

1 A comparison of nine machine learning models based on an  
2 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 Random forest, support vector machine, logistic regression, neural  
11 networks and k-nearest neighbor (**lazar**) algorithms, were applied to new  
12 *Salmonella* mutagenicity dataset with 8309 unique chemical structures. The  
13 best prediction accuracies in 10-fold-crossvalidation were obtained with  
14 **lazar** models and MolPrint2D descriptors, that gave accuracies (84%)  
15 similar to the interlaboratory variability of the Ames test.

16 **TODO:** PA results

## 17 Introduction

18 **TODO:** rationale for investigation

19 The main objectives of this study were

- to generate a new mutagenicity training dataset, by combining the most comprehensive public datasets
- to compare the performance of MolPrint2D (*MP2D*) fingerprints with PaDEL descriptors
- to compare the performance of global QSAR models (random forests (*RF*), support vector machines (*SVM*), logistic regression (*LR*), neural nets (*NN*)) with local models (**lazar**)
- 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>.

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

The testing dataset consisted of 602 different PAs.

**TODO: Verena** Kannst Du kurz die Quellen und Auswahlkriterien zusammenfassen?

The compilation of the PA dataset is described in detail in Schöning et al. (2017).

## **Descriptors**

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

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.

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

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

### 73 **PaDEL descriptors**

74 Molecular 1D and 2D descriptors were calculated with the PaDEL-Descriptors program  
75 (<http://www.yapcwsoft.com> version 2.21, Yap (2011)).

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

80 During feature selection, descriptors with near zero variance were removed using ‘*NearZe-*  
81 *ro Var*’-function (package ‘caret’). If the percentage of the most common value was more  
82 than 90% or when the frequency ratio of the most common value to the second most  
83 common value was greater than 95:5 (e.g. 95 instances of the most common value and  
84 only 5 or less instances of the second most common value), a descriptor was classified  
85 as having a near zero variance. After that, highly correlated descriptors were removed  
86 using the ‘*findCorrelation*’-function (package ‘caret’) with a cut-off of 0.9. This resulted  
87 in a training dataset with 516 descriptors. These descriptors were scaled to be in the  
88 range between 0 and 1 using the ‘*preProcess*’-function (package ‘caret’). The scaling  
89 routine was saved in order to apply the same scaling on the testing dataset. As these  
90 three steps did not consider the dependent variable (experimental mutagenicity), it was

91 decided that they do not need to be included in the cross-validation of the model. To  
92 further reduce the number of features, a LASSO (*least absolute shrinkage and selection*  
93 *operator*) regression was performed using the ‘*glmnet*’-function (package ‘*glmnet*’). The  
94 reduced dataset was used for the generation of the pre-trained models.

95 PaDEL descriptors were used in global (RF, SVM, LR, NN) and local (**lazar**) models.

## 96 **Algorithms**

### 97 **lazar**

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

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

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

106 Apart from this basic workflow, **lazar** is completely modular and allows the researcher to  
107 use arbitrary algorithms for similarity searches and local QSAR (*Quantitative structure–*  
108 *activity relationship*) modelling. Algorithms used within this study are described in the  
109 following sections.

### 110 **Neighbour identification**

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

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

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

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

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

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

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

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

## 132 Local QSAR models and predictions

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

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

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

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

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

## 141 Applicability domain

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

## 150 Availability

- 151 • **lazar** experiments for this manuscript: <https://git.in-silico.ch/mutagenicity-paper>  
152 (source code, GPL3)

- 153 • **lazar** framework: <https://git.in-silico.ch/lazar> (source code, GPL3)
- 154 • **lazar** GUI: <https://git.in-silico.ch/lazar-gui> (source code, GPL3)
- 155 • Public web interface: <https://lazar.in-silico.ch>

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

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

### 160 **Random Forest (*RF*)**

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

### 163 **Support Vector Machines (*SVM*)**

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

166 **TODO: Verena, Phillip** Sollen wir die DL Modelle ebenso wie die Tensorflow als  
167 Neural Nets (NN) bezeichnen?

### 168 **Deep Learning**

169 The DL model was generated using the ‘*h2o.deeplearning*’-function (package ‘*h2o*’). The  
170 DL contained four hidden layer with 70, 50, 50, and 10 neurons, respectively. Other  
171 hyperparameter were set as follows:  $l1=1.0E-7$ ,  $l2=1.0E-11$ ,  $\epsilon = 1.0E-10$ ,  $\rho =$   
172  $0.8$ , and  $\text{quantile\_alpha} = 0.5$ . For all other hyperparameter, the default values were



173 used. Weights and biases were in a first step determined with an unsupervised DL model.  
174 These values were then used for the actual, supervised DL model.

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

182 **TODO: Verena** kannst Du bitte ueberpruefen, ob das noch stimmt und ggf die Figure  
183 1 anpassen

