A comparison of twelve machine learning models based on an expanded mutagenicity dataset and their application for predicting pyrrolizidine alkaloid mutagenicity

Christoph Helma[*1], Verena Schöning[2], Philipp Boss[3], and Jürgen Drewe[2]

[1]in silico toxicology gmbh, Rastatterstrasse 41, 4057 Basel, Switzerland

[2]Zeller AG, Seeblickstrasse 4, 8590 Romanshorn, Switzerland

[3]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association, Robert-Rössle-Strasse 10, Berlin, 13125, Germany

[*] Correspondence: Christoph Helma <helma@in-silico.ch>

Introduction

TODO

The main objectives of this study were

- to generate a new training dataset, by combining the most comprehensive public mutagenicity datasets
- to compare the performance of global models (RF, SVM, LR, NN) with local models (`lazar`)
- to compare the performance of MolPrint2D fingerprints with PaDEL descriptors
- to apply these models for the prediction of pyrrolizidine alkaloid mutagenicity

Materials and Methods

Data

Mutagenicity training data

An identical training dataset was used for all models. The training dataset was compiled from the following sources:

- Kazius/Bursi Dataset (4337 compounds, Kazius, McGuire, and Bursi (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 (2016)): 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 with 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 8309 unique structures.

Source code for all data download, extraction and merge operations is publicly available from the git repository https://git.in-silico.ch/mutagenicity-paper under a GPL3 License. The new combined dataset can be found at https://git.in-silico.ch/mutagenicity-paper/data/mutagenicity.csv.

Pyrrolizidine alkaloid (PA) dataset

The testing dataset consisted of 602 different PAs. The compilation of the PA dataset is described in detail in Schöning et al. (2017).

TODO: **Verena** Quellen und Auswahlkriterien

Descriptors

MolPrint2D fingerprints (*MP2D*)

MolPrint2D fingerprints (O'Boyle et al. (2011)) use atom environments as molecular representation. They determine for each atom in a molecule, the atom types of its connected atoms to represent their chemical environment. This resembles basically the chemical concept of functional groups.

In contrast to predefined lists of fragments (e.g. FP3, FP4 or MACCs fingerprints) or descriptors (e.g PaDEL) they are generated dynamically from chemical structures. This has the advantage that they can capture substructures of toxicological relevance that are not included in other descriptors.

Chemical similarities (e.g. Tanimoto indices) can be calculated very efficiently with MolPrint2D fingerprints. Using them as descriptors for global models leads however to huge, sparsely populated matrices that cannot be handled with traditional machine learning algorithms. In our experiments none of the R and Tensorflow algorithms was capable to use them as descriptors.

MolPrint2D fingerprints were calculated with the OpenBabel cheminformatics library (O'Boyle et al. (2011)).

PaDEL descriptors

For R and Tensorflow models, molecular 1D and 2D descriptors were calculated with the PaDEL-Descriptors program (http://www.yapcwsoft.com version 2.21, Yap (2011)).

As the training dataset contained over 8309 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.

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.

Algorithms

`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.

Neighbour identification

Utilizing this modularity, similarity calculations were based both on MolPrint2D fingerprints and on PaDEL descriptors.

For MolPrint2D fingerprints 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).

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

For PaDEL descriptors chemical similarity between two compounds a and b is expressed as the cosine similarity between the descriptor vectors A for a and B for b.

$$sim = \frac{A \cdot B}{|A||B|}$$

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. This are predictions with *high confidence*.
- 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 (*low confidence*).
- 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.

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.

$$p_c = \frac{\sum \mathrm{sim}_{n,c}}{\sum \mathrm{sim}_n}$$

p_c Probability of class c (e.g. mutagenic or non-mutagenic)

$\sum \text{sim}_{n,c}$ Sum of similarities of neighbours with class c

$\sum \text{sim}_n$ Sum of all neighbours

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 (*high confidence*) and predictions with warnings as more distant from the applicability domain (*low confidence*). Quantitative applicability domain information can be obtained from the similarities of individual neighbours.

