIs. Hepatotoxic events observed in patients treated with PKIs should be critically evaluated with regard to a causal relationship with drug therapy using validated tools like the RUCAM score (Danan & Benichou 1993; Regev et al. 2014).

6 Development of an *in vitro* screening method of acute cytotoxicity of the pyrrolizidine alkaloid lasiocarpine in human and rodent hepatic cell lines by increasing susceptibility

Authors

Kristina Forsch, Verena Schöning, Lucia Disch, Beate Siewert, Matthias Unger, Jürgen Drewe

Published in⁶:

Journal of Ethnopharmacology 217 (2018) 134–139, ISI Impact factor 3.115

Corresponding author:

Prof. Dr. Jürgen Drewe, MSc

6.1 Abstract

Ethnopharmacological relevance

Pyrrolizidine alkaloids (PAs) are secondary plant ingredients formed in many plant species to protect against predators. PAs are generally considered acutely hepatotoxic, genotoxic and

⁶ This is a pre-copyedited, author-produced version of an article accepted for publication in Journal of Ethnopharmacology following peer review. The version of record ‘Forsch K, Schöning V, Disch L, Siewert B, Unger M Drewe J. 2018. Development of an *in vitro* screening method of acute cytotoxicity of the pyrrolizidine alkaloid lasiocarpine in human and rodent hepatic cell lines by increasing susceptibility. J Ethnopharmacol 217, 134–139’ is available online at: <https://doi.org/10.1016/j.jep.2018.02.018>. In course of harmonisations for this manuscript, the numbering and sometimes also the allocations of figures, annexes, and supplementary material was amended. Furthermore, terms were harmonised. No other changes were made.

carcinogenic. Up to now, only few *in vitro* and *in vivo* investigations were performed to evaluate their relative toxic potential.

Aim of the study

The aim was to develop an *in vitro* screening method of their cytotoxicity.

Materials and Methods

Human and rodent hepatocyte cell lines (HepG2 and H-4-II-E) were used to assess cytotoxicity of the PA lasiocarpine. At concentrations of 25 μM up to even 2400 μM , no toxic effects in neither cell line was observed with standard cell culture media. Therefore, different approaches were investigated to enhance the susceptibility of cells to PA toxicity (using high-glucose or galactose-based media, induction of toxifying cytochromes, inhibition of metabolic carboxylesterase, and inhibition of glutathione-mediated detoxification).

Results

Galactose-based culture medium (11.1 mM) increased cell susceptibility in both cell-lines. Cytochrome P450-induction by rifampicin showed no effect. Inhibition of carboxylesterase-mediated PA detoxification by specific carboxylesterase 2 inhibitor loperamide (2.5 μM) enhanced lasiocarpine toxicity, whereas the unspecific carboxylesterase inhibitor bis(4-nitrophenyl)phosphate (BNPP, 100 μM) had a weaker effect. Finally, the inhibition of glutathione-mediated detoxification by buthionine sulphoximine (BSO, 100 μM) strongly enhanced lasiocarpine toxicity in H-4-II-E cells in low and medium, but not in high concentrations.

Conclusions

If no toxicity is observed under standard conditions, susceptibility enhancement by using galactose-based media, loperamide, and BSO may be useful to assess relative acute cytotoxicity of PAs in different cell lines.

6.2 Introduction

Pyrrolizidine alkaloids (PAs) are secondary plant ingredients formed in many plant species to protect against predators (Hartmann & Witte 1995; Langel et al. 2011). PAs, ester alkaloids composed of a necine base (two fused five-membered rings joined by a single nitrogen atom) and a necic acid (one or two carboxylic ester arms) are generally considered acutely hepatotoxic, genotoxic and carcinogenic.

The main organ of PA metabolism and target of toxicological effects is the liver (Bull & Dick 1959; Bull et al. 1958; Butler et al. 1970; DeLeve et al. 1996; Jago 1971; Li et al. 2011; Neumann et al. 2015). Three principal metabolic pathways for 1,2-unsaturated PAs of the retronecine-type are known (Chen et al. 2010): (i) *Detoxification*: Hydrolysis of the C7/ C9 ester bond by non-specific esterases to release necine base and necic acid. These intermediates are then subjected to further phase II-conjugation and excretion. (ii) *Detoxification*: *N*-oxidation of the necine base to form PA *N*-oxides. (iii) *Metabolic activation/ toxification*: Oxidation and/or oxidative *N*-demethylation, resulting after cleaving the ester bond(s) by esterases in the formation of reactive pyrroles (also known as dehydropyrrolizidine or pyrrolic ester) (Figure 23). This pathway is mainly catalyzed by hepatic cytochrome P450 (CYP) isoforms CYP2B and 3A (Ruan et al. 2014b). Reactive pyrroles cause damage in the cells in which they are formed, usually hepatocytes, but can pass from the hepatocytes into the adjacent sinusoids and damage the endothelial lining (Gao et al. 2015) mainly by reaction with DNA,

protein, and lipids. Due to the ability of 1,2-unsaturated PAs to form DNA adducts, DNA crosslinks and DNA breaks, they are generally considered genotoxic and carcinogenic (Chen et al. 2010; EFSA 2011; Fu et al. 2004; Li et al. 2011; Takanashi et al. 1980; Yan et al. 2008; Zhao et al. 2012).

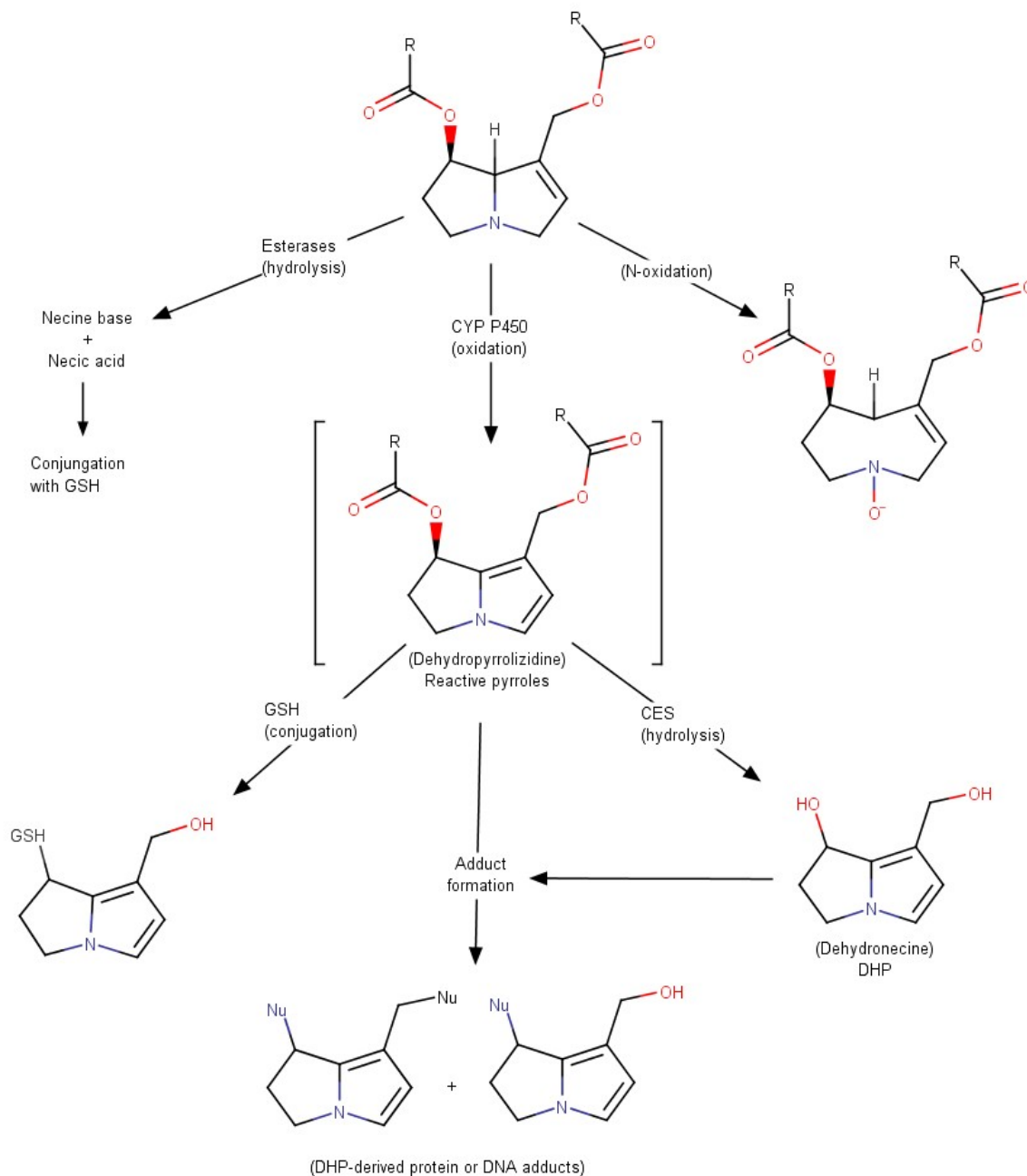


Figure 23: Metabolic pathways of retronecine-type PAs

CES: carboxylesterase, GSH: glutathione; Nu =nucleophilic targets, adapted according to (Chen et al. 2010)

After acute intoxication of humans, the most common lesions in the liver are hemorrhagic necrosis, lesions in the central and sublobular veins of the liver, and acute venoocclusive disease (DeLeve et al. 2003; EFSA 2011).

However, up until to now, only few *in vitro* and *in vivo* investigations were performed to evaluate the relative toxic potential of PAs (Field et al. 2015; Li et al. 2011; Tamta et al. 2012). Especially in *in vitro* studies, the susceptibility of different cells lines to acute toxic effects of PAs was low (Field et al. 2015), which further complicates those kind of studies. This could be due to the fact that some cell lines switch their metabolism according to the so-called Crabtree effect (Crabtree 1928). This effect describes that in presence of low-glucose concentrations cells in culture derive all their energy from anaerobic glycolysis rather than *via* mitochondrial oxidative phosphorylation (OXPHOS) despite of aerobic conditions. This leads to a high resistance against mitochondrial toxins (Marroquin et al. 2007). Mitochondrial toxicity, which may be related to the acute toxicity of PAs, was shown among others by for the PAs clivorine and senecionine (Ji et al. 2008), retrorsine (Gordon et al. 2000), lasiocarpine (Armstrong et al. 1972) and dehydromonocrotaline (Mingatto et al. 2007).

On this account the aim of this study was to develop a suitable screening system by reducing the threshold of susceptibility to toxic effects. The acute toxic effects of PAs were then studied in those sensitized cells. This included a general approach of modification of the cell culture medium, and specific alterations to the activity of enzymes, which are involved in the metabolism of PAs.

