

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

4 Christoph Helma<sup>\*1</sup>, Verena Schöning<sup>2</sup>, Philipp Boss<sup>3</sup>, and Jürgen Drewe<sup>2</sup>

5 <sup>1</sup>in silico toxicology gmbh, Rastatterstrasse 41, 4057 Basel, Switzerland

6 <sup>2</sup>Zeller AG, Seeblickstrasse 4, 8590 Romanshorn, Switzerland

7 <sup>3</sup>Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular  
8 Medicine in the Helmholtz Association, Robert-Rössle-Strasse 10, Berlin, 13125, Germany

9 <sup>\*</sup> Correspondence: Christoph Helma <helma@in-silico.ch>

10 **Introduction**

11 TODO

12 The main objectives of this study were

- 13 • to generate a new training dataset, by combining the most comprehensive public  
14 mutagenicity datasets
- 15 • to compare the performance of global models (RF, SVM, LR, NN) with local  
16 models (**lazar**)
- 17 • to compare the performance of MolPrint2D fingerprints with PaDEL descriptors
- 18 • 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](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](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>.

## 43 **Pyrrolizidine alkaloid (PA) dataset**

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

46 TODO: Verena Quellen und Auswahlkriterien

## 47 **Descriptors**

### 48 **MolPrint2D fingerprints (*MP2D*)**

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

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

57 Chemical similarities (e.g. Tanimoto indices) can be calculated very efficiently with Mol-  
58 Print2D fingerprints. Using them as descriptors for global models leads however to huge,  
59 sparsely populated matrices that cannot be handled with traditional machine learning  
60 algorithms. In our experiments none of the R and Tensorflow algorithms was capable to  
61 use them as descriptors.

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

### 64 **PaDEL descriptors**

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

67 As the training dataset contained over 8309 instances, it was decided to delete instances  
68 with missing values during data pre-processing. Furthermore, substances with equivocal  
69 outcome were removed. The final training dataset contained 8080 instances with known  
70 mutagenic potential.

71 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  
72 than 90% or when the frequency ratio of the most common value to the second most  
73 common value was greater than 95:5 (e.g. 95 instances of the most common value and  
74 only 5 or less instances of the second most common value), a descriptor was classified  
75 as having a near zero variance. After that, highly correlated descriptors were removed  
76 using the ‘*findCorrelation*’-function (package ‘*caret*’) with a cut-off of 0.9. This resulted  
77 in a training dataset with 516 descriptors. These descriptors were scaled to be in the  
78 range between 0 and 1 using the ‘*preProcess*’-function (package ‘*caret*’). The scaling  
79 routine was saved in order to apply the same scaling on the testing dataset. As these  
80 three steps did not consider the outcome, it was decided that they do not need to be  
81 included in the cross-validation of the model. To further reduce the number of features,  
82 a LASSO (*least absolute shrinkage and selection operator*) regression was performed  
83 using the ‘*glmnet*’-function (package ‘*glmnet*’). The reduced dataset was used for the  
84 generation of the pre-trained models.  
85

## 86 Algorithms

### 87 **lazar**

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

90 structure **lazar**:

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

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

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

## 100 Neighbour identification

101 Utilizing this modularity, similarity calculations were based both on MolPrint2D finger-  
102 prints and on PaDEL descriptors.

103 For MolPrint2D fingerprints chemical similarity between two compounds  $a$  and  $b$  is  
104 expressed as the proportion between atom environments common in both structures  
105  $A \cap B$  and the total number of atom environments  $A \cup B$  (Jaccard/Tanimoto index).

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

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

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

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

- 112 • First a similarity threshold of 0.5 is used to collect neighbours, to create a local  
113 QSAR model and to make a prediction for the query compound. This are predic-  
114 tions with *high confidence*.
- 115 • If any of these steps fails, the procedure is repeated with a similarity threshold  
116 of 0.2 and the prediction is flagged with a warning that it might be out of the  
117 applicability domain of the training data (*low confidence*).
- 118 • Similarity thresholds of 0.5 and 0.2 are the default values chosen by the software  
119 developers and remained unchanged during the course of these experiments.