Availability

- `lazar` experiments for this manuscript: https://git.in-silico.ch/mutagenicity-paper (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

R Random Forest, Support Vector Machines, and Deep Learning

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.

Random Forest

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.

Support Vector Machines

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

Deep Learning

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 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.

TODO: **Verena** kannst Du bitte ueberpruefen, ob das noch stimmt und ggf die Figure 1 anpassen

To validate these models, an internal cross-validation approach was chosen. 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.

Applicability domain

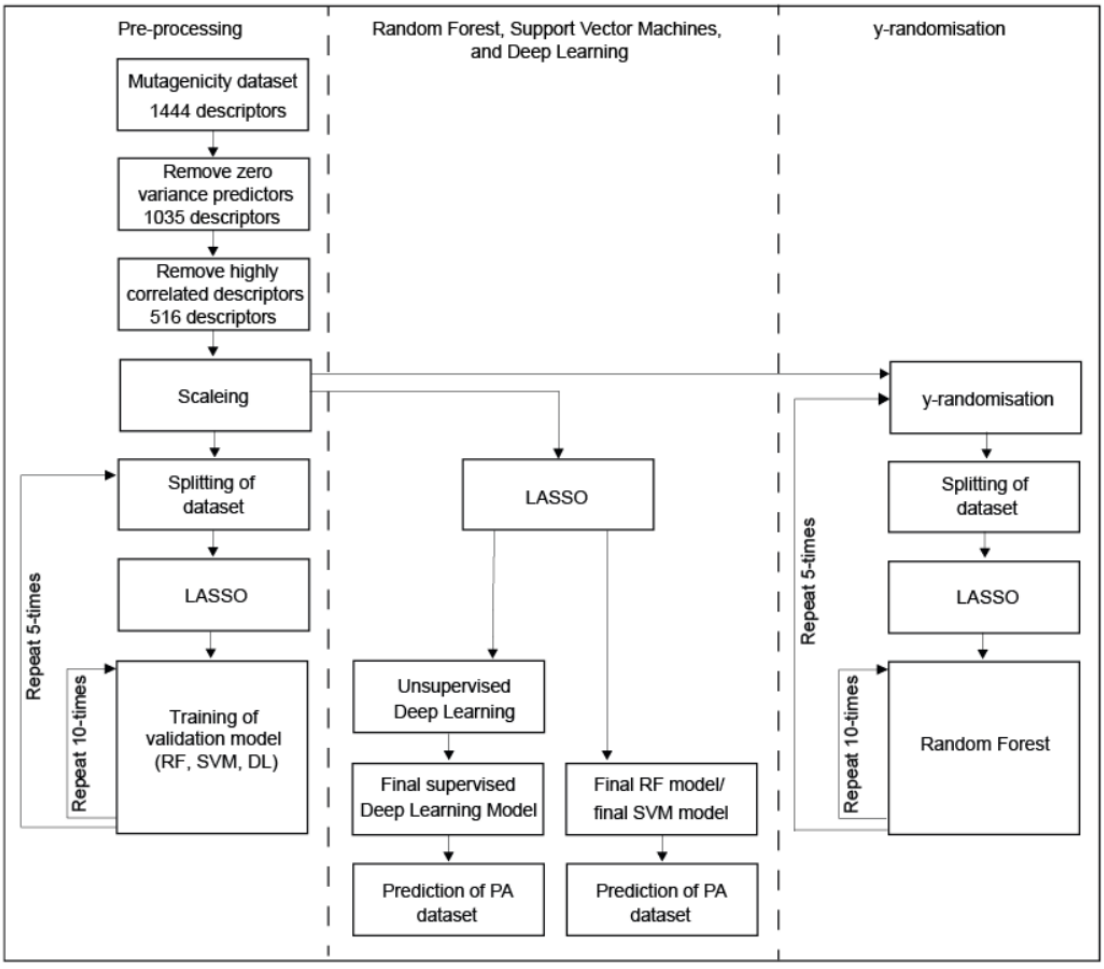

Figure 1: Flowchart of the generation and validation of the models generated in R-project

TODO: **Verena**: Mit welchen Deskriptoren hast Du den Jaccard index berechnet? Fuer den Jaccard index braucht man binaere Deskriptoren (zB MP2D), mit PaDEL Deskriptoren koennte man zB eine euklidische oder cosinus Distanz berechnen.