6.3 Materials and Methods

6.3.1 Chemical and reagents

Lasiocarpine, was obtained from Phytolab (Vestenbergsgreuth, Germany).

The specific carboxylesterase (CES) 2 inhibitor loperamide hydrochloride, the unspecific CES inhibitor bis(4-nitrophenyl)phosphate (BNPP), DL-buthionine sulphoximine (BSO) and rifampicin were purchased from Sigma Aldrich (St. Louis, Missouri, USA) in the highest grade available. Media MEM-Glutamax media, DMEM-Glutamax media, sodium pyruvate, MEM non-essential amino acids, penicillin/streptomycin, L-glutamine, 2-(4-(2-hydroxyethyl)-1-piperazinyl)-ethansulphoic acid (HEPES) and fetal bovine serum (FBS) were obtained from Gibco (Carlsbad, Californian, USA). The positive control digitonin was purchased from Sigma Aldrich (St. Louis, Missouri, USA). WST-1 kit was purchased from BioVision (Milpitas, California, USA).

6.3.2 Cells

The human hepatocellular carcinoma (HepG2) and the rat hepatocellular carcinoma (H-4-II-E) cell lines, purchased from ATCC (LGC Standards, Middlesex TW11 0LY, UK) were maintained in 75 cm² culture flask (Semadeni, Ostermundigen, CH) as adherent cell lines in low-glucose MEM-Glutamax (5.5 mM D-glucose) media with 10% (V/V) FBS, 0.5 mM sodium pyruvate, MEM non-essential amino acids and 1% (V/V) penicillin/ streptomycin. The high-glucose DMEM-Glutamax media (25 mM D-glucose) consisted of 10% (V/V) FBS, 0.5 mM sodium pyruvate, and MEM non-essential amino acids. The galactose-based DMEM-Glutamax media (without D-glucose) consisted of 10% (V/V) FBS, 0.5 mM sodium pyruvate, MEM non-essential amino acids, 11.1 mM galactose, 1% (V/V) L-glutamine and 0.5% (V/V)

HEPES. The values of glucose or galactose were defined as the glucose or galactose concentration in the media, not in the cells.

Galactose-based DMEM-Glutamax media were used to prevent the energy production *via* glycolysis in the cultivated cells (Crabtree effect (Crabtree 1928)).

6.3.3 Treatment conditions

6.3.3.1 PA toxicity without pre-treatment

H-4-II-E and HepG2 cells were differentiated for 72 h and then the cells were incubated in low-glucose (5.5 mM) media with lasiocarpine up to 2400 μ M every 24 h for further three days. This experiment was entirely performed in low-glucose (5.5 mM) media.

6.3.3.2 Induction of cytochromes

H-4-II-E and HepG2 cells were differentiated for 72 h. Then the cells were induced every 24 h for three days with rifampicin (25 μ M) to increase cytochrome expression and activity. Afterwards, the cells were additionally incubated with lasiocarpine up to 900 μ M and rifampicin every 24 h for further three days. This experiment was entirely performed in low-glucose (5.5 mM) media.

6.3.3.3 Change in media

To increase susceptibility, H-4-II-E cells were cultured in two different media: (1) Cells were cultivated and differentiated for 24 h and treated with lasiocarpine every 24 h for three days in high-glucose-based media (25 mM D-glucose). (2) Cells were cultivated and differentiated over 24 h in high-glucose (25 mM D-glucose) based media and then switched to galactose-

based medium (11.1 mM) 24 h prior to and during the lasiocarpine treatment. Cells were treated with lasiocarpine every 24 h for three days.

6.3.3.4 Inhibition of detoxification pathways

6.3.3.4.1 Inhibition of carboxylesterases

To increase susceptibility in H-4-II-E and HepG2 cells, carboxylesterases (CES), which are involved in the detoxification of PAs, were inhibited. Cells were cultivated and differentiated for 24 h in high-glucose-based media (25 mM D-glucose), and then switched to galactose-based medium (11.1 mM) 24 h prior to treatment. Cells were treated every 24 h for three days with (1) lasiocarpine up to 900 μ M and the unspecific CES inhibitor BNPP (100 μ M) or (2) lasiocarpine up to 900 μ M and the the specific CES-2 inhibitor loperamide (2.5 μ M).

6.3.3.4.2 Inhibition of GSH formation

The detoxification of lasiocarpine was inhibited by reducing the glutathione synthesis with BSO), an inhibitor of γ -glutamylcysteine synthetase (γ -GCS). BSO lowers tissue glutathione (GSH) concentrations. Cells were cultivated and differentiated for 24 h in high-glucose-based media (25 mM D-glucose), and then switched to galactose-based medium (11.1 mM) 24 h prior to treatment. Cells were treated every 24 h for three days with lasiocarpine up to 900 μ M and BSO (100 μ M).

6.3.4 WST-1 assay

The WST-1 test was used to measure the metabolic activity in the two cell lines (H-4-II-E and HepG2) as a marker for cellular toxicity. The toxicity was defined as decrease of metabolic activity of $\geq 20\%$. At the end of the experiment, 10 μ L WST-1 reagent was added to each well. Plates were then incubated for 4 h to allow for the reduction of WST-1 reagent. Absorbance

was measured by the microplate absorbance reader (Infinite M 200, Tecan Trading Ltd., Männedorf, CH) at 450 nm, reference wavelength of 620 nm (n = 3).

The validity of the WST-1 assay was verified using digitonin (100 μ M) as positive control and the metabolic activity was reduced to 0.9 - 12.5%.

6.4 Results

6.4.1 Susceptibility of cells to PAs without pre-treatment

The susceptibility of immortalized cell lines to toxic effect of lasiocarpine was examined in H-4-II-E and HepG2 cells after 72 h incubation without any pre-treatment in low-glucose media. Lasiocarpine was applied in doses of 25 μ M to up to excessive concentrations of 2400 μ M. The toxic effect of lasiocarpine was measured as decrease in metabolic activity evaluated by WST-1 assay. At the highest concentration, the metabolic activity decreased in both cell lines. The effect was more pronounced in H-4-II-E than in HepG2 cells (Figure 24).

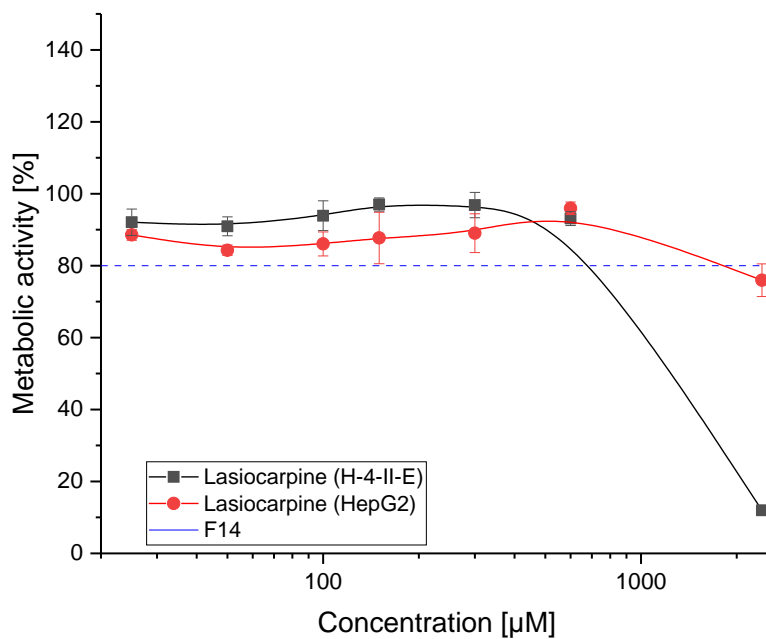


Figure 24: Results of WST-I assay in H-4-II-E and HepG2 cells

Only at the highest concentration of 2400 µM a toxic decrease in metabolic activity ($\geq 20\%$) was observed (mean \pm SEM; n=3).

6.4.2 Enhancement of susceptibility by induction of metabolic activation (rifampicin)

The induction of cytochromes did not increase the susceptibility of H-4-II-E and HepG2 cells to lasiocarpine toxicity (Figure 25).

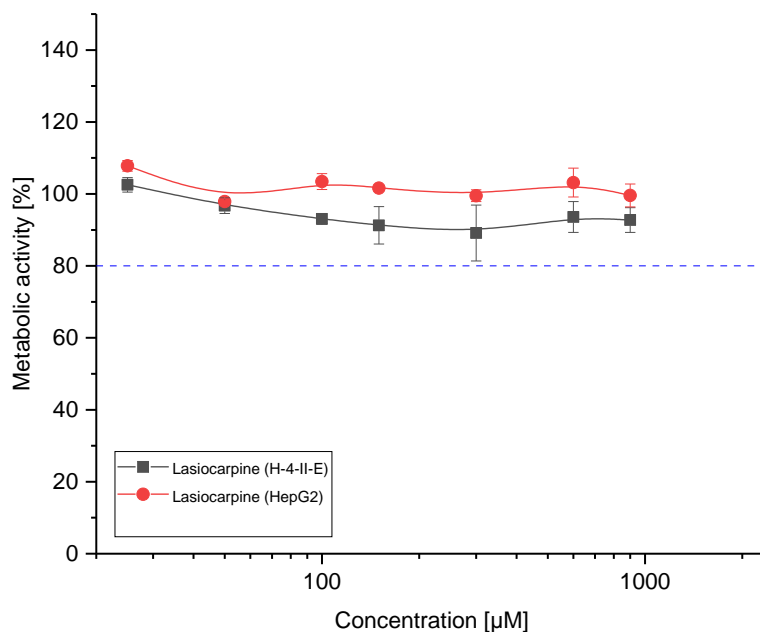


Figure 25: Results of WST-I assay in H-4-II-E und HepG2 cells
 No toxic effects of lasiocarpine after pre-incubation with rifampicin (25 µM) over 72 h. No decrease in metabolic activity after 72 h incubation with lasiocarpine was observed in both cell lines (means ± SEM; n=3).

6.4.3 Enhancement of susceptibility by changes in the medium (high-glucose *versus* galactose)

As H-4-II-E cells were more susceptible to lasiocarpine toxicity, the influence of the media was investigated in this cell line. A toxic effect as decrease in metabolic activity is observed in galactose and the high-glucose approach at lasiocarpine concentrations of 600 µM. The effect was more pronounced in cells treated in galactose media than in high-glucose media (Figure 26). However, high-glucose media did also increase the susceptibility to lasiocarpine toxicity of H-4-II-E cells compared to the first experiment in low-glucose media (5.5 mM D-glucose).