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

## 122 Local QSAR models and predictions

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

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

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

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

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

### 131 **Applicability domain**

132 The applicability domain (AD) of **lazar** models is determined by the structural diver-  
133 sity of the training data. If no similar compounds are found in the training data no  
134 predictions will be generated. Warnings are issued if the similarity threshold had to be  
135 lowered from 0.5 to 0.2 in order to enable predictions. Predictions without warnings  
136 can be considered as close to the applicability domain (*high confidence*) and predictions  
137 with warnings as more distant from the applicability domain (*low confidence*). Quantita-  
138 tive applicability domain information can be obtained from the similarities of individual  
139 neighbours.

### 140 **Availability**

- 141 • **lazar** experiments for this manuscript: <https://git.in-silico.ch/mutagenicity-paper>  
142 (source code, GPL3)
- 143 • **lazar** framework: <https://git.in-silico.ch/lazar> (source code, GPL3)
- 144 • **lazar** GUI: <https://git.in-silico.ch/lazar-gui> (source code, GPL3)
- 145 • Public web interface: <https://lazar.in-silico.ch>

### 146 **R Random Forest, Support Vector Machines, and Deep Learning**

147 The RF, SVM, and DL models were generated using the R software (R-project for  
148 Statistical Computing, <https://www.r-project.org/>; version 3.3.1), specific R packages  
149 used are identified for each step in the description below.

## 150 **Random Forest**

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

## 153 **Support Vector Machines**

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

## 156 **Deep Learning**

157 The DL model was generated using the ‘*h2o.deeplearning*’-function (package ‘*h2o*’). The  
158 DL contained four hidden layer with 70, 50, 50, and 10 neurons, respectively. Other  
159 hyperparameter were set as follows:  $l1=1.0E-7$ ,  $l2=1.0E-11$ ,  $\epsilon = 1.0E-10$ ,  $\rho =$   
160  $0.8$ , and  $\text{quantile\_alpha} = 0.5$ . For all other hyperparameter, the default values were  
161 used. Weights and biases were in a first step determined with an unsupervised DL model.  
162 These values were then used for the actual, supervised DL model.

163 TODO: **Verena** kannst Du bitte ueberpruefen, ob das noch stimmt und ggf die Figure  
164 1 anpassen

165 To validate these models, an internal cross-validation approach was chosen. The training  
166 dataset was randomly split in training data, which contained 95% of the data, and  
167 validation data, which contain 5% of the data. A feature selection with LASSO on the  
168 training data was performed, reducing the number of descriptors to approximately 100.  
169 This step was repeated five times. Based on each of the five different training data,  
170 the predictive models were trained and the performance tested with the validation data.  
171 This step was repeated 10 times.