#### 184 **Applicability domain**

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

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

#### 194 **Availability**

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

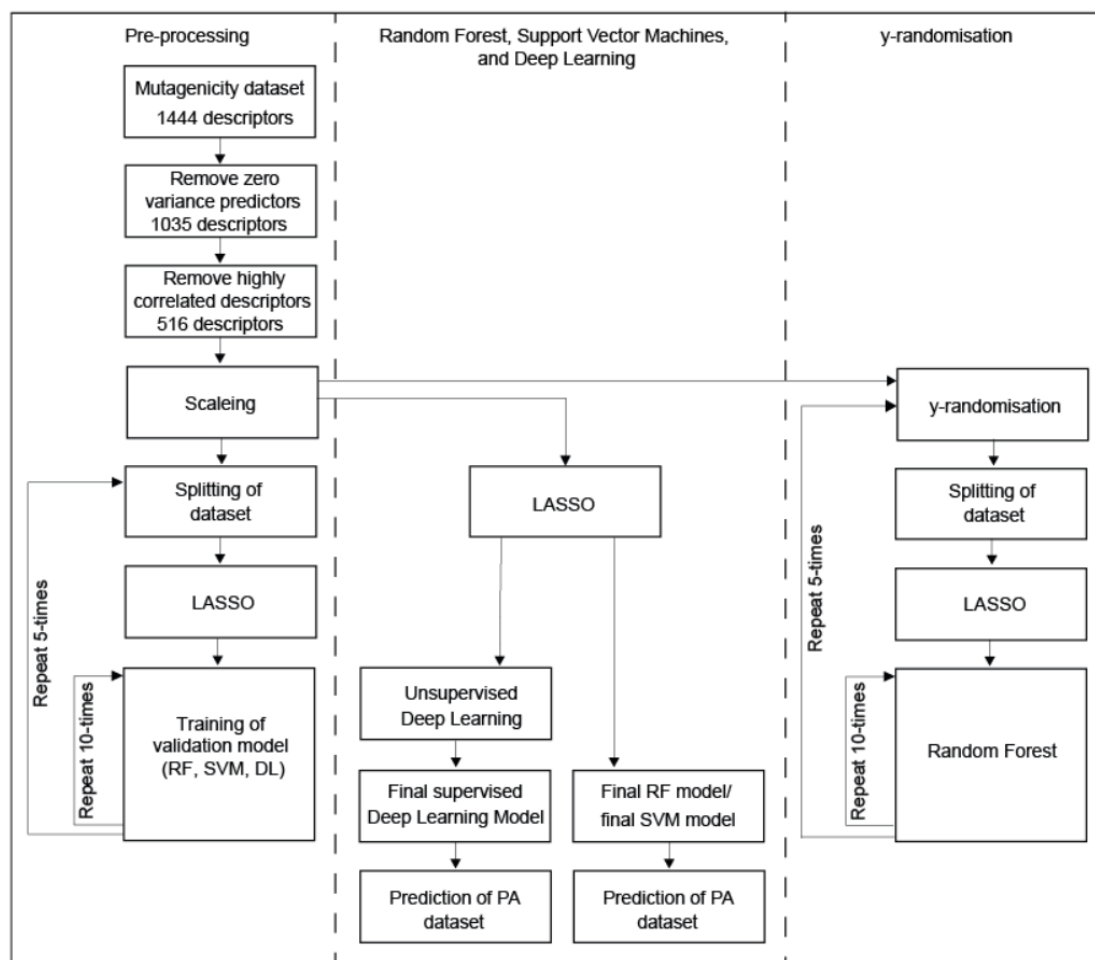


Figure 1: Flowchart of the generation and validation of the models generated in R-project

## 197 **Tensorflow models**

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

207 **TODO: Philipp** Ich hab die alten Ergebnisse mit feature selection weggelassen, ist das  
208 ok? Dann muesste auch dieser Absatz gestrichen werden, oder?

209 **TODO: Philipp** Kannst Du bitte die folgenden Absaetze ergaenzen

## 210 **Random forests (*RF*)**

## 211 **Logistic regression (SGD) (*LR-sgd*)**

## 212 **Logistic regression (scikit) (*LR-scikit*)**

213 **TODO: Philipp, Verena** DL oder NN?

## 214 **Neural Nets (*NN*)**

215 Alternatively, a DL model was established with Python-based Tensorflow program ([https:](https://www.tensorflow.org/)  
216 [//www.tensorflow.org/](https://www.tensorflow.org/)) using the high-level API Keras ([https://www.tensorflow.org/](https://www.tensorflow.org/guide/keras)  
217 [guide/keras](https://www.tensorflow.org/guide/keras)) to build the models.

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

## 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/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 standard **lazar**

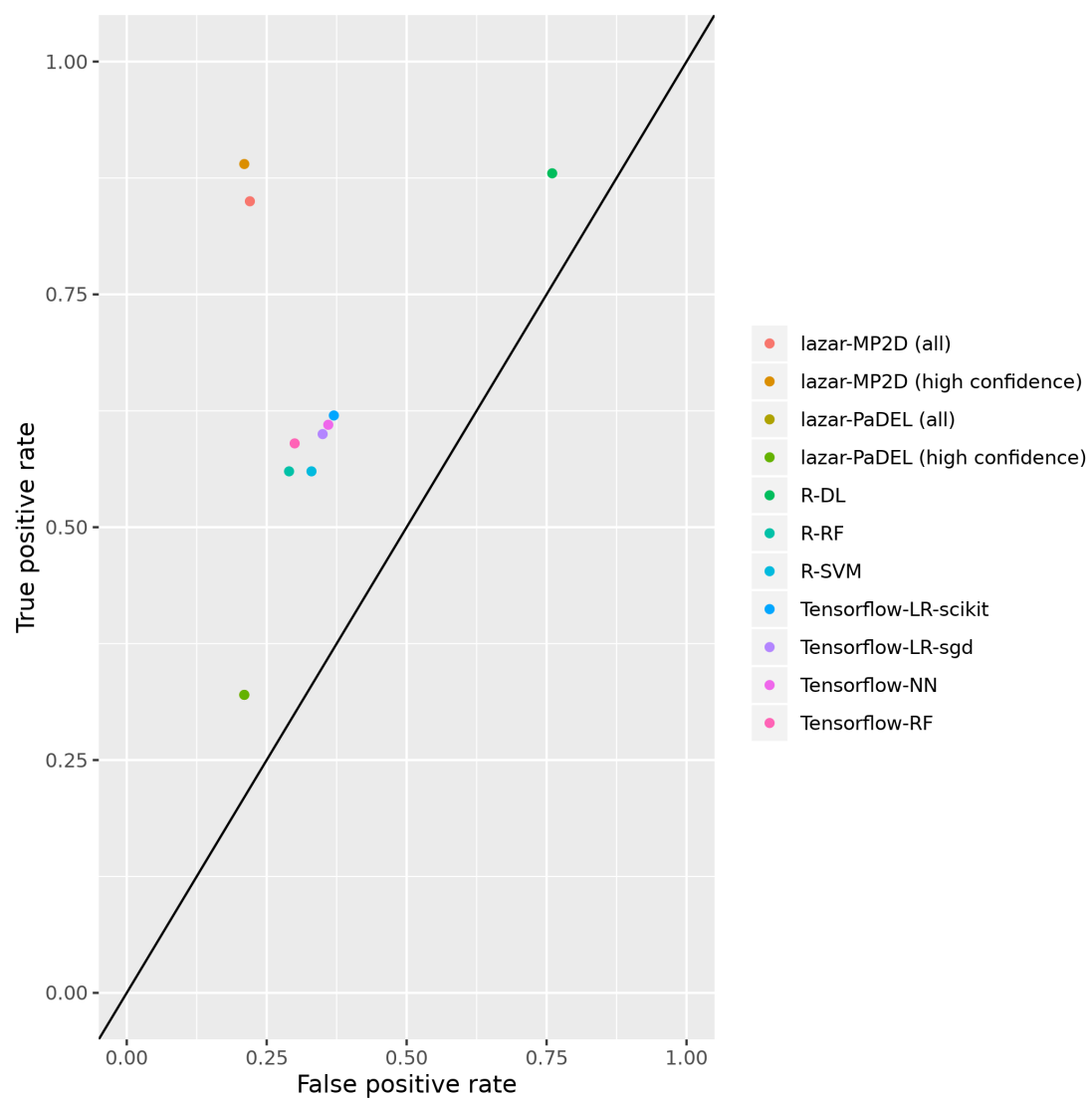


Figure 2: ROC plot of crossvalidation results.