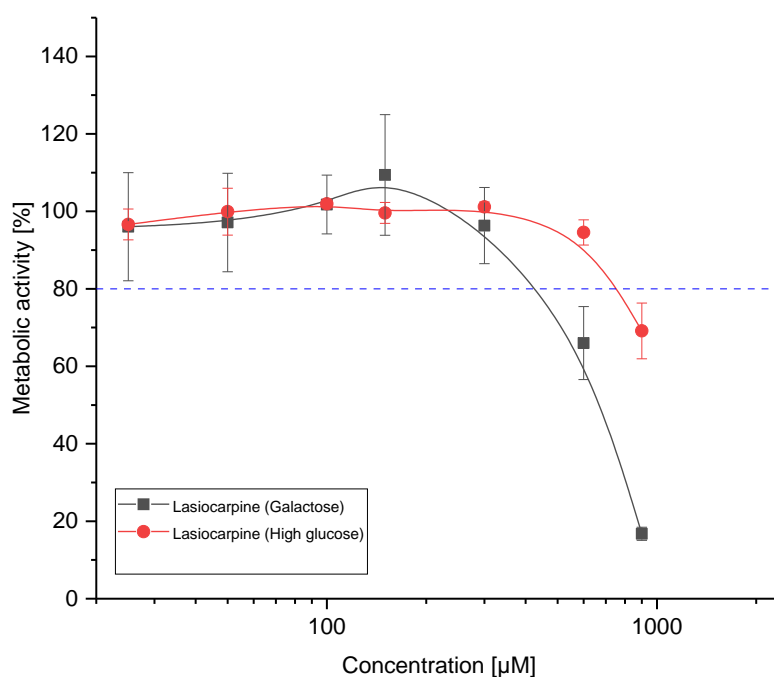


Figure 26: Results of WST-1 assay in H-4-II-E cells
Treatment with lasiocarpine (25 to 900 µM) in high-glucose (25 mM) compared to galactose (11.1 mM) medium (means ± SEM; n = 3).

6.4.4 Enhancement of susceptibility by inhibition of detoxification (carboxylesterases and glutathione formation)

6.4.4.1 Inhibition of carboxylesterases (CES)

A further approach to increase susceptibility of immortalized cells to toxic effects of PAs is to increase the number of reactive pyrroles by inhibition of their detoxification pathways. Treatment with both CES-inhibitors led to a decrease in the metabolic activity. At the two highest concentrations of lasiocarpine (600 µM and 900 µM) with loperamide, a reduction in metabolic activity down to 59 and 38% in H-4-II-E cells, and 66 and 49% in HepG2 cells, respectively, was observed (Figure 27 and Figure 28).

Furthermore, treatment with lasiocarpine in combination with the unspecific CES inhibitor BNPP also reduced, but less effective, metabolic activity. At the highest concentration of

900 μM , lasiocarpine reduced metabolic activity down to 56 and 73% in H-4-II-E and HepG2 cells, respectively.

Loperamide alone had no effect on the metabolic activity (H-4-II-E: 120.4% and HepG2: 91.0%) of both cell lines.

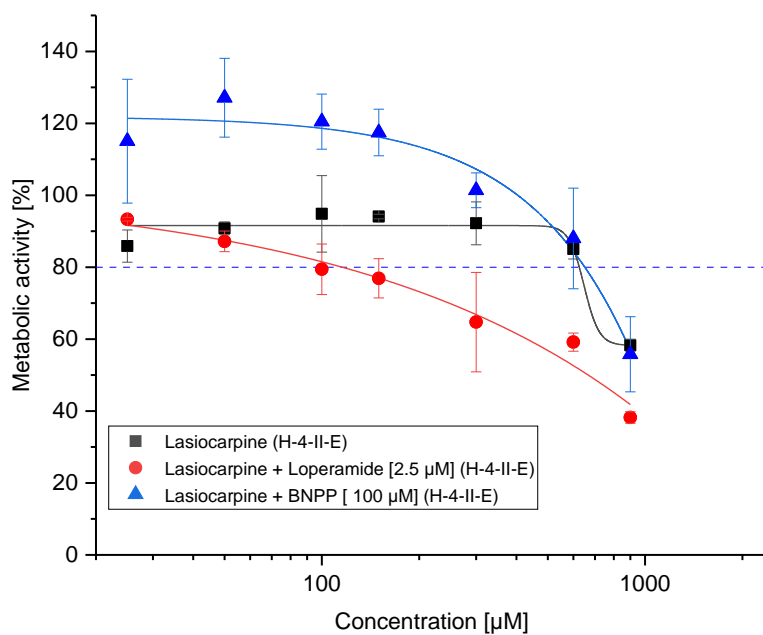


Figure 27: Results of WST-I assay in H-4-II-E cells
Treatment with lasiocarpine (25 to 900 μM) only and in combination with two carboxylesterase inhibitors loperamide (2.5 μM) and BNPP (100 μM).

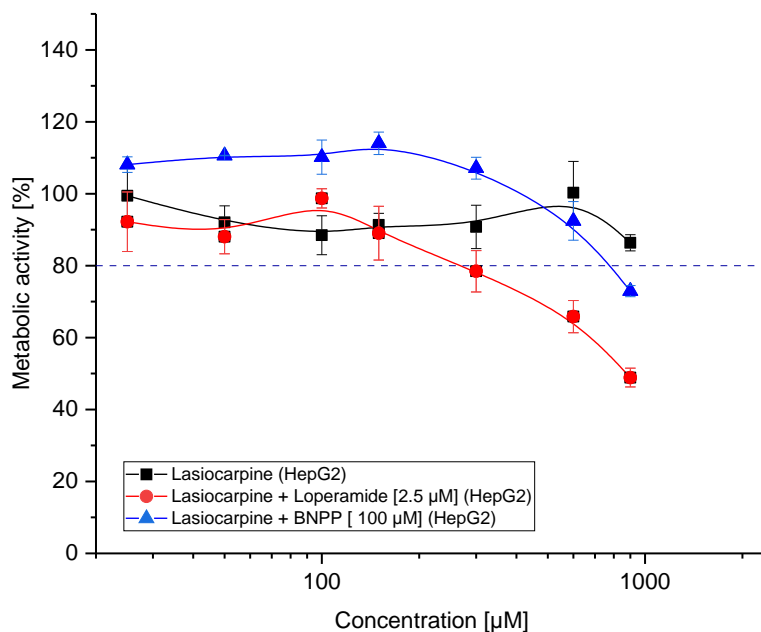


Figure 28: Results of WST-I assay in HepG2 cells
 Treatment with lasiocarpine (25 to 900 µM) only and in combination with two carboxylesterase inhibitors loperamide (2.5 µM) and BNPP (100 µM).

6.4.4.2 Inhibition of glutathione formation

Inhibition of glutathione synthesis with BSO (100 µM) revealed a strong decrease in the metabolic activity of H-4-II-E at the low and medium concentrations of lasiocarpine (50-300 µM). However, at higher concentrations (600 µM and 900 µM), the metabolic activity increased to 126% (Figure 29). The visual control of the cells (data not shown) also confirmed this result: at low and medium concentrations, the cell density was lower and gaps in the monolayer were visible; at the highest concentration, the cell density was comparable with solvent control and the monolayer was confluent.

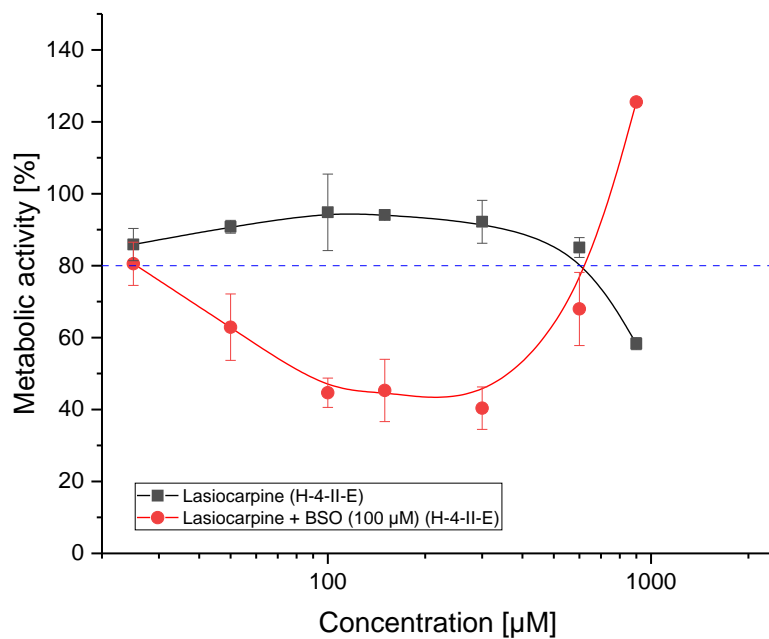


Figure 29: Results of WST-I assay in H-4-II-E cells
Treatment with lasiocarpine only and in combination with glutathione synthesis inhibitor BSO
(means \pm SEM; n=4-6).

In contrast, the inhibition of glutathione in HepG2 cells did not show any effect on the metabolic activity (Figure 30).

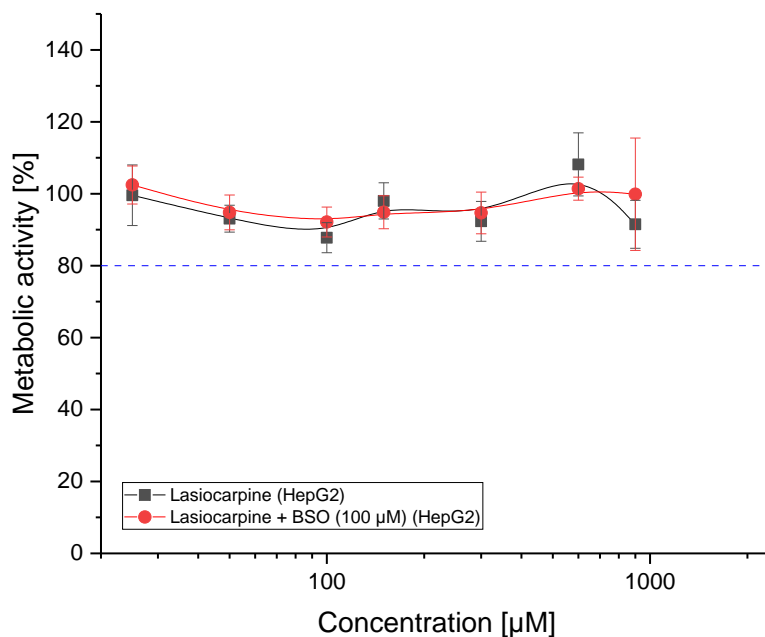


Figure 30: Results of WST-I assay in HepG2 cells
 Treatment with lasiocarpine only and in combination with glutathione synthesis inhibitor BSO
 (means \pm SEM; n=4-6).