## 172 **Applicability domain**



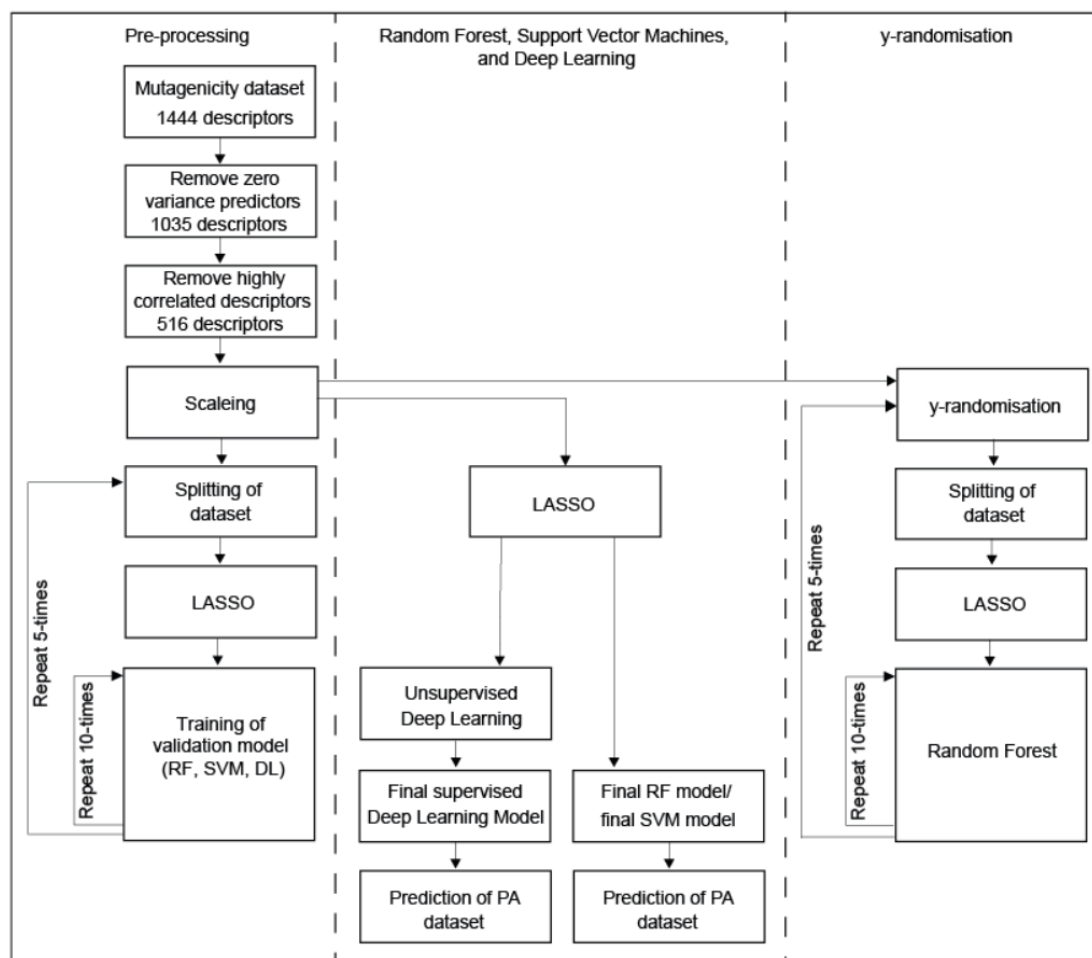


Figure 1: Flowchart of the generation and validation of the models generated in R-project

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

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

## 182 **Availability**

183 R scripts for these experiments can be found in [https://git.in-silico.ch/mutagenicity-](https://git.in-silico.ch/mutagenicity-paper/scripts/R)  
184 [paper/scripts/R](https://git.in-silico.ch/mutagenicity-paper/scripts/R).

## 185 **Tensorflow models**

186 TODO: **Philipp** bitte ergaenzen

## 187 **Logistic regression (SGD)**

## 188 **Logistic regression (scikit)**

## 189 **Random forests**

## 190 **Deep Learning**

191 Alternatively, a DL model was established with Python-based Tensorflow program ([https:](https://www.tensorflow.org/)  
192 [//www.tensorflow.org/](https://www.tensorflow.org/)) using the high-level API Keras (<https://www.tensorflow.org/>

193 guide/keras) to build the models.

194 Tensorflow models used the same PaDEL descriptors as the R models.

195 Data pre-processing was done by rank transformation using the ‘*QuantileTransformer*’  
196 procedure. A sequential model has been used. Four layers have been used: input layer,  
197 two hidden layers (with 12, 8 and 8 nodes, respectively) and one output layer. For the  
198 output layer, a sigmoidal activation function and for all other layers the ReLU (‘*Rectified*  
199 *Linear Unit*’) activation function was used. Additionally, a  $L^2$ -penalty of 0.001 was used  
200 for the input layer. For training of the model, the ADAM algorithm was used to minimise  
201 the cross-entropy loss using the default parameters of Keras. Training was performed  
202 for 100 epochs with a batch size of 64. The model was implemented with Python 3.6  
203 and Keras.