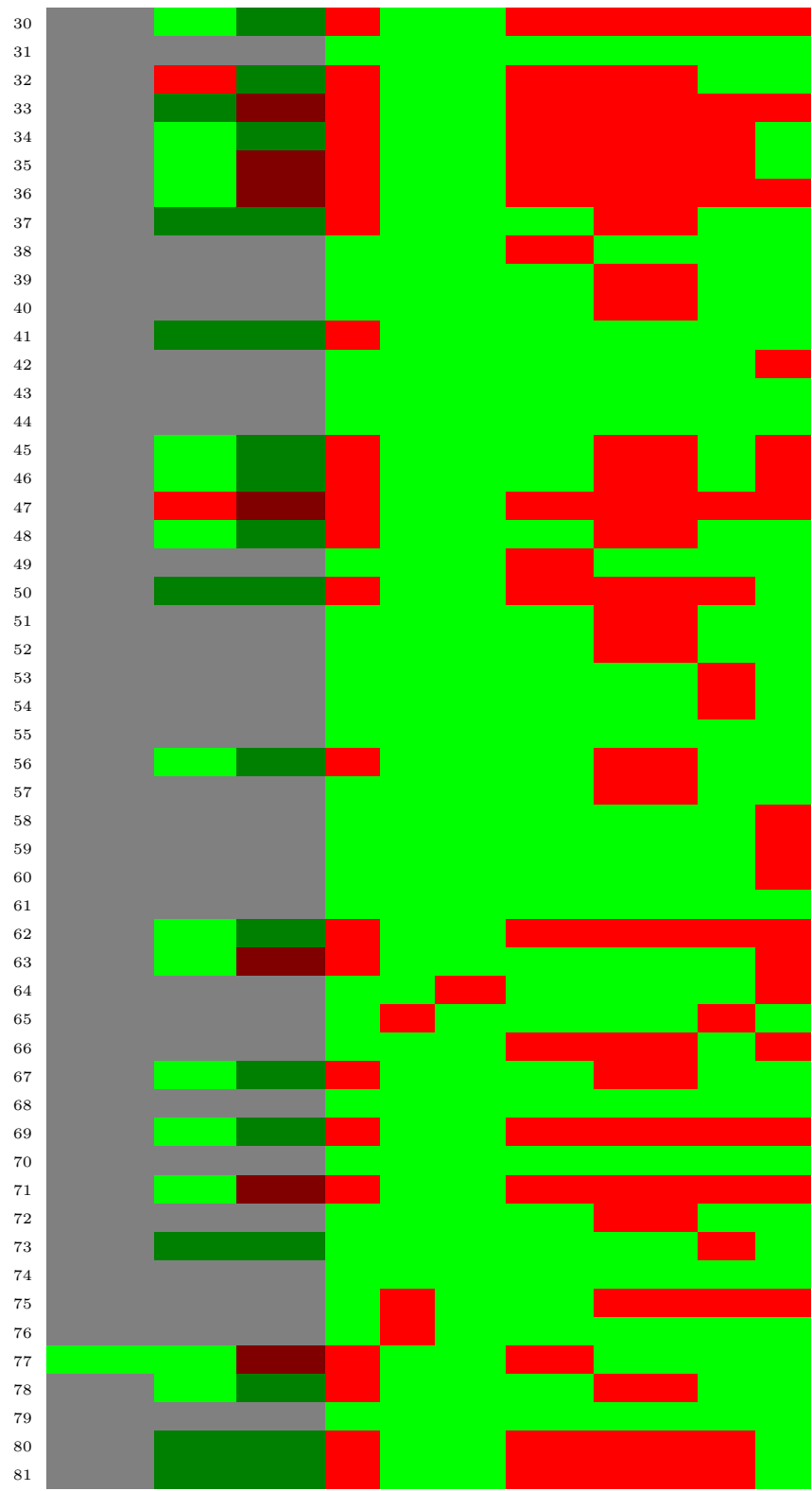
models using MolPrint2D descriptors (0.84 for predictions with high confidence, 0.82 for all predictions). Models utilizing PaDEL descriptors have generally lower accuracies ranging from 0.56 (R deep learning) to 0.64 (R/Tensorflow random forests). Sensitivity and specificity is generally well balanced with the exception of **lazar**-PaDEL (low sensitivity) and R deep learning (low specificity) models.

## 241 Pyrrolizidine alkaloid mutagenicity predictions

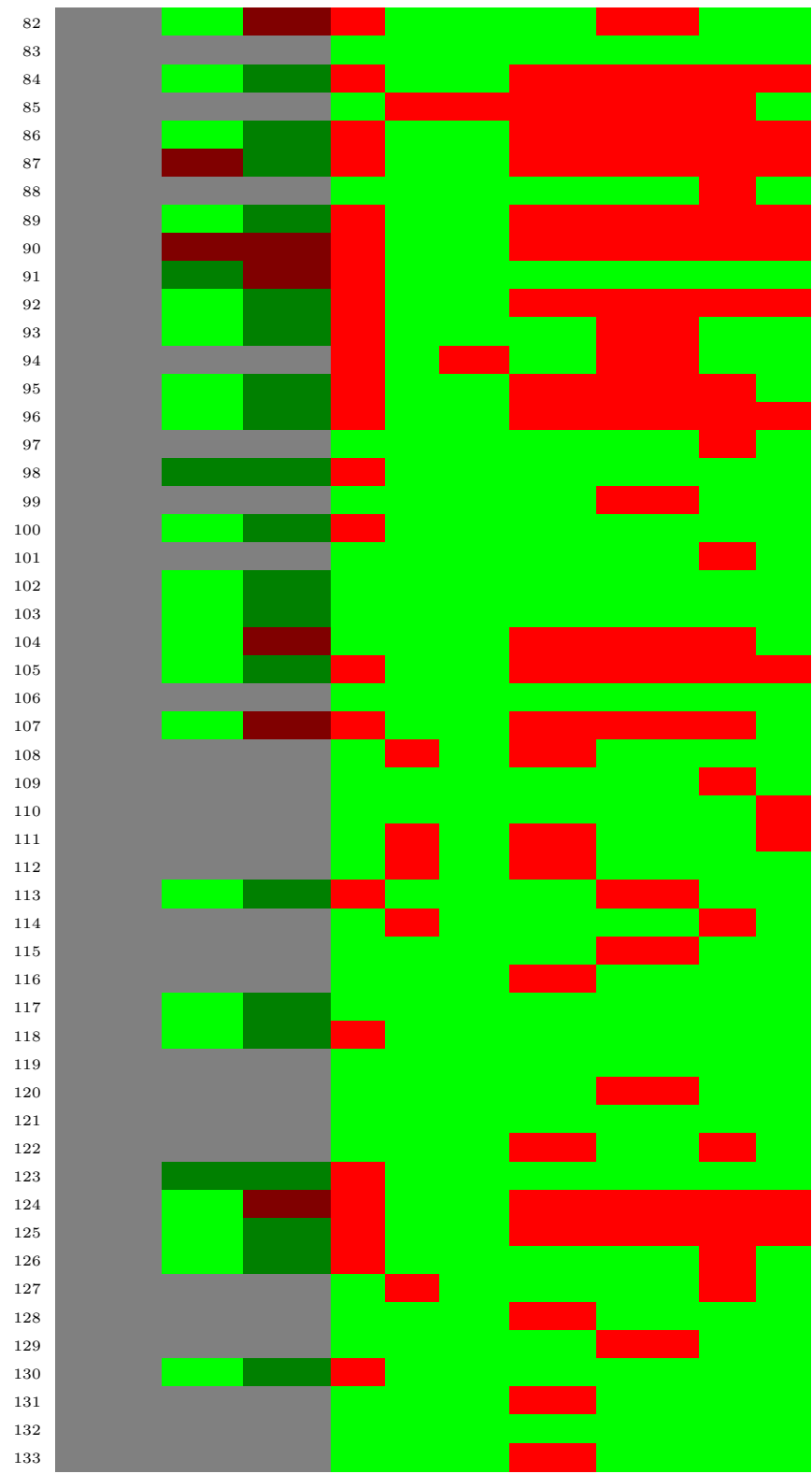
242 Mutagenicity predictions from all investigated models for 602 pyrrolizidine alkaloids are  
243 summarized in Table 4.