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.

Availability

R scripts for these experiments can be found in https://git.in-silico.ch/mutagenicity-paper/scripts/R.

Tensorflow models

TODO: **Philipp** bitte ergaenzen

Logistic regression (SGD)

Logistic regression (scikit)

Random forests

Deep Learning

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.

Tensorflow models used the same PaDEL descriptors as the R 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.

TODO: **Philipp** kannst Du bitte ueberpruefen ob die Beschreibung noch stimmt und ob der Ablauf von Verena (Figure 1) auch fuer Deine Modelle gilt

Validation

10-fold cross-validation was used for all Tensorflow models.

Availability

Jupyter notebooks for these experiments can be found in https://git.in-silico.ch/mutagenicity-paper/scripts/tensorflow.

Results

10-fold crossvalidations

Crossvalidation results are summarized in the following tables: Table 1 shows `lazar` results with MolPrint2D and PaDEL descriptors, Table 2 R results and Table 3 Tensorflow results.

Table 1: Summary of lazar crossvalidation results (all predictions/high confidence predictions

	MP2D	PaDEL
Accuracy	0.82/0.84	0.58/0.58
True positive rate/Sensitivity	0.85/0.89	0.32/0.32
True negative rate/Specificity	0.78/0.79	0.79/0.79
Positive predictive value/Precision	0.8/0.83	0.56/0.56
Negative predictive value	0.84/0.85	0.59/0.59
Nr. predictions	7781/5890	4089/4081

Table 2: Summary of R crossvalidation results

	RF	SVM	DL
Accuracy	0.64	0.61	0.56
True positive rate/Sensitivity	0.56	0.56	0.88
True negative rate/Specificity	0.71	0.67	0.24
Positive predictive value/Precision	0.66	0.62	0.53
Negative predictive value	0.62	0.61	0.67
Nr. predictions	8070	8070	8070

Table 3: Summary of tensorflow crossvalidation results

	RF	LR (SGD)	LR (SCIKIT)	NN
Accuracy	0.62	0.63	0.63	
True positive rate/Sensitivity	0.6	0.62	0.61	
True negative rate/Specificity	0.65	0.63	0.64	
Positive predictive value/Precision	0.63	0.62	0.63	
Negative predictive value	0.62	0.63	0.63	
Nr. predictions	8080	8080	8080	

Figure 2 depicts the position of all crossvalidation results in receiver operating characteristic (ROC) space.

Confusion matrices for all models are available from the git repository http://git.in-silico.ch/mutagenicity-paper/10-fold-crossvalidations/confusion-matrices/, individual predictions can be found in http://git.in-silico.ch/mutagenicity-paper/10-fold-crossvalidations/predictions/.

The most accurate crossvalidation predictions have been obtained with `lazar` models with MolPrint2D descriptors (0.84 for predictions with high confidence, 0.82 for all predictions). Models utilizing PaDEL descriptors have generally lower accuracies ranging from TODO to TODO. Sensitivity and specificity is generally well balanced with the exception of `lazar`-PaDEL (low sensitivity) and R deep learning (low specificity) models.

Pyrrolizidine alkaloid mutagenicity predictions

Pyrrolizidine alkaloid mutagenicity predictions are summarized in Table (**???**).