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

## 206 **Validation**

207 10-fold cross-validation was used for all Tensorflow models.

## 208 **Availability**

209 Jupyter notebooks for these experiments can be found in [https://git.in-silico.ch/mutagenicity-](https://git.in-silico.ch/mutagenicity-paper/scripts/tensorflow)  
210 [paper/scripts/tensorflow](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 lazarus 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	0.64	0.62	0.63	0.63
True positive rate/Sensitivity	0.59	0.6	0.62	0.61
True negative rate/Specificity	0.7	0.65	0.63	0.64
Positive predictive value/Precision	0.66	0.63	0.62	0.63
Negative predictive value	0.63	0.62	0.63	0.63
Nr. predictions	8080	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

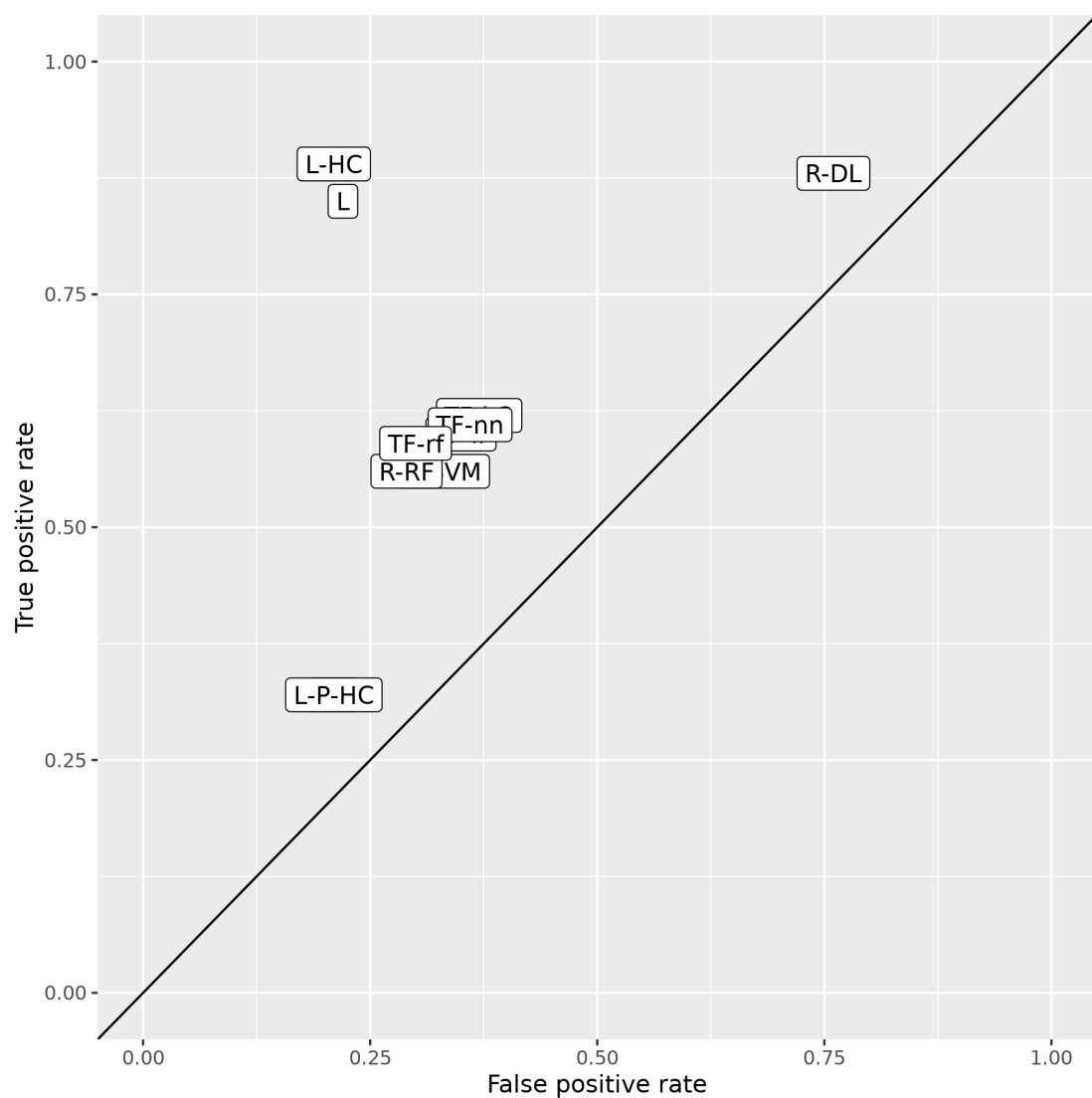


Figure 2: ROC plot of crossvalidation results. *R-RF*: R Random Forests, *R-SVM*: R Support Vector Machines, *R-DL*: R Deep Learning, *TF*: Tensorflow without feature selection, *TF-FS*: Tensorflow with feature selection, *L*: lazar, *L-HC*: lazar high confidence predictions, *L-P*: lazar with PaDEL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