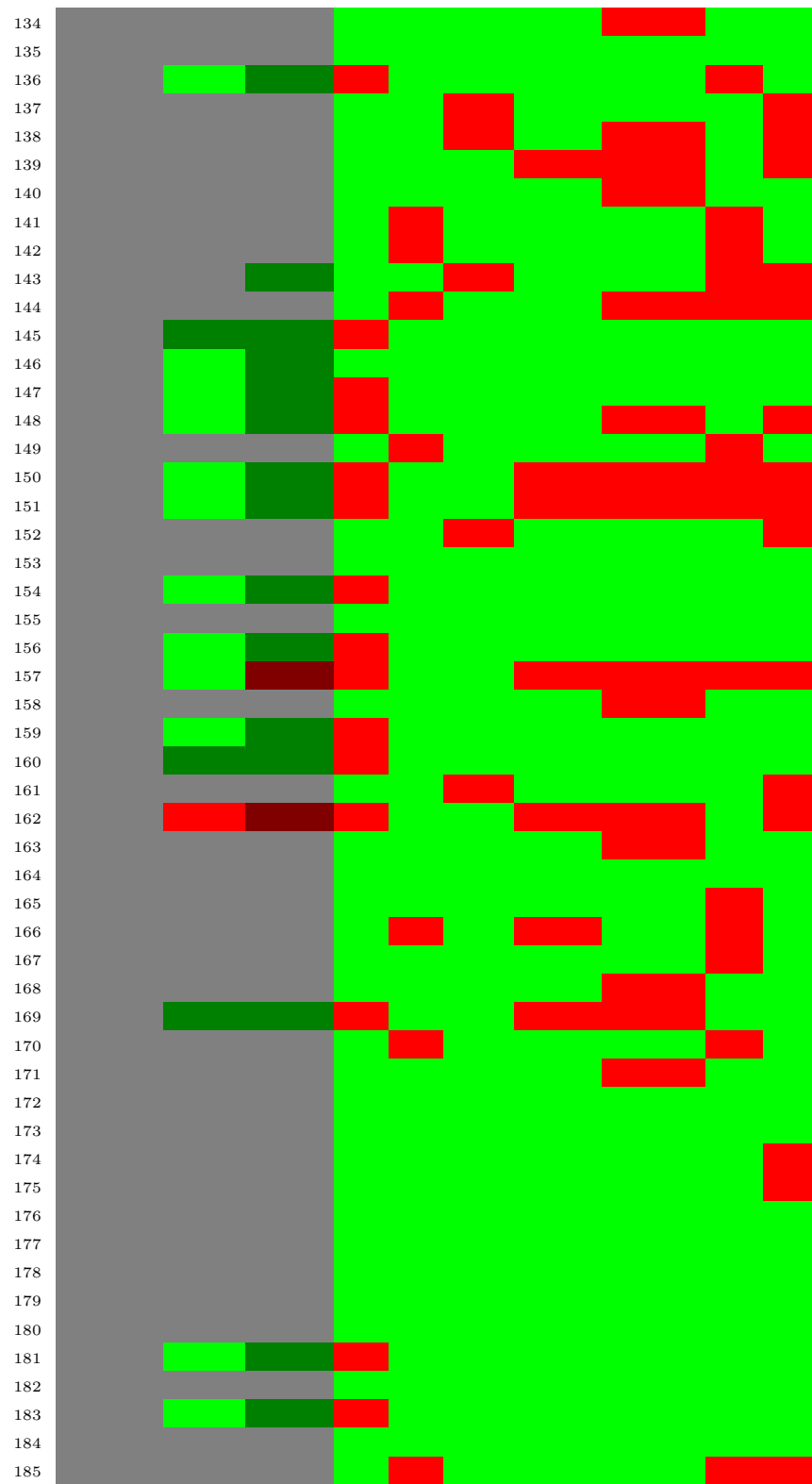
Table 4: Summary of pyrrolizidine alkaloid predictions: red: mutagen, green: non-mutagen, grey: no prediction, dark red/green: low confidence