6.5 Discussion

The use of primary human hepatocytes is to date the standard *in vitro* model to investigate cytotoxicity. But these cells are very cost-intensive and difficult to handle. Human and rodent hepatic immortalized cell lines are alternative systems, which are easier in handling and cost-effective. The human hepatic cell line HepG2 was used for many investigations of cytotoxic compounds including PA toxicity (Li et al. 2013; Tamta et al. 2012). These studies revealed that for a suitable and well-established *in vitro* screening system for PA toxicity, the susceptibility of the cells needs to be increased.

Different alterations, e.g. by induction or inhibition of metabolic pathways in immortalized human HepG2 and rodent H-4-II-E cells, were evaluated with regard to their influence on the cytotoxicity of the PA lasiocarpine using WST-1 assay. The general assumption implies that

cytotoxicity of PAs primary depends on their metabolic activation by CYP enzymes to form reactive pyrroles, which leads to covalent adduct formation with cellular nucleophiles (Li et al. 2013). In the first experiment, we examined the susceptibility of cells to lasiocarpine toxicity up to excessively high concentrations of 2400 μM without any pre-treatment. The results revealed a decrease in metabolic activity compared to the solvent in both cell lines at the highest concentration only. This effect was more pronounced in H-4-II-E cells than in HepG2 cells. This result suggests that H-4-II-E cells may have a higher metabolic activity than HepG2 cells, which leads to a higher metabolic toxification of toxic pyrroles and thus increased susceptibility. However, both cell lines can be considered as resistant to lasiocarpine toxicity.

Induction of CYP enzyme activity with rifampicin prior to treatment did not increase susceptibility for lasiocarpine toxicity in both cell lines. This could be due to two different reasons: (1) rifampicin: this substance is known to also induce phase-II enzymes (Doostdar et al. 1993; Westerink & Schoonen 2007), which would increase the detoxification of the reactive pyrroles. (2) the Crabtree effect (Aguer et al. 2011; Marroquin et al. 2007): in this case, although cultured under aerobic conditions, cell lines with low supply of glucose metabolically rely on anaerobic glycolysis rather than mitochondrial OXPHOS to generate the required energy. This phenomenon increases the resistance to toxic effects of many mitochondrial function impairing drugs.

Therefore, in the third experiment, we successfully tried to circumvent the Crabtree effect with two different approaches in H-4-II-E cells: (1) Switching the cells from a high-glucose media (cultivation) to a glucose-free, galactose-based media during treatment (Iyer et al. 2010). (2) Cultivation and treatment in high-glucose media, as high concentrations of glucose inhibit the hexokinase (which is an important enzyme in the glycolytic pathway) (Marin-Hernandez et al. 2006). Toxic effects were already seen at moderate lasiocarpine concentrations of 600 μM in

both approaches, but the galactose-based media resulted in a higher increase in the susceptibility to PAs.

A further approach of our study was the inhibition of the detoxification pathways including hydrolysis by carboxylesterases and glutathione conjugation of reactive pyrroles. Our present findings demonstrated that the inhibition of carboxylesterases by loperamide and BNPP leads to a significant decrease in metabolic activity in H-4-II-E cells and HepG2 cells, with the specific inhibitor loperamide being more effective. Therefore, carboxylesterase activity has a strong impact on the detoxification of PAs and consequently on the susceptibility of the cells.

The inhibition of glutathione synthesis by BSO in H-4-II-E cells resulted in a strong decrease in metabolic activity at low and medium concentrations and an increase at the two highest concentrations. Cytotoxicity is a complex interplay of several physiological mechanisms. One possible explanation for this observed hormetic effect may be the induction of phase-II enzymes, e.g. UDP-glucuronosyltransferase, by oxidative stress at high concentrations (Kalthoff et al. 2010), which would then increase the detoxification of the reactive pyrroles. For the PA senecionine it was shown, that UDP-glucuronosyltransferase 1A4 is involved in its detoxification (Galeotti et al. 2010). However, the investigation of the exact reason for this phenomenon is out of scope of this study and will be investigated separately.

In HepG2 cells, inhibition of glutathione synthesis did not lead to a change in metabolic activity. This is due to the lower CYP enzyme activity in this cell line, and therefore a lower toxification of lasiocarpine to reactive pyrroles, and lower toxic effects (Kalthoff et al. 2010; Westerink & Schoonen 2007).

6.6 Conclusions

In the present *in vitro* study, the susceptibility to lasiocarpine in HepG2 and H-4-II-E cells under different conditions was investigated. Inhibition of glycolysis by treating the cells in galactose-based media and inhibition of carboxylesterases increased the susceptibility of both cell lines, whereas the inhibition of GSH was only suitable in H-4-II-E cells. Especially, the former two approaches provide a useful method to perform *in vitro* screening of PA toxicity in immortalized cell lines.

Furthermore, this study emphasizes the necessity to proof the suitability and susceptibility of the *in vitro* test system before assessing compound toxicity.

7 Overall discussion

Predictive models based on machine learning methods are a very useful tool for toxicological investigations. They provide a fast, cost-efficient way to predict specific pharmacological and toxicological endpoints compared to *in vitro* and *in vivo* approaches. Especially in the field of drug development, they can be utilised for large scaled screening of potential candidates (virtual screening (VS)). Furthermore, the predictions can provide evidence to guide further mechanistic investigations. It is also useful in providing insights in toxicological properties and relationships of large substance groups. As an example, we investigated the substance group of pyrrolizidine alkaloids (PAs). Currently, over 600 different PA structures are known. They are secondary metabolites of some plant families, common contaminants of different food products (e.g. honey and herbal teas) and therefore part of the human food chain. More importantly, PAs are also present in herbal medicinal products, either as contaminant or as natural constituent.

As humans are exposed to PAs on a regular basis, different risk assessment were done by various authorities (EFSA 2011; EMA 2014; 2016). However, all of these risk assessments applied the most conservative approach and identified the most toxic PA (lasiocarpine) from *in vivo* investigations in the most sensitive animal species (rat) (NTP 1978) and used that as starting point for all PAs. However, especially for PAs, this might massively overestimate the potential risk in humans. Different PAs exhibit very different toxicological potencies based on their structure. Furthermore, the sensitivity to PA toxicity is highly species-specific. Setting very low limits for PAs exposure, as proposed by the EMA (EMA 2014; 2016) for herbal medicinal products, would drastically reduce the number of herbal medicinal products on the market and thus unnecessarily reduce treatment options for patients. Therefore, it is important

to assess the toxicity of PAs more differentiated and establish a toxicological rank order of single PAs and PA groups (based on their structure). Unfortunately, it is not feasible nor practically achievable to investigate every single PA on their toxicological potential *in vitro* or *in vivo*. From the over 600 known PAs, currently, only about 30 different PAs are commercially available, limiting *in vitro* and *in vivo* testing to those substances. This gap can be overcome by applying machine learning methods. In this work, two different machine learning methods were successfully applied to predict the hepatotoxic potential of over 600 PAs. The prediction of the hepatotoxic potential fitted very well with the known metabolism of PAs and already published literature. This means also that the conclusion drawn of the structural dependent toxicity from testing of the commercially available PAs can be transferred to the whole substance group. Furthermore, due to the high number of investigated PAs, the toxicological share of necine base and necic acid could be assessed, which was not possible before due to the limited PAs tested. The confirmatory *in vitro* study, performed with the PA lasiocarpine, clearly showed that PA toxicity depends in large parts on the animal origin of cell system and the overall experimental conditions. Taking the *in silico* and the *in vitro* study together, it can be concluded that it is not possible to assign one toxicological threshold to all PAs, as the toxicity of PAs depends on different experimental conditions, animal species and structural features. Considering this, the establishment of a Relative Potency Factor (RFP), as proposed by Merz and Schrenk (2016) for PAs seems to be justified.

A further successful application of machine learning techniques provided the study on acute hepatotoxic potential of protein kinase inhibitors (PKIs). The predicted probabilities of PKIs when compared with published data showed a positive correlation. Therefore, the obtained predicted probabilities can be used to rank the substances regarding their hepatotoxic potential.

The *in silico* study, which focussed on the mutagenicity of PAs, however, highlighted constrains of machine learning. Even though a significant modelling, compared to a by chance prediction, could be obtained, interpretation of the data was difficult due to the low performance of the used models. Training of predictive models is not always straightforward and the performance of the models might be low. Using the same machine learning algorithm in different applications (e.g. R-project and TensorFlow), allowing different ways to training the model, can make a significant difference in the model performance. This emphasis even more the importance of model validation and performance assessment, not only accuracy, but also sensitivity and specificity. Without this knowledge, the results could be interpreted overly confidently. Another issue, which has to be address before interpretation of the results, is the applicability domain. The testing dataset has to be within the applicability domain defined by the training dataset. This also needs to be tested and confirmed. Even when these points are considered, the results need to be subjected to critical assessment and review of plausibility. Machine learning cannot be used for confirmatory studies, but are useful in hypothesis generation and risk assessment. When confirmation with literature is not possible, own *in vitro* and *in vivo* studies need to be performed.

8 Overall conclusion

Machine learning is a useful tool in the evaluation of drug-induced toxicity. It is cost- and time-efficient way to study pharmacological and toxicological endpoints compared to *in vitro* and *in vivo* testing. It is especially suitable for large-scaled screening of substance groups and identification of potential candidates for further testing. It is also able to reveal relationships between structural features and pharmacological properties. This helps to deduce relative potency of substances within a substance group to each other. However, while *in silico*