Figure 3 shows the position of pyrrolizidine alkaloids (PA) in the mutagenicity training dataset in MP2D space

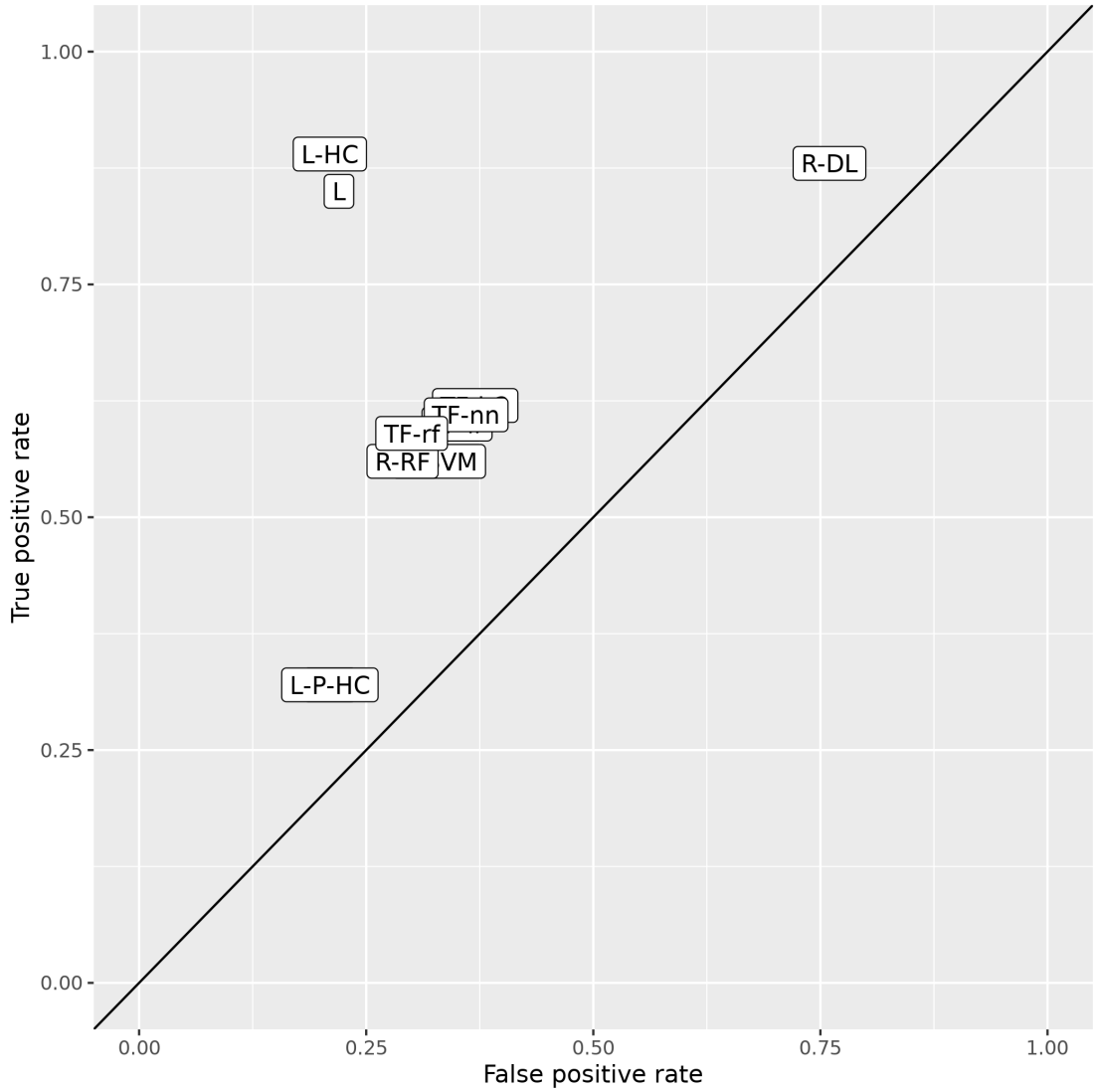

Figure 2: ROC plot of crossvalidation results. *R-RF*: R Random Forests, *R-SVM*: R Support Vector Machines, *R-DL*: R Deep Learning, *TF*: Tensorflow without feature selection, *TF-FS*: Tensorflow with feature selection, *L*: lazar, *L-HC*: lazar high confidence predictions, *L-P*: lazar with PaDEL descriptors, *L-P-HC*: lazar PaDEL high confidence predictions (overlaps with L-P)