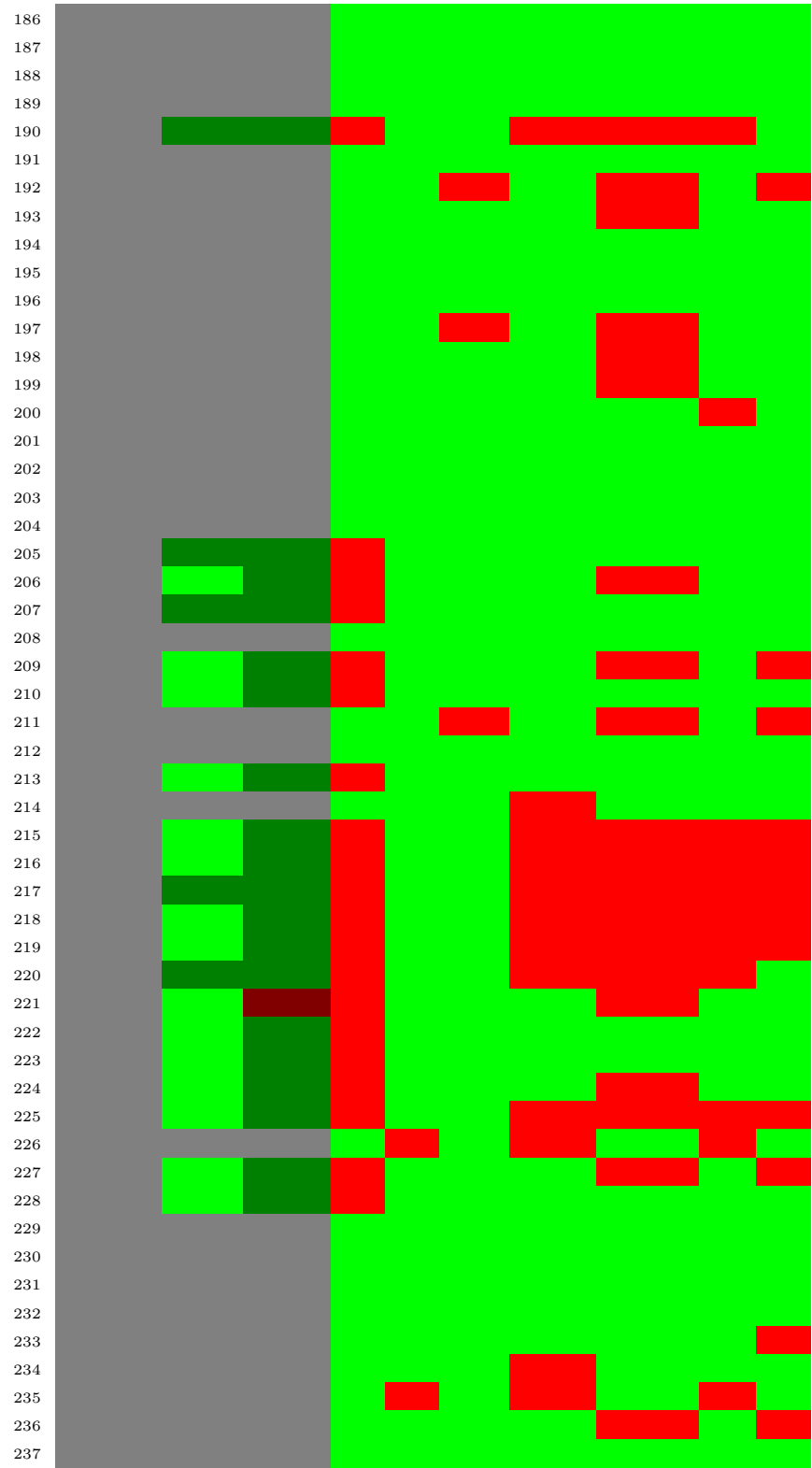
ID	Measured	lazar		R			Tensorflow			
		MP2D	PaDEL	DL	RF	SVM	LR-sgd	LR-scikit	NN	RF
1	grey	green	darkgreen	red	red	green	red	red	red	red
2		red	darkgreen	red	red	green	red	red	red	red
3		green	darkgreen	red	red	green	green	red	red	green
4		green	darkred	red	red	green	red	red	red	red
5		darkgreen	darkgreen	red	red	green	red	red	red	green
6		green	darkgreen	red	red	green	red	red	red	red
7		green	darkgreen	red	red	green	red	red	red	red
8		green	darkgreen	red	red	green	red	red	red	red
9		red	darkgreen	red	red	red	green	red	red	red
10		green	darkgreen	red	red	green	red	red	red	red
11		green	darkgreen	red	red	red	green	red	red	red
12		red	darkgreen	red	red	red	green	red	red	green
13		red	darkgreen	red	red	green	green	red	red	green
14		red	darkred	red	red	green	red	red	red	red
15		darkgreen	darkgreen	red	red	green	red	red	red	red
16		green	darkgreen	red	red	red	green	red	red	red
17	green	darkgreen	red	red	red	red	red	red	red	
18	grey	green	darkgreen	red	red	green	red	red	red	red
19		darkgreen	darkgreen	red	red	green	red	red	red	red
20		darkgreen	darkgreen	red	red	green	red	red	red	green
21		green	darkgreen	red	red	green	red	red	red	red
22		green	darkgreen	red	red	green	green	red	red	green
23		green	darkgreen	red	red	green	red	red	red	red
24		green	darkgreen	red	red	green	red	red	red	red
25		green	darkgreen	red	red	green	red	red	red	red
26		red	darkgreen	red	red	red	red	red	red	red
27		grey	grey	red	red	red	green	red	red	green
28		green	darkgreen	red	red	green	red	red	red	red
29		red	darkgreen	red	red	green	red	red	red	green