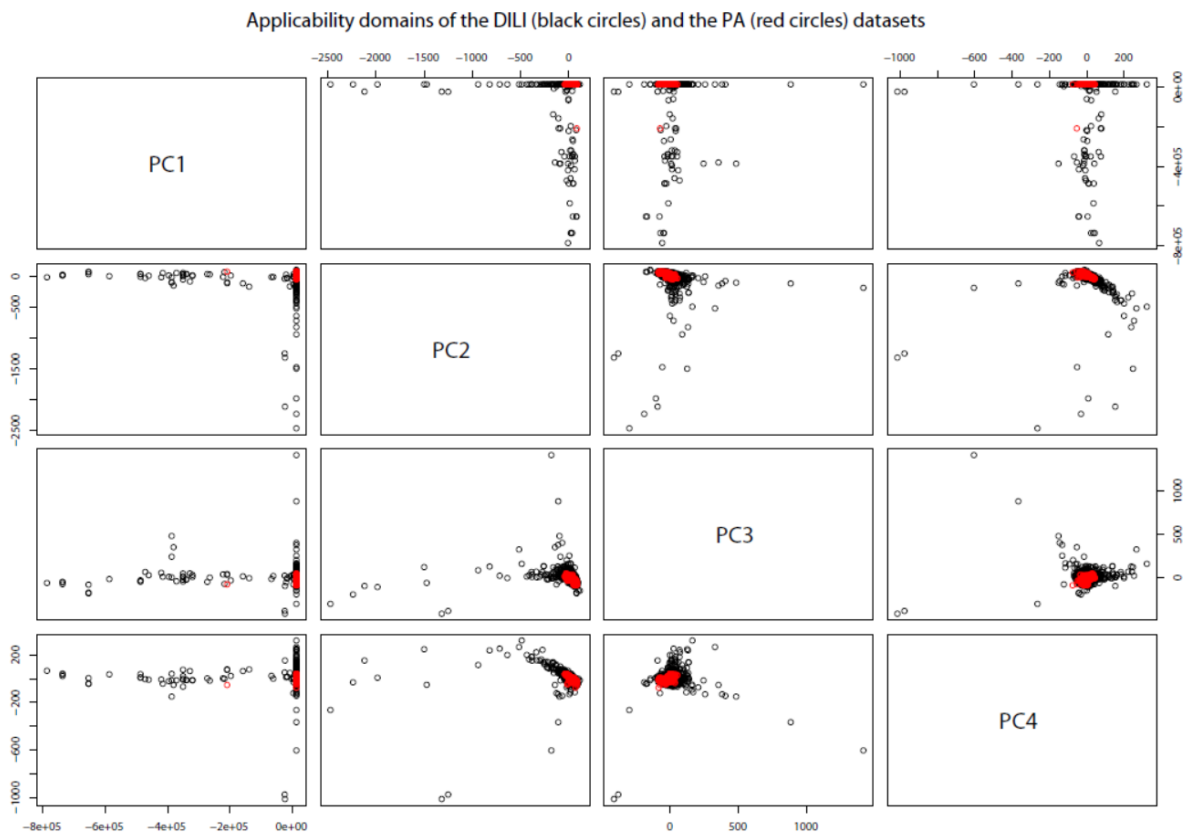
modelling complements other experiments and provides additional information, it is not able to replace *in vitro* and *in vivo* testing completely.

The greatest challenge is the performance of the models. This has to be validated e.g. by cross-validation before the model can be used on the substances of interest. Also group statement could be easily obtained, due caution has to be taken while interpreting the results of predictive models for singly compounds and if possible, comparison to alright published data is advisable, as a form of external validation.

9 Software

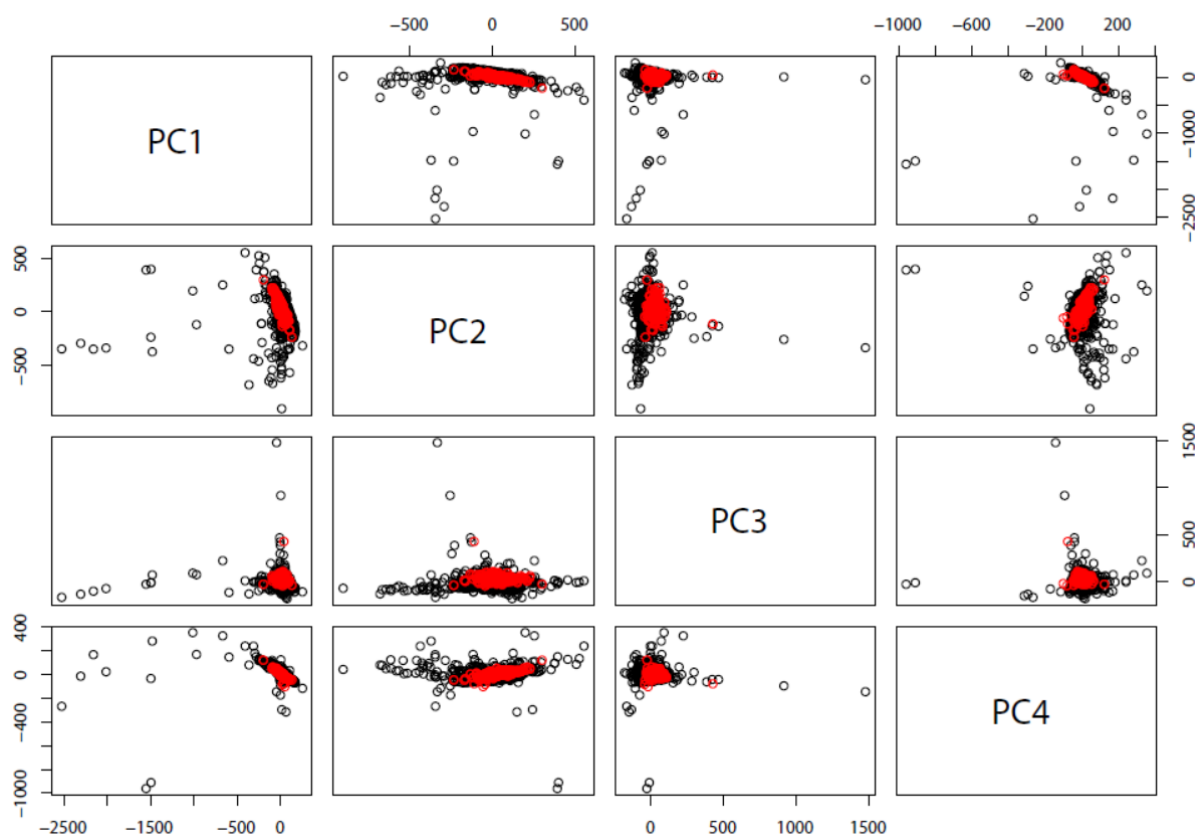
- Gnu R 3.3.3 (<http://www.R-project.org>)
- Gephi 0.82 (<https://gephi.org>)
- PaDEL-Descriptor 2.21 (<http://www.yapcsoft.com/dd/padeldescriptor>)
- IBM SPSS Statistics version 25
- LAZAR version 1.3.1 (<https://lazar.in-silico.ch/predict>)
- OpenBabel version 2.3.1 (<https://openbabel.org>)
- TensorFlow program (<https://www.tensorflow.org/>)
- Keras (<https://www.tensorflow.org/guide/keras>)
- Python (<https://www.python.org>)

10 Annex



Annex 1: Applicability domain of the DILI and the PA dataset, showing the four first principal components (PC1 – PC4).

Applicability domains of the DILI (black circles) and the PKI (red circles) datasets



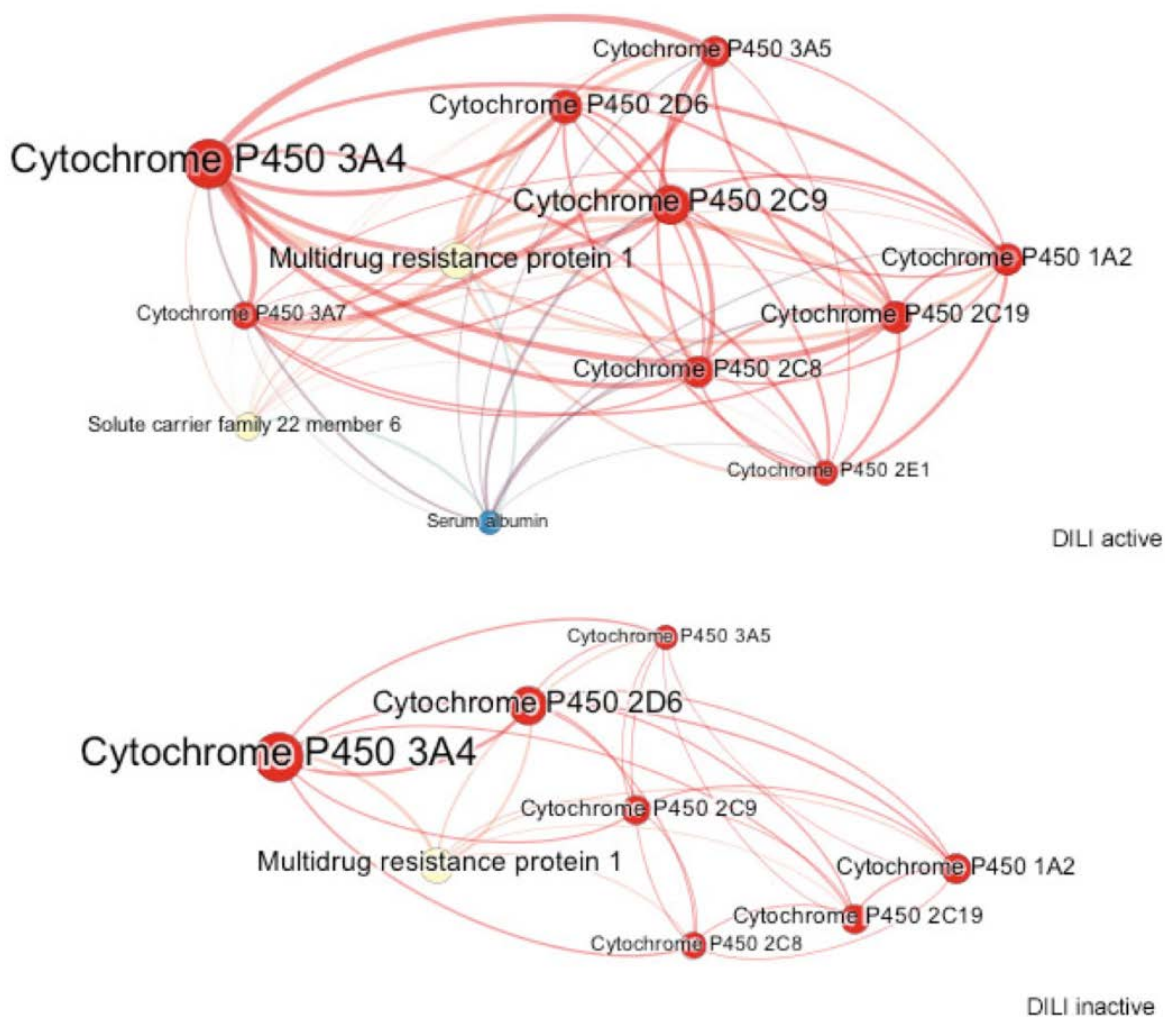
Annex 2: Comparison of the applicability domain of the DILI and the PKI dataset, showing the four first principal components (PC1 – PC4).

Black circles denote model-defining DILI compounds, red circles denote tested PKIs. The plots indicate that the PKI dataset is completely included in the applicability domain of the DILI dataset.

Descriptor importance of the RF model



Annex 3: Predictor importance: list of the 30 most important predictors and the mean decrease in accuracy, which would result by removal of this descriptor from the model.



Annex 4: Visualization of the drug-bioentity networks for hepatotoxic drugs ('DILI active') and non-hepatotoxic drugs ('DILI inactive')

11 Supplementary material

Supplementary material S1: Information on the substructure search performed in PubChem to obtain the PA dataset.