Figure 3: t-sne visualisation of mutagenicty training data and pyrrolizidine alkaloids (PA)

Figure 4 shows the position of pyrrolizidine alkaloids (PA) in the mutagenicity training dataset in PADEL space

Discussion

Data

A new training dataset for *Salmonella* mutagenicity was created from three different sources (Kazius, McGuire, and Bursi (2005), Hansen et al. (2009), EFSA (2016)). It contains 8309 unique chemical structures, which is according to our knowledge the largest public mutagenicity dataset presently available. The new training data can be downloaded from https://git.in-silico.ch/mutagenicity-paper/data/mutagenicity.csv.

Model performance

Table **??** and Figure 2 show that the standard `lazar` algorithm (with MP2D fingerprints) give the most accurate crossvalidation results. R Random Forests, Support Vector Machines and Tensorflow models have similar accuracies with balanced sensitivity (true position rate) and specificity (true negative rate). `lazar` models with PaDEL descriptors have low sensitivity and R Deep Learning models have low specificity.

The accuracy of `lazar` *in-silico* predictions are comparable to the interlaboratory variability of the Ames test (80-85% according to Benigni and Giuliani (1988)), especially for predictions with high confidence (84%). This is a clear indication that *in-silico* predictions can be as reliable as the bioassays, if the compounds are close to the applicability domain. This conclusion is also supported by our analysis of `lazar` lowest observed effect level predictions, which are also similar to the experimental variability (Helma et al. (2018)).

The lowest number of predictions (4081) has been obtained from `lazar`/PaDEL high