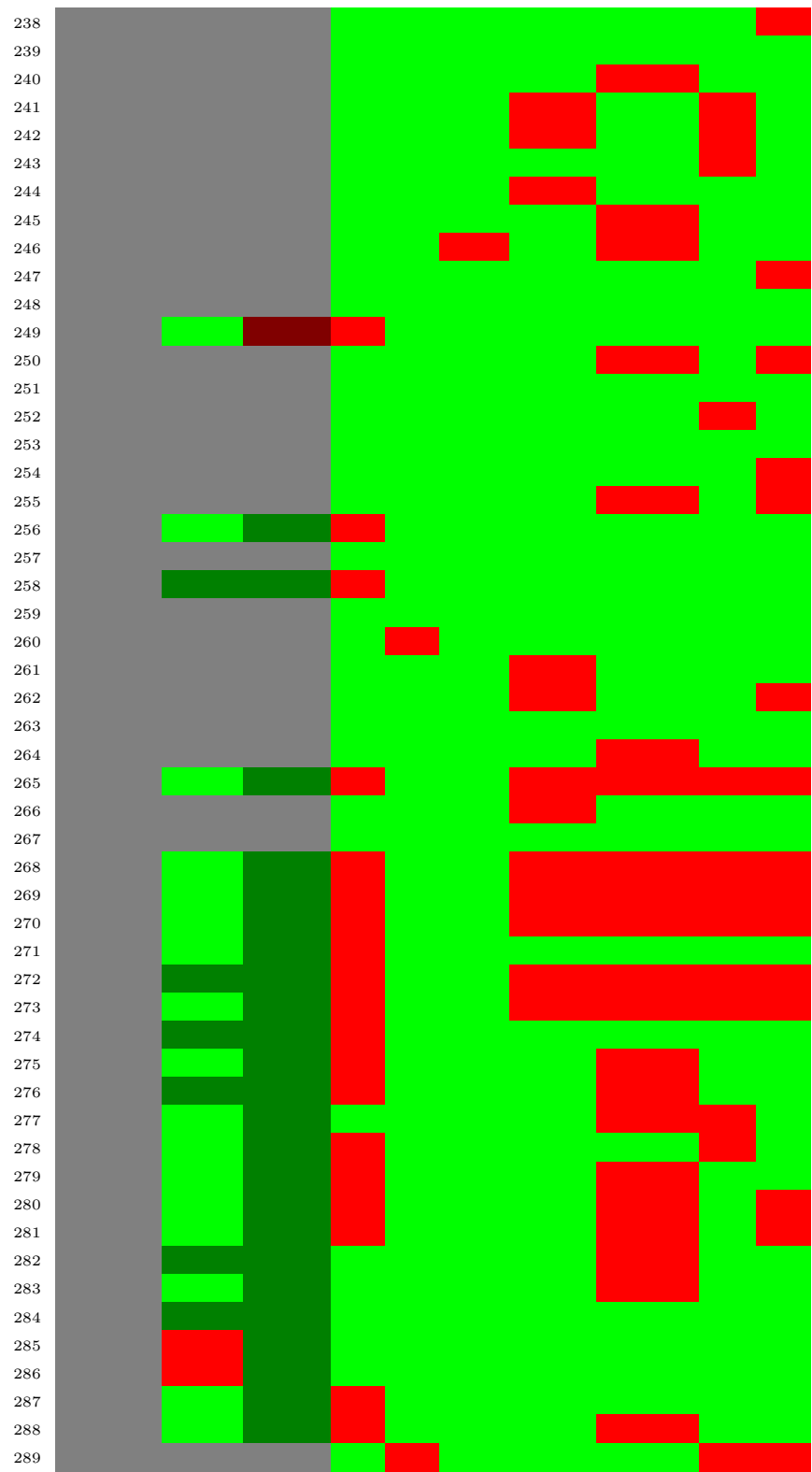








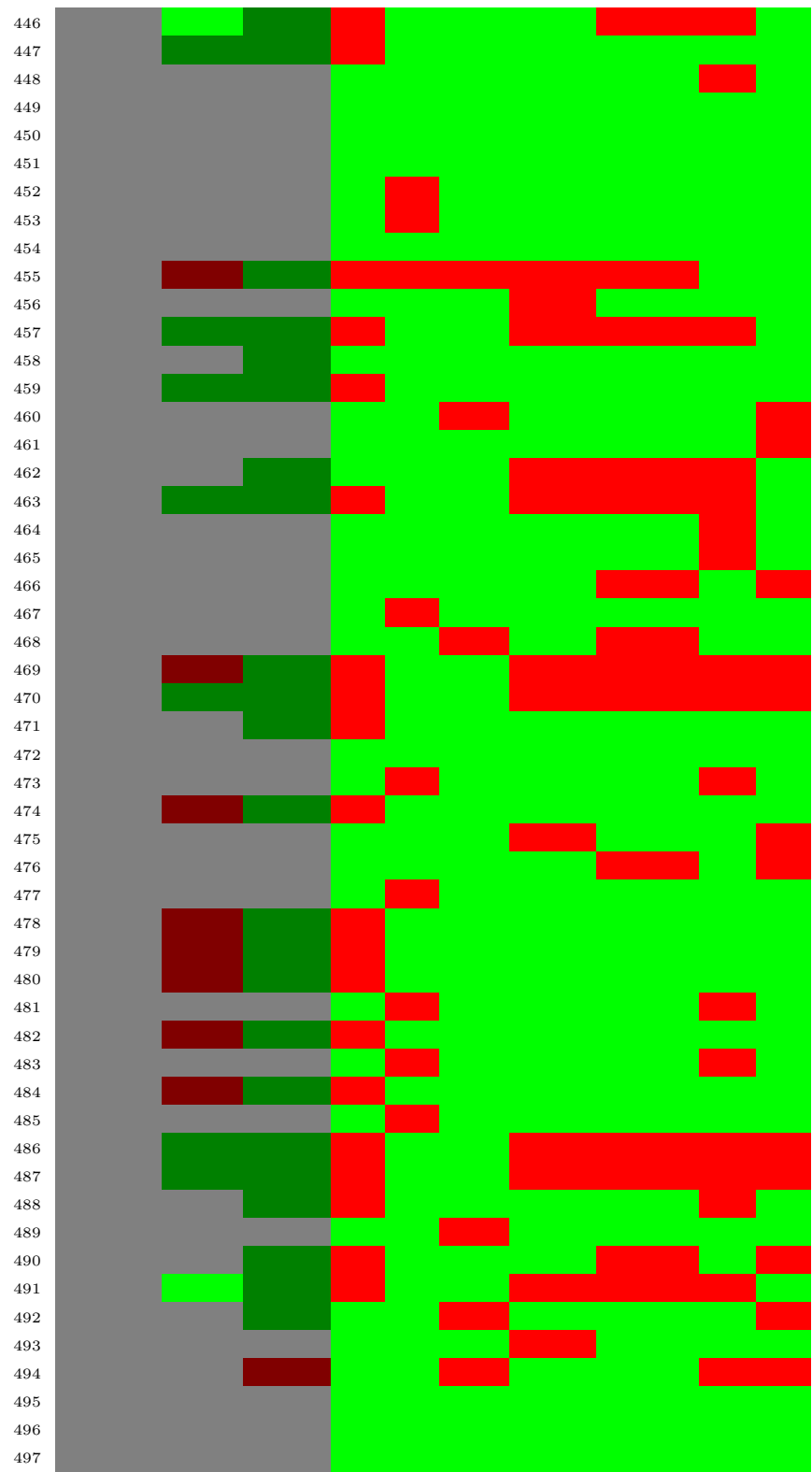




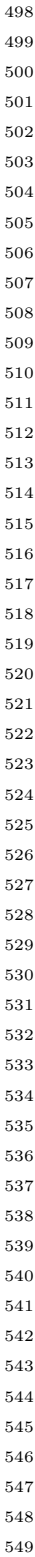


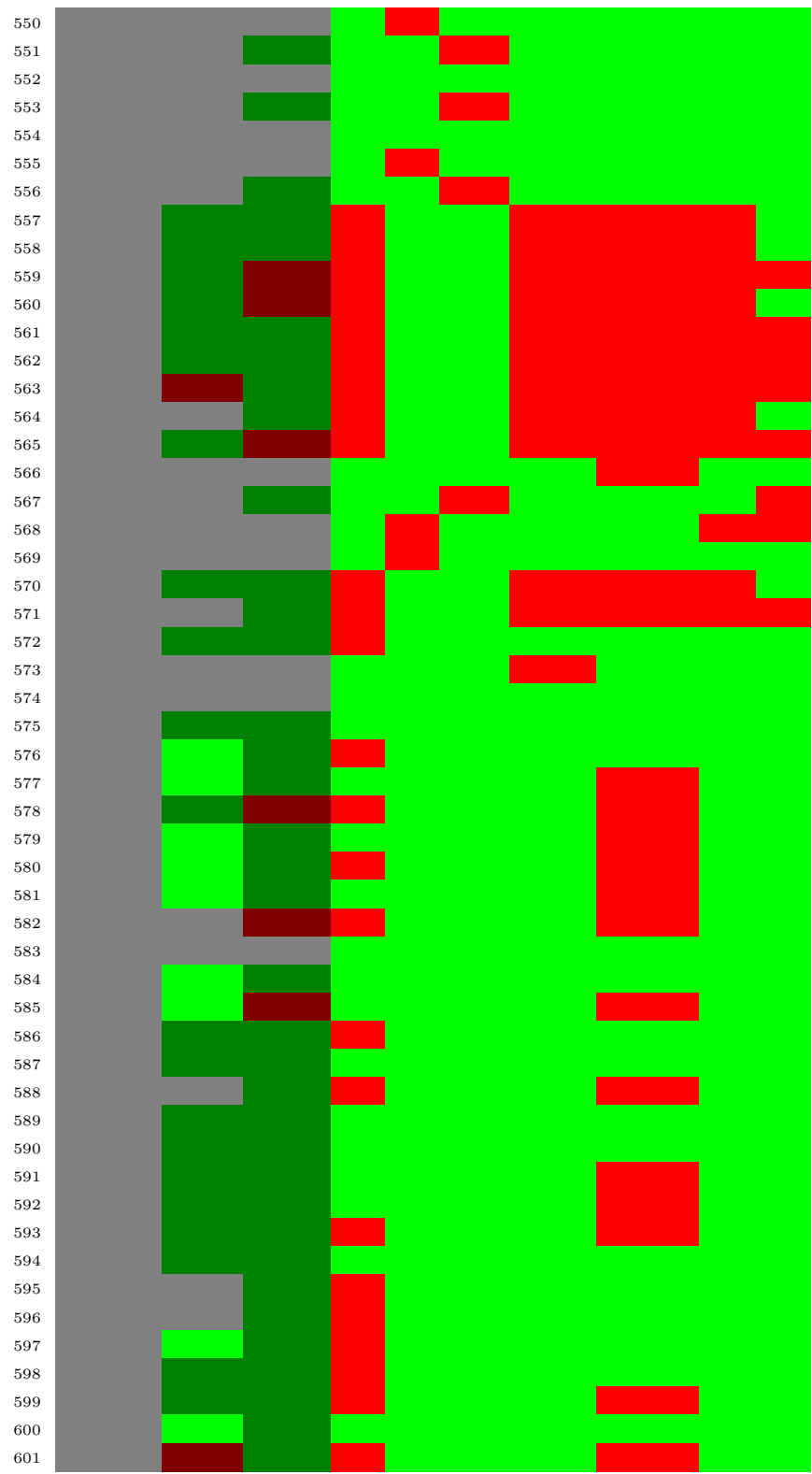












244 Training data and pyrrolizidine alkaloids were visualised with t-distributed stochastic  
245 neighbor embedding (t-SNE, Maaten and Hinton (2008)) for MolPrint2D and PaDEL  
246 descriptors. t-SNA maps each high-dimensional object (chemical) to a two-dimensional  
247 point. Similar objects are represented by nearby points and dissimilar objects are repre-  
248 sented by distant points.

249 Figure 3 shows the t-SNE of pyrrolizidine alkaloids (PA) and the mutagenicity training  