Supplementary material S2: PKI dataset. Spreadsheet 1: Overview of substances, SMILES and CID (compound identifier from PubChem). Spreadsheet 2: Protein kinase targets of the substances, '1': is target of protein kinase mentioned in column title, '0': is no target of protein kinase mentioned in column title. Spreadsheet 3: Probability of hepatotoxicity with RF and aNN model.

Supplementary material S3: Database of substances used for model training, with DILI-outcome indicated

Supplementary material S4: Complete architecture of drug - bioentity networks for hepatotoxic drugs ('DILI active') and non-hepatotoxic drugs ('DILI inactive')

Supplementary material S5: Full list of descriptor sets used in model building.

Supplementary material S6: Literature review of recent machine learning efforts for drug-induced liver injury

12 References

- Aguer C, Gambarotta D, Mailloux RJ, Moffat C, Dent R, et al. 2011. Galactose enhances oxidative metabolism and reveals mitochondrial dysfunction in human primary muscle cells. *PLoS One* 6:e28536
- Ahmed SN, Siddiqi ZA. 2006. Antiepileptic drugs and liver disease. *Seizure* 15:156-64
- Aleo MD, Luo Y, Swiss R, Bonin PD, Potter DM, Will Y. 2014. Human drug-induced liver injury severity is highly associated with dual inhibition of liver mitochondrial function and bile salt export pump. *Hepatology (Baltimore, Md)* 60:1015-22
- ANZFA. 2001. Pyrrolizidine alkaloids in food. *A Toxicological Review and Risk Assessment*. ed. Authority, ANZF, pp. 1-16
- Armstrong SJ, Zuckerman AJ, Bird RG. 1972. Induction of morphological changes in human embryo liver cells by the pyrrolizidine alkaloid lasiocarpine. *British journal of experimental pathology* 53:145-9
- Barysz M, Jashari G, Lall RS, Srivastava AK, Trinajstic N. 1983. On the distance matrix of molecules containing heteroatoms. In *Chemical Applications of Topology and Graph Theory*, pp. 222-30. Amsterdam, The Netherlands: Elsevier
- Basak SC, Harriss DK, Magnuson VR. Comparative Study of Lipophilicity *versus* Topological Molecular Descriptors in Biological Correlations. *Journal of Pharmaceutical Sciences* 73:429-37
- Bender A, Mussa HY, Glen RC, Reiling S. 2004. Molecular similarity searching using atom environments, information-based feature selection, and a naive Bayesian classifier. *J Chem Inf Comput Sci* 44:170-8
- Benichou C, Danan G, Flahault A. 1993. Causality assessment of adverse reactions to drugs--II. An original model for validation of drug causality assessment methods: case reports with positive rechallenge. *J Clin Epidemiol* 46:1331-6
- Bergmeir C, Benítez JM. 2012. Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. *Journal of Statistical Software* 46:1-26
- Bishop-Bailey D, Thomson S, Askari A, Faulkner A, Wheeler-Jones C. 2014. Lipid-metabolizing CYPs in the regulation and dysregulation of metabolism. *Annu Rev Nutr* 34:261-79
- Blower PE, Cross KP. 2006. Decision Tree Methods in Pharmaceutical Research. *Current topics in medicinal chemistry* 6:31-9
- Boelsterli UA, Lee KK. 2014. Mechanisms of isoniazid-induced idiosyncratic liver injury: emerging role of mitochondrial stress. *Journal of gastroenterology and hepatology* 29:678-87
- Bramer M. 2013. *Principles of Data Mining*. p. 444: Springer-Verlag
- Breimann L. 2001. Random Forests. *Machine Learning* 45:5-32
- Breimann L. 2003. *Manual-Setting Up, Using, And Understanding Random Forests V4.0.1-33*
- Bull LB, Dick AT. 1959. The chronic pathological effects on the liver of the rat of the pyrrolizidine alkaloids heliotrine, lasiocarpine and their N-oxides. *J Path Bact* 78:483-502
- Bull LB, Dick AT, McKenzie JS. 1958. The acute toxic effects of heliotrine and lasiocarpine, and their N-oxides, on the rat. *J Path Bact* 75:17-25
- Burden FR. 1989. Molecular identification number for substructure searches. *Journal of Chemical Information and Computer Sciences* 29:225-7

- Butler WH, Mattocks AR, Barnes JM. 1970. Lesions in the liver and lungs of rats given pyrrole derivatives of pyrrolizidine alkaloids. *J Path* 100:169-75
- Chai J, He Y, Cai SY, Jiang Z, Wang H, et al. 2012. Elevated hepatic multidrug resistance-associated protein 3/ATP-binding cassette subfamily C 3 expression in human obstructive cholestasis is mediated through tumor necrosis factor alpha and c-Jun NH2-terminal kinase/stress-activated protein kinase-signaling pathway. *Hepatology* 55:1485-94
- Chalhoub WM, Sliman KD, Arumuganathan M, Lewis JH. 2014. Drug-induced liver injury: what was new in 2013? *Expert Opin Drug Metab Toxicol* 10:959-80
- Chawla NV, Bowyer KW, Hall LO. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16:321-57
- Chen M, Borlak J, Tong W. 2013. High lipophilicity and high daily dose of oral medications are associated with significant risk for drug-induced liver injury. *Hepatology (Baltimore, Md)* 58:388-96
- Chen M, Suzuki A, Thakkar S, Yu K, Hu C, Tong W. 2016. DILrank: the largest reference drug list ranked by the risk for developing drug-induced liver injury in humans. *Drug Discov Today* 21:648-53
- Chen T, Mei N, Fu PP. 2010. Genotoxicity of pyrrolizidine alkaloids. *J Appl Toxicol* 30:183-96
- Crabtree HG. 1928. The carbohydrate metabolism of certain pathological overgrowths *Biochem J* 22:1289-98
- Daly AK, Donaldson PT, Bhatnagar P, Shen Y, Pe'er I, et al. 2009. HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. *Nature genetics* 41:816-9
- Danan G, Benichou C. 1993. Causality assessment of adverse reactions to drugs--I. A novel method based on the conclusions of international consensus meetings: application to drug-induced liver injuries. *J Clin Epidemiol* 46:1323-30
- Dar AC, Shokat KM. 2011. The evolution of protein kinase inhibitors from antagonists to agonists of cellular signaling. *Annu Rev Biochem* 80:769-95
- de Wildt SN, Kearns GL, Leeder JS, van den Anker JN. 1999. Cytochrome P450 3A: ontogeny and drug disposition. *Clin Pharmacokinet* 37:485-505
- DeLeve LD, Ito Y, Bethea NW, McCuskey MK, Wang X, McCuskey RS. 2003. Embolization by sinusoidal lining cells obstructs the microcirculation in rat sinusoidal obstruction syndrome. *Am J Physiol Gastrointest Liver Physiol* 284:G1045-G52
- DeLeve LD, Wang X, Kuhlenkamp JF, Kaplowitz N. 1996. Toxicity of Azathioprine and Monocrotaline in Murine Sinusoidal Endothelial Cells and Hepatocytes: The Role of Glutathione and Relevance to Hepatic Venooclusive Disease. *Hepatology* 23:589-99
- Dong H, Haining RL, Thummel KE, Rettie AE, Nelson SD. 2000. Involvement of human cytochrome P450 2D6 in the bioactivation of acetaminophen. *Drug Metab Dispos* 28:1397-400
- Doostdar H, Grant MH, Melvin WT, Wolf CR, Burke MD. 1993. The effects of inducing agents on cytochrome P450 and UDP-glucuronyltransferase activities in human HEPG2 hepatoma cells. *Biochemical pharmacology* 46:629-35
- EFSA. 2011. Scientific Opinion on Pyrrolizidine alkaloids in food and feed. *EFSA Journal* 9:1-134
- Ekins S, Williams AJ, Xu JJ. 2010. A predictive ligand-based Bayesian model for human drug-induced liver injury. *Drug Metab. Dispos.* 38:2302-8
- EMA. 2014. EMA/HMPC/893108/2011: Public statement on the use of herbal medicinal products containing toxic, unsaturated pyrrolizidine alkaloids (PAs).1-24

- EMA. 2016. EMA/HMPC/328782/2016: Public statement on contamination of herbal medicinal products/traditional herbal medicinal products with pyrrolizidine alkaloids. 1-11
- Fashe MM, Juvonen RO, Petsalo A, Vepsalainen J, Pasanen M, Rahnasto-Rilla M. 2015. In silico prediction of the site of oxidation by cytochrome P450 3A4 that leads to the formation of the toxic metabolites of pyrrolizidine alkaloids. *Chem Res Toxicol* 28:702-10
- Field RA, Stegelmeier BL, Colegate SM, Brown AW, Green BT. 2015. An in vitro comparison of the cytotoxic potential of selected dehydropyrrolizidine alkaloids and some N-oxides. *Toxicol* 97:36-45
- Fleming I. 2014. The pharmacology of the cytochrome P450 epoxygenase/soluble epoxide hydrolase axis in the vasculature and cardiovascular disease. *Pharmacol Rev* 66:1106-40
- Fonti V. 2017. *Feature Selection using LASSO*. Research paper. VU Amsterdam. 26 pp.
- Fu PP, Chou MW, Churchwell M, Wang Y, Zhao Y, et al. 2010. High-Performance Liquid Chromatography Electrospray Ionization Tandem Mass Spectrometry for the Detection and Quantitation of Pyrrolizidine Alkaloid-Derived DNA Adducts in Vitro and in Vivo. *Chem Res Toxicol* 23:637-52
- Fu PP, Xia Q, Lin G, Chou MW. 2004. Pyrrolizidine alkaloids--genotoxicity, metabolism enzymes, metabolic activation, and mechanisms. *Drug Metab Rev* 36:1-55
- Galeotti N, Vivoli E, Bilia AR, Vincieri FF, Ghelardini C. 2010. St. John's wort reduces neuropathic pain through a hypericin-mediated inhibition of the protein kinase Cgamma and epsilon activity. *Biochem Pharmacol* 79:1327-36
- Ganesan S, Tekwani BL, Sahu R, Tripathi LM, Walker LA. 2009. Cytochrome P(450)-dependent toxic effects of primaquine on human erythrocytes. *Toxicol Appl Pharmacol* 241:14-22
- Gao H, Ruan JQ, Chen J, Li N, Ke CQ, et al. 2015. Blood pyrrole-protein adducts as a diagnostic and prognostic index in pyrrolizidine alkaloid-hepatic sinusoidal obstruction syndrome. *Drug Des Devel Ther* 9:4861-8
- Gitlin N. 1980. Salicylate hepatotoxicity: the potential role of hypoalbuminemia. *J Clin Gastroenterol* 2:281-5
- Gordon GJ, Coleman WB, Grisham JW. 2000. Bax-mediated apoptosis in the livers of rats after partial hepatectomy in the retrorsine model of hepatocellular injury. *Hepatology* 32:312-20
- Gradhand U, Lang T, Schaeffeler E, Glaeser H, Tegude H, et al. 2008. Variability in human hepatic MRP4 expression: influence of cholestasis and genotype. *Pharmacogenomics J* 8:42-52
- Gramatica P, Corradi M, Consonni V. 2000. Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors. *Chemosphere* 41:763-77
- Greene N, Fisk L, Naven RT, Note RR, Patel ML, Pelletier DJ. 2010. Developing structure-activity relationships for the prediction of hepatotoxicity. *Chemical Research in Toxicology* 23:1215-22
- Guo YX, Xu XF, Zhang QZ, Li C, Deng Y, et al. 2015. The inhibition of hepatic bile acids transporters Ntcp and Bsep is involved in the pathogenesis of isoniazid/rifampicin-induced hepatotoxicity. *Toxicology mechanisms and methods* 25:382-7
- Hall LH, Kier LB. 1995. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *Journal of Chemical Information and Computer Sciences* 35:1039-45