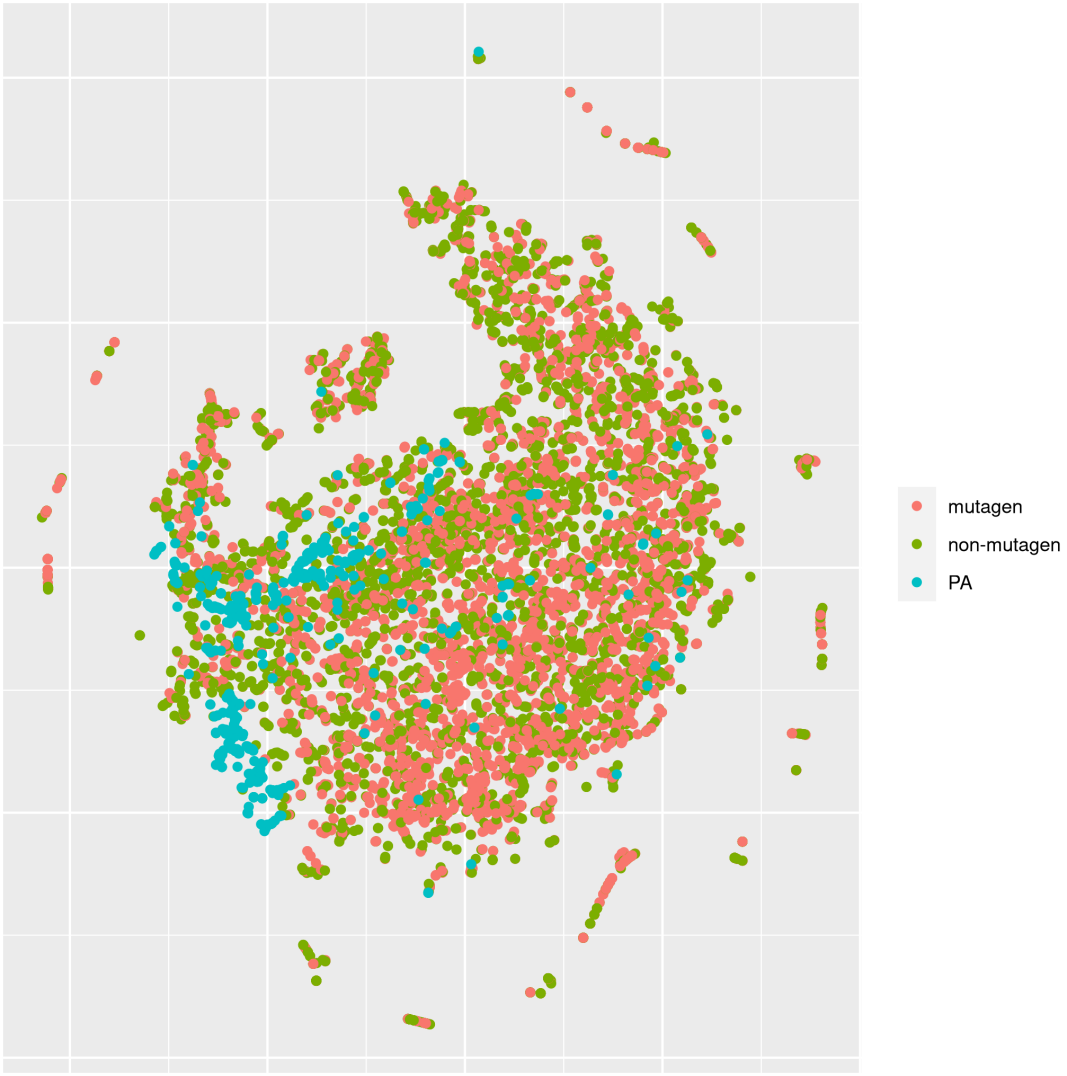

Figure 4: t-sne visualisation of mutagenicty training data and pyrrolizidine alkaloids (PA)

confidence predictions, the largest number of predictions comes from Tensorflow models (). Standard `lazar` give a slightly lower number of predictions (7781) than R and Tensorflow models. This is not necessarily a disadvantage, because `lazar` abstains from predictions, if the query compound is very dissimilar from the compounds in the training set and thus avoids to make predictions for compounds that do not fall into its applicability domain.

There are two major differences between `lazar` and R/Tensorflow models, which might explain the different prediction accuracies:

- `lazar` uses MolPrint2D fingerprints, while all other models use PaDEL descriptors
- `lazar` creates local models for each query compound and the other models use a single global model for all predictions

We will discuss both options in the following sections.

Descriptors

This study uses two types of descriptors to characterize chemical structures.

MolPrint2D fingerprints (MP2D, Bender et al. (2004)) use atom environments (i.e. connected atoms for all atoms in a molecule) as molecular representation, which resembles basically the chemical concept of functional groups. MP2D descriptors are used to determine chemical similarities in lazar, and previous experiments have shown, that they give more accurate results than predefined descriptors (e.g. MACCS, FP2-4) for all investigated endpoints.

PaDEL calculates topological and physical-chemical descriptors.

TODO: **Verena** kannst Du bitte die Deskriptoren nochmals kurz beschreiben

PaDEL descriptors were used for the R and Tensorflow models. In addition we have used PaDEL descriptors to calculate cosine similarities for the `lazar` algorithm and compared

the results with standard MP2D similarities, which led to a significant decrease of `lazar` prediction accuracies. Based on this result we can conclude, that PaDEL descriptors are less suited for similarity calculations than MP2D descriptors.

In order to investigate, if MP2D fingerprints are also a better option for global models we have tried to build R and Tensorflow models both with and without unsupervised feature selection. Unfortunately none of the algorithms was capable to deal with the large and sparsely populated descriptor matrix. Based on this result we can conclude, that MP2D descriptors are at the moment unsuitable for standard global machine learning algorithms. Please note that `lazar` does not suffer from the sparseness problem, because (a) it utilizes internally a much more efficient occurrence based representation and (b) it uses fingerprints only for similarity calculations and mot as model parameters.