250 data in MP2D space (Tanimoto/Jaccard similarity).

251 Figure 4 shows the t-SNE of pyrrolizidine alkaloids (PA) and the mutagenicity training  
252 data in PaDEL space (Euclidean similarity).

## 253 Discussion

### 254 Data

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

### 260 Model performance

261 Table 1, Table 2, Table 3 and Figure 2 show that the standard **lazar** algorithm (with  
262 MP2D fingerprints) give the most accurate crossvalidation results. R Random Forests,  
263 Support Vector Machines and Tensorflow models have similar accuracies with balanced

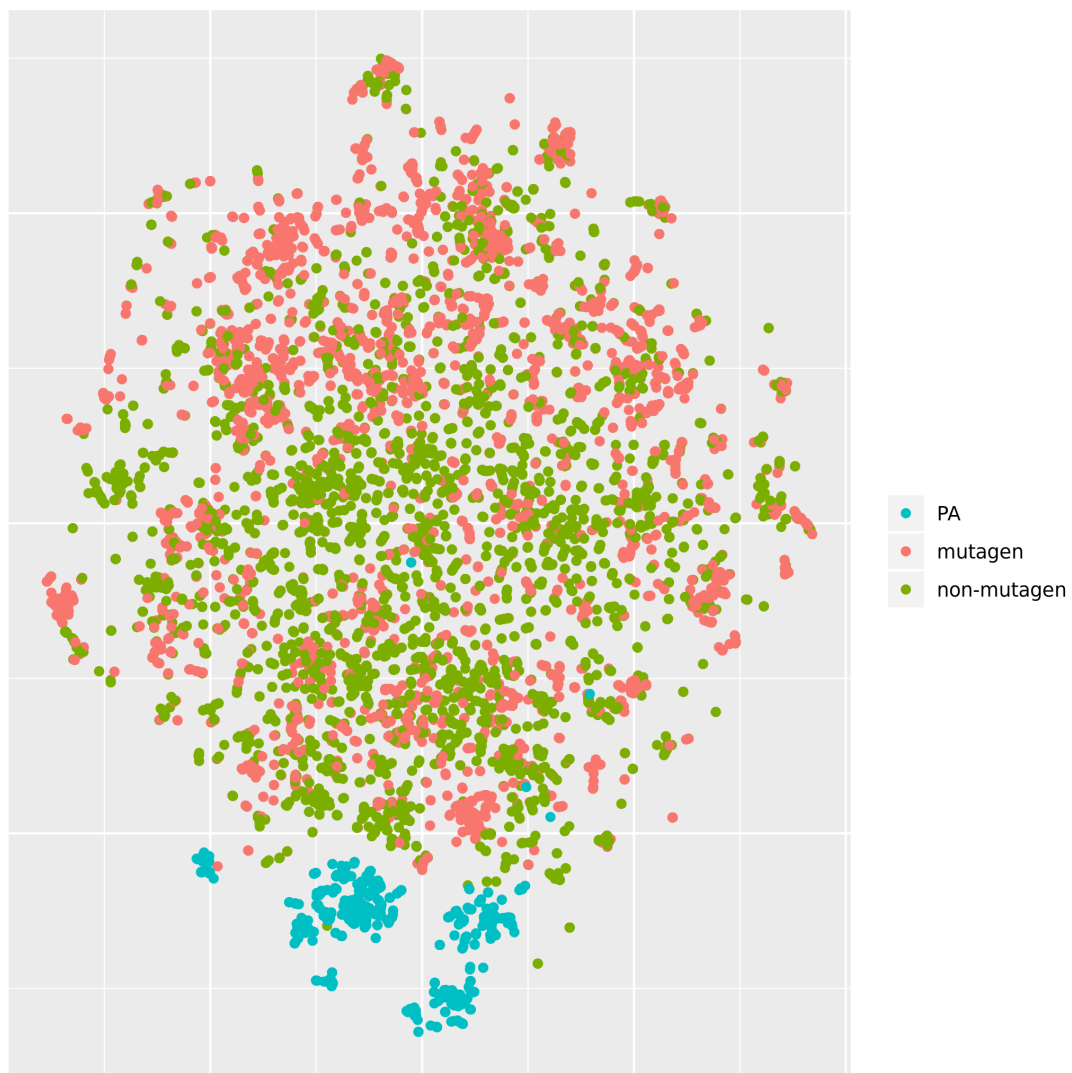


Figure 3: t-SNE visualisation of mutagenicity training data and pyrrolizidine alkaloids (PA)