- Hammann F, Schoning V, Drewe J. 2018. Prediction of clinically relevant drug-induced liver injury from structure using machine learning. *J Appl Toxicol*
- Hansen K, Mika S, Schroeter T, Sutter A, ter Laak A, et al. 2009. Benchmark data set for in silico prediction of Ames mutagenicity. *J Chem Inf Model* 49:2077-81
- Hartmann T, Ehmke A, Eilert U, von Borstel K, Thuring C. 1989. Sites of synthesis, translocation and accumulation of pyrrolizidine alkaloid N-oxides in *Senecio vulgaris* L. *Planta* 177:98-107
- Hartmann T, Witte L. 1995. Chemistry, Biology and Chemoecology of the Pyrrolizidine Alkaloids. In *Alkaloids: Chemical and Biological Perspectives*, ed. Pelletier, pp. 155-233. Pergamon, London, New York
- Hessel S, Gottschalk C, Schumann D, These A, Preiss-Weigert A, Lampen A. 2014. Structure-activity relationship in the passage of different pyrrolizidine alkaloids through the gastrointestinal barrier: ABCB1 excretes heliotrine and echimidine. *Mol Nutr Food Res* 58:995-1004
- Hunt CM, Westerkam WR, Stave GM. 1992. Effect of age and gender on the activity of human hepatic CYP3A. *Biochemical pharmacology* 44:275-83
- Ibanez L, Perez E, Vidal X, Laporte JR, Grup d'Estudi Multicentric d'Hepatotoxicitat Aguda de B. 2002. Prospective surveillance of acute serious liver disease unrelated to infectious, obstructive, or metabolic diseases: epidemiological and clinical features, and exposure to drugs. *J Hepatol* 37:592-600
- ICH. 2011. Guidance on genotoxicity testing and data interpretation for pharmaceuticals intended for human use S2(R1). p. 29
- Iyer VV, Yang H, Ierapetritou MG, Roth CM. 2010. Effects of glucose and insulin on HepG2-C3A cell metabolism. *Biotechnol Bioeng* 107:347-56
- Jago MV. 1971. Factors affecting the chronic hepatotoxicity of pyrrolizidine alkaloids. *The Journal of Pathology* 105:1-11
- Jeon JY, Sparreboom A, Baker SD. 2017. Kinase Inhibitors: The Reality Behind the Success. *Clin Pharmacol Ther* 102:726-30
- Jeong W, Doroshow JH, Kummar S. 2013. United States Food and Drug Administration approved oral kinase inhibitors for the treatment of malignancies. *Curr Probl Cancer* 37:110-44
- Ji L, Chen Y, Liu T, Wang Z. 2008. Involvement of Bcl-xL degradation and mitochondrial-mediated apoptotic pathway in pyrrolizidine alkaloids-induced apoptosis in hepatocytes. *Toxicol Appl Pharmacol* 231:393-400
- Jornil J, Nielsen TS, Rosendal I, Ahlner J, Zackrisson AL, et al. 2013. A poor metabolizer of both CYP2C19 and CYP2D6 identified by mechanistic pharmacokinetic simulation in a fatal drug poisoning case involving venlafaxine. *Forensic Sci Int* 226:e26-31
- Kalthoff S, Ehmer U, Freiberg N, Manns MP, Strassburg CP. 2010. Interaction between oxidative stress sensor Nrf2 and xenobiotic-activated aryl hydrocarbon receptor in the regulation of the human phase II detoxifying UDP-glucuronosyltransferase 1A10. *J Biol Chem* 285:5993-6002
- Kazius J, McGuire R, Bursi R. 2005. Derivation and validation of toxicophores for mutagenicity prediction. *J Med Chem* 48:312-20
- Khan D, Khan AU. 2016. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discov Today* 21:1291-302
- Kim HY, Stermitz FR, Molyneux RJ, Wilson DW, Taylor D, Coulombe RA, Jr. 1993. Structural influences on pyrrolizidine alkaloid-induced cytopathology. *Toxicol Appl Pharmacol* 122:61-9

- Kock K, Ferslew BC, Netterberg I, Yang K, Urban TJ, et al. 2014. Risk factors for development of cholestatic drug-induced liver injury: inhibition of hepatic basolateral bile acid transporters multidrug resistance-associated proteins 3 and 4. *Drug Metab Dispos* 42:665-74
- Lammert C, Einarsson S, Saha C, Niklasson A, Bjornsson E, Chalasani N. 2008. Relationship between daily dose of oral medications and idiosyncratic drug-induced liver injury: search for signals. *Hepatology* 47:2003-9
- Langel D, Ober D, Pelser PB. 2011. The evolution of pyrrolizidine alkaloid biosynthesis and diversity in the Senecioneae. *Phytochemistry Reviews* 10:3-74
- Lasser KE, Allen PD, Woolhandler SJ, Himmelstein DU, Wolfe SM, Bor DH. 2002. Timing of new black box warnings and withdrawals for prescription medications. *JAMA* 287:2215-20
- Li N, Xia Q, Ruan J, Fu PP, Lin G. 2011. Hepatotoxicity and Tumorigenicity Induced by Metabolic Activation of Pyrrolizidine Alkaloids in Herbs. *Current Drug Metabolism* 12
- Li X, Cameron MD. 2012. Potential role of a quetiapine metabolite in quetiapine-induced neutropenia and agranulocytosis. *Chem Res Toxicol* 25:1004-11
- Li YH, Kan WL, Li N, Lin G. 2013. Assessment of pyrrolizidine alkaloid-induced toxicity in an in vitro screening model. *J Ethnopharmacol* 150:560-7
- Lima A, Bernardes M, Azevedo R, Medeiros R, Seabra V. 2015. Pharmacogenomics of Methotrexate Membrane Transport Pathway: Can Clinical Response to Methotrexate in Rheumatoid Arthritis Be Predicted? *Int J Mol Sci* 16:13760-80
- Lin G. 1998. Microsomal Formation of a Pyrrolic Alcohol Glutathione Conjugate of Clivorine Firm Evidence for the Formation of a Pyrrolic Metabolite of an Otonecine-Type Pyrrolizidine Alkaloid. *Drug Metabolism and Disposition* 26:181-4
- Lindigkeit R, Biller A, Buch M, Schiebel H-M, Boppré M, Hartmann T. 1997. The two faces of pyrrolizidine alkaloids: the role of the tertiary amine and its N-oxide in chemical defense of insects with acquired plant alkaloids. *Eur J Biochem* 245
- Makhlouf HA, Helmy A, Fawzy E, El-Attar M, Rashed HA. 2008. A prospective study of antituberculous drug-induced hepatotoxicity in an area endemic for liver diseases. *Hepatol Int* 2:353-60
- Marin-Hernandez A, Rodriguez-Enriquez S, Vital-Gonzalez PA, Flores-Rodriguez FL, Macias-Silva M, et al. 2006. Determining and understanding the control of glycolysis in fast-growth tumor cells. Flux control by an over-expressed but strongly product-inhibited hexokinase. *FEBS J* 273:1975-88
- Marroquin LD, Hynes J, Dykens JA, Jamieson JD, Will Y. 2007. Circumventing the Crabtree effect: replacing media glucose with galactose increases susceptibility of HepG2 cells to mitochondrial toxicants. *Toxicol Sci* 97:539-47
- Mattocks AR. 1986. *Chemistry and Toxicology of Pyrrolizidine Alkaloids*: Academic Press
- Meharena HS, Chang P, Keshwani MM, Oruganty K, Nene AK, et al. 2013. Deciphering the structural basis of eukaryotic protein kinase regulation. *PLoS Biol* 11:e1001680
- Merz KH, Schrenk D. 2016. Interim relative potency factors for the toxicological risk assessment of pyrrolizidine alkaloids in food and herbal medicines. *Toxicol Lett* 263:44-57
- Miners JO, Birkett DJ. 1998. Cytochrome P4502C9: an enzyme of major importance in human drug metabolism. *British Journal of Clinical Pharmacology* 45:525-38
- Mingard C, Paech F, Bouitbir J, Krahenbuhl S. 2018. Mechanisms of toxicity associated with six tyrosine kinase inhibitors in human hepatocyte cell lines. *J Appl Toxicol* 38:418-31