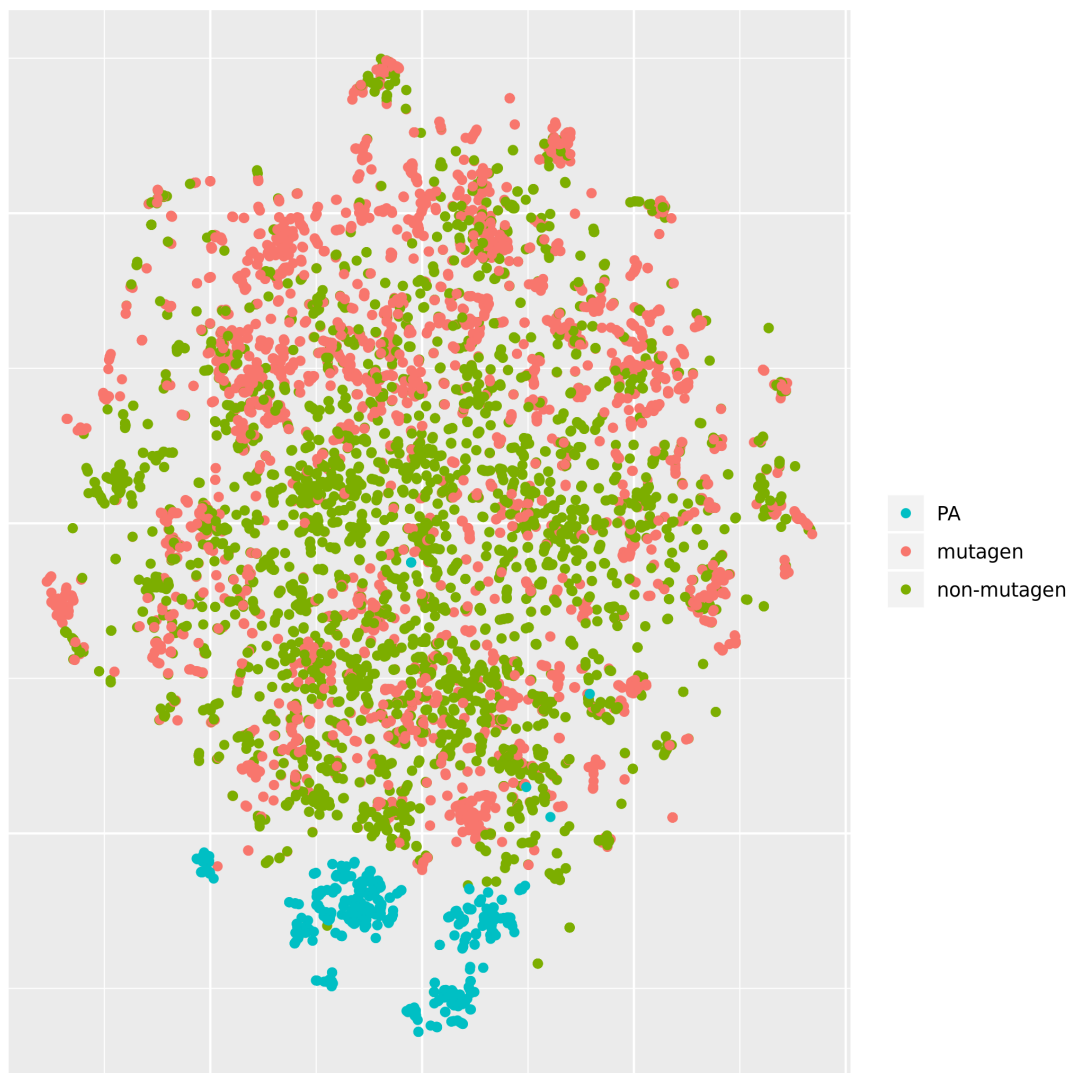


Figure 3: t-sne visualisation of mutagenicity training data and pyrrolizidine alkaloids (PA)

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

## 233 Discussion

### 234 Data

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

### 240 Model performance

241 Table ?? and Figure 2 show that the standard **lazar** algorithm (with MP2D fingerprints)  
242 give the most accurate crossvalidation results. R Random Forests, Support Vector Ma-  
243 chines and Tensorflow models have similar accuracies with balanced sensitivity (true  
244 position rate) and specificity (true negative rate). **lazar** models with PaDEL descrip-  
245 tors have low sensitivity and R Deep Learning models have low specificity.