Based on these results we can conclude, that PaDEL descriptors are less suited for similarity calculations than MP2D fingerprints and that current standard machine learning algorithms are not capable to utilize chemical fingerprints.

Algorithms

`lazar` is formally a *k-nearest-neighbor* algorithm that searches for similar structures for a given compound and calculates the prediction based on the experimental data for these structures. The QSAR literature calls such models frequently *local models*, because models are generated specifically for each query compound. R and Tensorflow models are in contrast *global models*, i.e. a single model is used to make predictions for all compounds. It has been postulated in the past, that local models are more accurate, because they can account better for mechanisms, that affect only a subset of the training data. Our results seem to support this assumption, because `lazar` models perform better than global models. Both types of models use however different descriptors, and for this reason we cannot draw a definitive conclusion if the model algorithm or the descriptor

type are the reason for the observed differences. In order to answer this question, we would have to use global modelling algorithms that are capable to handle large, sparse binary matrices.

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 2011Fu et al. 2004Ruan 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 bounds 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 that tertiary PAs or dehydropyrrolizidines. As PA *N*-oxides are easily conjugated for extraction, 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.

Conclusions

A new public *Salmonella* mutagenicity training dataset with 8309 compounds was created and used it to train `lazar`, R and Tensorflow models. The best performance was obtained with `lazar` models using MolPrint2D descriptors, with prediction accuracies comparable to the interlaboratory variability of the Ames test. Differences between algorithms (local vs. global models) and/or descriptors (MolPrint2D vs PaDEL) may be responsible for the different prediction accuracies.

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.

References

Bender, Andreas, Hamse Y. Mussa, Robert C. Glen, and Stephan Reiling. 2004. "Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier." *Journal of Chemical Information and Computer Sciences* 44 (1): 170–78. https://doi.org/10.1021/ci034207y.

Benigni, R., and A. Giuliani. 1988. "Computer-assisted Analysis of Interlaboratory Ames Test Variability." *Journal of Toxicology and Environmental Health* 25 (1): 135–48. https://doi.org/10.1080/15287398809531194.

EFSA. 2016. "Guidance on the Establishment of the Residue Definition for Dietary Assessment: EFSA Panel on Plant Protect Products and Their Residues (PPR)." *EFSA Journal*, no. 14: 1–12.

Hansen, Katja, Sebastian Mika, Timon Schroeter, Andreas Sutter, Antonius ter Laak, Thomas Steger-Hartmann, Nikolaus Heinrich, and Klaus-Robert Müller. 2009. "Benchmark Data Set for in Silico Prediction of Ames Mutagenicity." *Journal of Chemical Information and Modeling* 49 (9): 2077–81. https://doi.org/10.1021/ci900161g.

Helma, Christoph, David Vorgrimmler, Denis Gebele, Martin Gütlein, Barbara Engeli, Jürg Zarn, Benoit Schilter, and Elena Lo Piparo. 2018. "Modeling Chronic Toxicity: A Comparison of Experimental Variability with (Q)SAR/Read-Across Predictions." *Frontiers in Pharmacology*, no. 9: 413.

Kazius, J., R. McGuire, and R. Bursi. 2005. "Derivation and Validation of Toxicophores for Mutagenicity Prediction." *J Med Chem*, no. 48: 312–20.

O'Boyle, Noel, Michael Banck, Craig James, Chris Morley, Tim Vandermeersch, and Geoffrey Hutchison. 2011. "Open Babel: An open chemical toolbox." *J. Cheminf.* 3 (1): 33. https://doi.org/doi:10.1186/1758-2946-3-33.

427 Yap, CW. 2011. “PaDEL-Descriptor: An Open Source Software to Calculate Molecular
428 Descriptors and Fingerprints.” *Journal of Computational Chemistry*, no. 32: 1466–74.