- Mingatto FE, Dorta DJ, dos Santos AB, Carvalho I, da Silva CH, et al. 2007. Dehydromonocrotaline inhibits mitochondrial complex I. A potential mechanism accounting for hepatotoxicity of monocrotaline. *Toxicol* 50:724-30
- Mitchell JB. 2014. Machine learning methods in chemoinformatics. *Wiley Interdiscip Rev Comput Mol Sci* 4:468-81
- Morgan RE, Trauner M, van Staden CJ, Lee PH, Ramachandran B, et al. 2010. Interference with bile salt export pump function is a susceptibility factor for human liver injury in drug development. *Toxicol Sci* 118:485-500
- Muegge I, Mukherjee P. 2016. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opin Drug Discov* 11:137-48
- Najibi A, Heidari R, Zarifi J, Jamshidzadeh A, Firoozabadi N, Niknahad H. 2016. Evaluating the Role of Drug Metabolism and Reactive Intermediates in Trazodone-Induced Cytotoxicity toward Freshly-Isolated Rat Hepatocytes. *Drug Res (Stuttg)* 66:592-6
- Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. 2009. A Practical Overview of Quantitative Structure-Activity Relationship. *EXCLI Journal* 8:74-88
- National Cancer Institute. 2006. Common Terminology Criteria for Adverse Events v3.0 (CTCAE). ed. Program, CTE
- Neumann MG, Cohen LB, Opris M, Nanau R, Jeong H. 2015. Hepatotoxicity of Pyrrolizidine Alkaloids. *J Pharm Pharm Sci* 18:825-43
- Newby D, Freitas AA, Ghafourian T. 2015. Decision trees to characterise the roles of permeability and solubility on the prediction of oral absorption. *Eur J Med Chem* 90:751-65
- Niederer C, Behra R, Harder A, Schwarzenbach RP, Escher BI. 2004. Mechanistic approaches for evaluating the toxicity of reactive organochlorines and epoxides in green algae. *Environmental Toxicology and Chemistry* 23:697-704
- NTP. 1978. Bioassay of lasiocarpine for possible carcinogenicity. pp. 1-82
- NTP. 2003. Toxicology and Carcinogenesis Studies of Riddelliine (CAS No. 23246-96-0) in F344/N Rats And B6c3F₁ Mice (Gavage Studies). ed. Health, NIO
- O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. 2011. Open Babel: An open chemical toolbox. *J Cheminform* 3:33
- Open Babel community. 2011. *Molecular fingerprints and similarity searching — Open Babel v2.3.1 documentation.* *Openbabel.org.* <https://openbabel.org/docs/dev/Fingerprints/intro.html>, December 31, 2018
- Paech F, Bouitbir J, Krahenbuhl S. 2017. Hepatocellular Toxicity Associated with Tyrosine Kinase Inhibitors: Mitochondrial Damage and Inhibition of Glycolysis. *Front Pharmacol* 8:367
- Parkinson A, Mudra DR, Johnson C, Dwyer A, Carroll KM. 2004. The effects of gender, age, ethnicity, and liver cirrhosis on cytochrome P450 enzyme activity in human liver microsomes and inducibility in cultured human hepatocytes. *Toxicol Appl Pharmacol* 199:193-209
- Pellinen P, Honkakoski P, Stenback F, Niemitz M, Alhava E, et al. 1994. Cocaine N-demethylation and the metabolism-related hepatotoxicity can be prevented by cytochrome P450 3A inhibitors. *Eur J Pharmacol* 270:35-43
- Regev A, Seeff LB, Merz M, Ormarsdottir S, Aithal GP, et al. 2014. Causality assessment for suspected DILI during clinical phases of drug development. *Drug Saf* 37 Suppl 1:S47-56
- Rendic S. 2002. Summary of information on human CYP enzymes: human P450 metabolism data. *Drug Metab Rev* 34:83-448

- Reuben A, Koch DG, Lee WM, Acute Liver Failure Study G. 2010. Drug-induced acute liver failure: results of a U.S. multicenter, prospective study. *Hepatology* 52:2065-76
- Rodrigues AC. 2010. Efflux and uptake transporters as determinants of statin response. *Expert Opin Drug Metab Toxicol* 6:621-32
- Roskoski R, Jr. 2015. A historical overview of protein kinases and their targeted small molecule inhibitors. *Pharmacol Res* 100:1-23
- Ruan J, Liao C, Ye Y, Lin G. 2014a. Lack of metabolic activation and predominant formation of an excreted metabolite of nontoxic platynecine-type pyrrolizidine alkaloids. *Chem Res Toxicol* 27:7-16
- Ruan J, Yang M, Fu P, Ye Y, Lin G. 2014b. Metabolic activation of pyrrolizidine alkaloids: insights into the structural and enzymatic basis. *Chem Res Toxicol* 27:1030-9
- Rubiolo P, Pieters L, Calomme M, Bicchi C, Vlietinck A, Vanden Berghe D. 1992. Mutagenicity of pyrrolizidine alkaloids in the *Salmonella typhimurium*/mammalian microsome system. *Mutat Res* 281:143-7
- Rücker C, Rücker G, Meringer M. 2007. γ -Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* 47:2345-57
- Schoental R, Head MA. 1957. Progression of liver lesions produced in rats by temporary treatment with pyrrolizidine (senecio) alkaloids, and the effects of betaine and high casein diet. *Br J Cancer* 11:535-44
- Schöning V, Hammann F, Peinl M, Drewe J. 2017. Editor's Highlight: Identification of Any Structure-Specific Hepatotoxic Potential of Different Pyrrolizidine Alkaloids Using Random Forests and Artificial Neural Networks. *Toxicol Sci* 160:361-70
- Shah RR, Morganroth J, Shah DR. 2013. Hepatotoxicity of tyrosine kinase inhibitors: clinical and regulatory perspectives. *Drug Saf* 36:491-503
- Spjuth O, Alvarsson J, Berg A, Eklund M, Kuhn S, et al. 2009. Bioclipse 2: A scriptable integration platform for the life sciences. *BMC Bioinformatics* 10:1-5
- Spjuth O, Helmus T, Willighagen EL, Kuhn S, Eklund M, et al. 2007. Bioclipse: an open source workbench for chemo- and bioinformatics. *BMC Bioinformatics* 8:1-10
- Srinivas N, Sandeep KS, Anusha Y, Devendra BN. 2014. In Vitro Cytotoxic Evaluation and Detoxification of Monocrotaline (Mct) Alkaloid: An In Silico Approach. *International Invention Journal of Biochemistry and Bioinformatics* 2:20-9
- Stine JG, Chalasani NP. 2017. Drug Hepatotoxicity: Environmental Factors. *Clin Liver Dis* 21:103-13
- Stine JG, Lewis JH. 2011. Drug-induced liver injury: a summary of recent advances. *Expert Opin Drug Metab Toxicol* 7:875-90
- Takanashi H, Umeda M, Hirono I. 1980. Chromosomal aberrations and mutations in cultured mammalian cells induced by pyrrolizidine alkaloids. *Mutation Research* 78:67-77
- Takeda M, Okamoto I, Nakagawa K. 2015. Pooled safety analysis of EGFR-TKI treatment for EGFR mutation-positive non-small cell lung cancer. *Lung Cancer* 88:74-9
- Tamta H, Pawar RS, Wamer WG, Grundel E, Krynitsky AJ, Rader JI. 2012. Comparison of metabolism-mediated effects of pyrrolizidine alkaloids in a HepG2/C3A cell-S9 co-incubation system and quantification of their glutathione conjugates. *Xenobiotica* 42:1038-48
- Teh LK, Bertilsson L. 2012. Pharmacogenomics of CYP2D6: molecular genetics, interethnic differences and clinical importance. *Drug Metab Pharmacokinet* 27:55-67
- Teo YL, Ho HK, Chan A. 2013. Risk of tyrosine kinase inhibitors-induced hepatotoxicity in cancer patients: a meta-analysis. *Cancer Treat Rev* 39:199-206

- Teo YL, Ho HK, Chan A. 2015. Formation of reactive metabolites and management of tyrosine kinase inhibitor-induced hepatotoxicity: a literature review. *Expert Opin Drug Metab Toxicol* 11:231-42
- Thompson RA, Isin EM, Ogese MO, Mettetal JT, Williams DP. 2016. Reactive Metabolites: Current and Emerging Risk and Hazard Assessments. *Chem Res Toxicol* 29:505-33
- Walker K, Ginsberg G, Hattis D, Johns DO, Guyton KZ, Sonawane B. 2009. Genetic polymorphism in N-Acetyltransferase (NAT): Population distribution of NAT1 and NAT2 activity. *Journal of toxicology and environmental health. Part B, Critical reviews* 12:440-72
- Wang YP, Yan J, Fu PP, Chou MW. 2005. Human liver microsomal reduction of pyrrolizidine alkaloid N-oxides to form the corresponding carcinogenic parent alkaloid. *Toxicol Lett* 155:411-20
- Weininger D. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 28:31-6
- Westerink WM, Schoonen WG. 2007. Phase II enzyme levels in HepG2 cells and cryopreserved primary human hepatocytes and their induction in HepG2 cells. *Toxicol In Vitro* 21:1592-602
- Wu P, Nielsen TE, Clausen MH. 2015. FDA-approved small-molecule kinase inhibitors. *Trends Pharmacol Sci* 36:422-39
- Xia Q, Ma L, He X, Cai L, Fu PP. 2015. 7-glutathione pyrrole adduct: a potential DNA reactive metabolite of pyrrolizidine alkaloids. *Chem Res Toxicol* 28:615-20
- Xia Q, Zhao Y, Von Tungeln LS, Doerge DR, Lin G, et al. 2013. Pyrrolizidine alkaloid-derived DNA adducts as a common biological biomarker of pyrrolizidine alkaloid-induced tumorigenicity. *Chem Res Toxicol* 26:1384-96
- Yan J, Xia Q, Chou MW, Fu P. 2008. Metabolic activation of retronecine and retronecine N-oxide – formation of DHP-derived DNA adducts. *Toxicology and Industrial Health* 24
- Yang X, Li W, Sun Y, Guo X, Huang W, et al. 2017. Comparative Study of Hepatotoxicity of Pyrrolizidine Alkaloids Retrorsine and Monocrotaline. *Chem Res Toxicol* 30:532-9
- Yap CW. 2011. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry* 32:1466-74
- Yap CW. 2014. *Descriptors*. <http://www.yapcwsoft.com/dd/padeldescriptor/Descriptors.xls>, 27.10.2016
- Yu K, Geng X, Chen M, Zhang J, Wang B, et al. 2014a. High daily dose and being a substrate of cytochrome P450 enzymes are two important predictors of drug-induced liver injury. *Drug Metab Dispos* 42:744-50
- Yu K, Geng X, Chen M, Zhang J, Wang B, et al. 2014b. High daily dose and being a substrate of cytochrome P450 enzymes are two important predictors of drug-induced liver injury. *Drug Metab. Dispos.* 42:744-50
- Zanger UM, Turpeinen M, Klein K, Schwab M. 2008. Functional pharmacogenetics/genomics of human cytochromes P450 involved in drug biotransformation. *Anal Bioanal Chem* 392:1093-108
- Zhang J, Sheng Y, Shi L, Zheng Z, Chen M, et al. 2017. Quercetin and baicalein suppress monocrotaline-induced hepatic sinusoidal obstruction syndrome in rats. *Eur J Pharmacol* 795:160-8
- Zhao Y, Xia Q, Gamboa da Costa G, Yu H, Cai L, Fu PP. 2012. Full structure assignments of pyrrolizidine alkaloid DNA adducts and mechanism of tumor initiation. *Chem Res Toxicol* 25:1985-96
- Zheng Z, Shi L, Sheng Y, Zhang J, Lu B, Ji L. 2016. Chlorogenic acid suppresses monocrotaline-induced sinusoidal obstruction syndrome: The potential contribution of

- NFkappaB, Egr1, Nrf2, MAPKs and PI3K signals. *Environ Toxicol Pharmacol* 46:80-9
- Zhu XW, Xin YJ, Ge HL. 2015. Recursive Random Forests Enable Better Predictive Performance and Model Interpretation than Variable Selection by LASSO. *J Chem Inf Model* 55:736-46