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

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



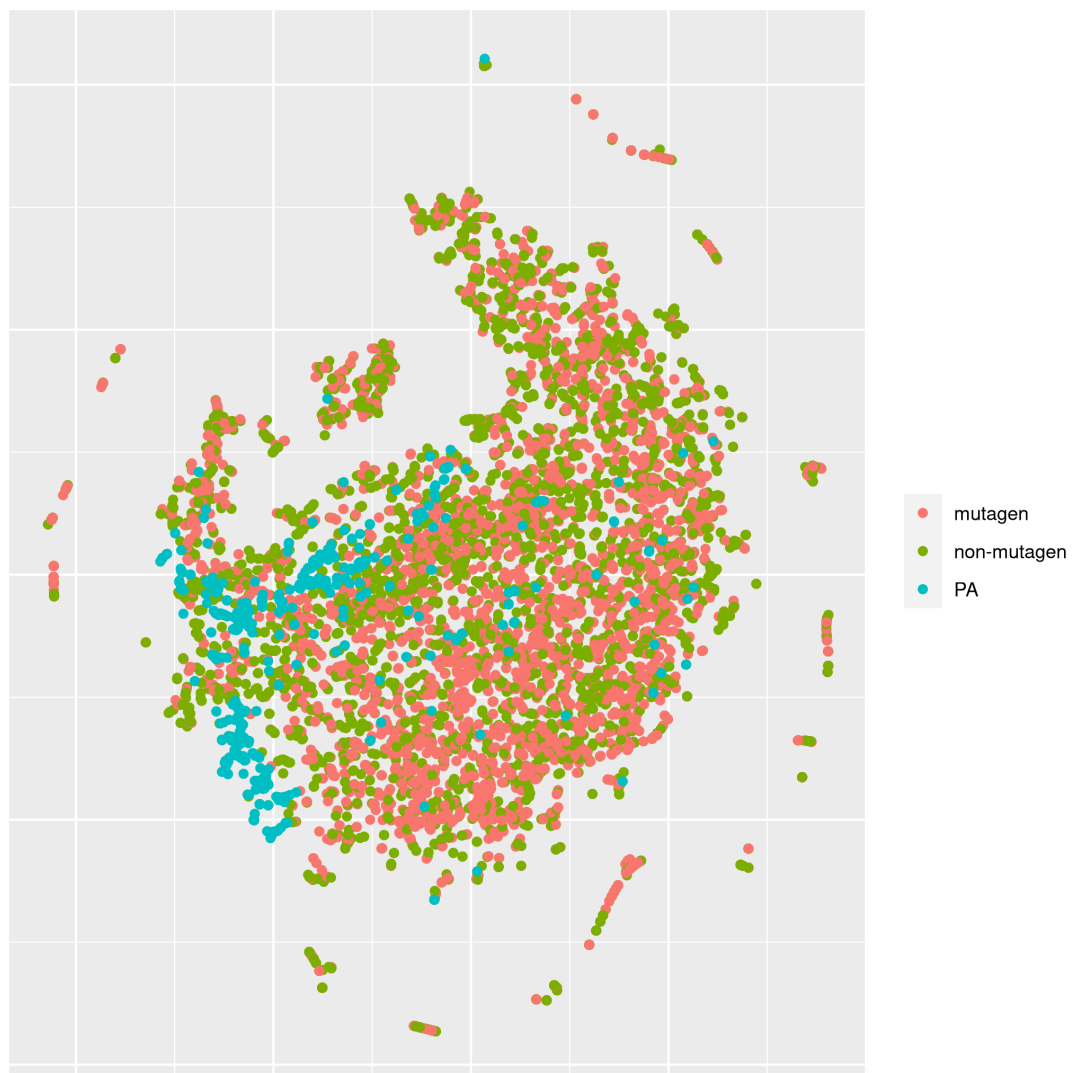


Figure 4: t-sne visualisation of mutagenicity training data and pyrrolizidine alkaloids (PA)

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 not 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

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

## 306 **Mutagenicity of PAs**

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

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

### 311 Necic acid

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

### 319 Necine base

320 The rank order of necine base is comparable in LAZAR, RF, and DL (R-project) models:  
321 with platynecine being less or as genotoxic as retronecine, and otonecine being the most  
322 genotoxic. In the Tensorflow-generate DL model, platynecine also has the lowest geno-  
323 toxic probability, but are then followed by the otonecines and last by retronecine. These  
324 results partly correspond to earlier published studies. Saturated PAs of the platynecine-  
325 type are generally accepted to be less or non-toxic and have been shown in *in vitro*  
326 experiments to form no DNA-adducts Xia et al. 2013(). Therefore, it is striking, that

327 1,2-unsaturated PAs of the retronecine-type should have an almost comparable genotoxic  
328 potential in the LAZAR and DL (R-project) model. In literature, otonecine-type PAs  
329 were shown to be more toxic than those of the retronecine-type Li et al. 2013().

330 Modifications of necine base

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

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

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

350 In the DL (R-project) and RF model, *N*-oxides have a by far more genotoxic potential  
351 than tertiary PAs or dehydropyrrolizidines. As PA *N*-oxides are easily conjugated for  
352 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

379 PAs, but in general, appeared to have a low sensitivity. The pre-incubation phase  
380 for metabolic activation of PAs by microsomal enzymes was the sensitivity-limiting  
381 step. This could very well mean that this is also reflected in the QSAR models.

## 382 Conclusions

383 A new public *Salmonella* mutagenicity training dataset with 8309 compounds was cre-  
384 ated and used it to train **lazar**, R and Tensorflow models. The best performance was  
385 obtained with **lazar** models using MolPrint2D descriptors, with prediction accuracies  
386 comparable to the interlaboratory variability of the Ames test. Differences between al-  
387 gorithms (local vs. global models) and/or descriptors (MolPrint2D vs PaDEL) may be  
388 responsible for the different prediction accuracies.

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

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

400 In further studies, additional machine learning techniques or a modified (extended) train-  
401 ing dataset should be used for an additional attempt to predict the genotoxic potential  
402 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.