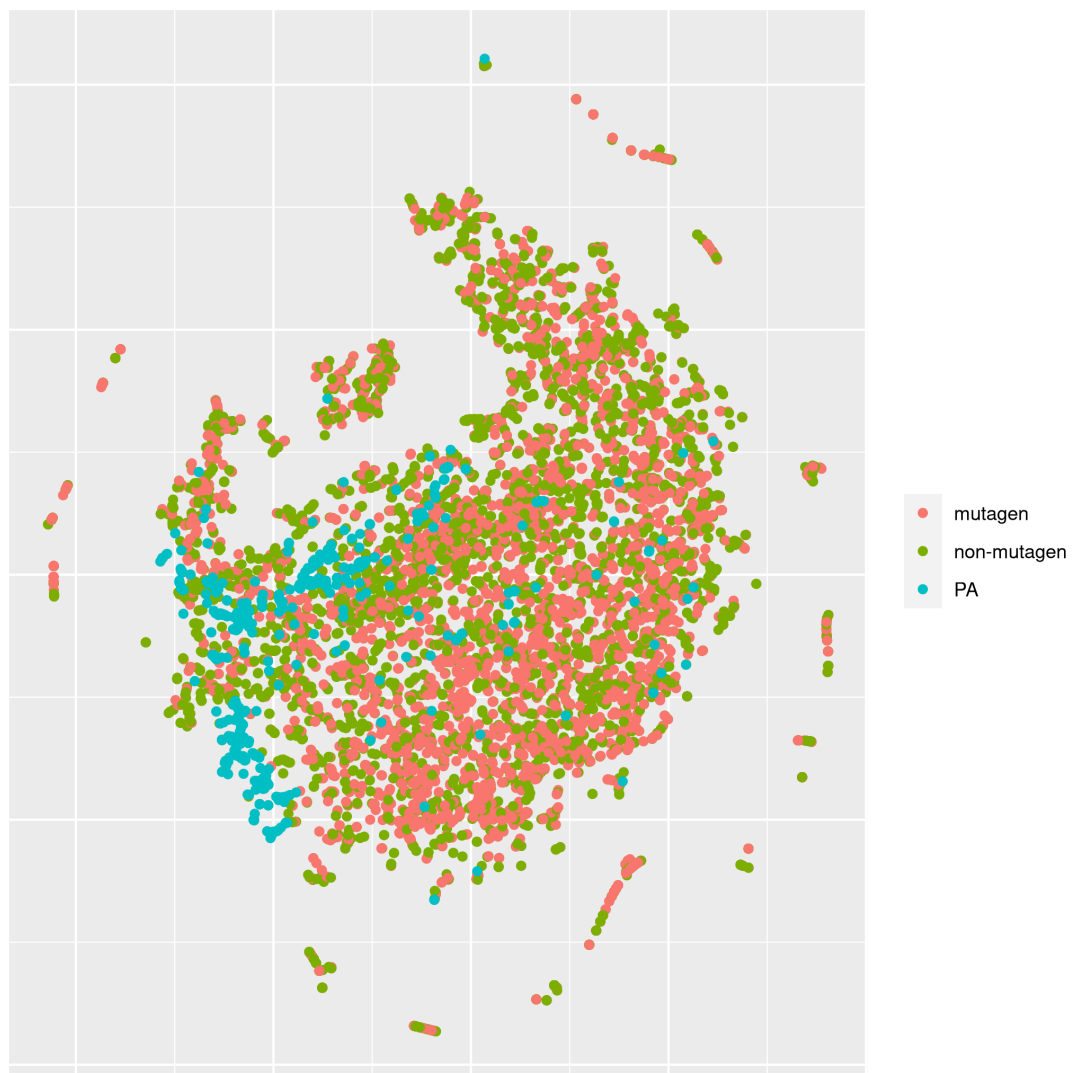


Figure 4: t-SNE visualisation of mutagenicity training data and pyrrolizidine alkaloids (PA)

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 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 out of the applicability domain.

## Descriptors

This study uses two types of descriptors for the characterisation of chemical structures: *MolPrint2D* fingerprints (MP2D, Bender et al. (2004)) use atom environments (i.e. connected atom types 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 the default **lazar** settings, and previous experiments have shown, that they give more accurate results than predefined fragments (e.g. MACCS, FP2-4).

In order to investigate, if MP2D fingerprints are also suitable 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 MolPrint2D descriptors are at the moment unsuitable for standard global machine learning algorithms.

**lazar** does not suffer from the size and 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.

PaDEL calculates topological and physical-chemical descriptors.

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

*PaDEL* descriptors were used for **lazar**, R and Tensorflow models. All models based on PaDEL descriptors had similar crossvalidation accuracies that were significantly lower than **lazar** MolPrint2D results. Direct comparisons are available only for the **lazar** algorithm, and also in this case PaDEL accuracies were lower than MolPrint2D accuracies.

Based on **lazar** results we can conclude, that PaDEL descriptors are less suited for chemical similarity calculations than MP2D descriptors. It is also likely that PaDEL descriptors lead to less accurate predictions for global models, but we cannot draw any definitive conclusion in the absence of MP2D models.

## 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 standard **lazar** models with MolPrint2D descriptors perform better than global models. The accuracy of **lazar** models with PaDEL descriptors is however substantially lower and comparable to global models with the same descriptors.

This observation may lead to the conclusion that the choice of suitable descriptors is more important for predictive accuracy than the modelling algorithm, but we were unable to obtain global MP2D models for direct comparisons. The selection of an appropriate modelling algorithm is still crucial, because it needs the capability to handle the descriptor space. Neighbour (and thus similarity) based algorithms like **lazar** have a clear advantage in this respect over global machine learning algorithms (e.g. RF, SVM, LR, NN), because Tanimoto/Jaccard similarities can be calculated efficiently with simple set operations.

## **Pyrrolizidine alkaloid mutagenicity predictions**

**TODO: Verena** Ich wuerde den Grossteil der Diskussion hier dir ueberlassen. Wenn Du **lazar** Ergebnisse konkret diskutieren willst, kann ich Dir ausfuehrliche Vorhersagen (mit aehnlichen Verbindungen und deren Aktivitaet) fuer einzelne Beispiele zusammenstellen

## **Conclusions**

A new public *Salmonella* mutagenicity training dataset with 8309 compounds was created and used it to train **lazar**, R and Tensorflow models with MolPrint2D and PaDEL descriptors. The best performance was obtained with **lazar** models using MolPrint2D descriptors, with prediction accuracies (84%) comparable to the interlaboratory variability of the Ames test (80-85%). Models based on PaDEL descriptors had lower accuracies



337 than MolPrint2D models, but only the `lazar` algorithm could use MolPrint2D descrip-  
338 tors.

339 **TODO:** PA Vorhersagen

## 340 References

341 Bender, Andreas, Hamse Y. Mussa, Robert C. Glen, and Stephan Reiling. 2004. “Molec-  
342 ular Similarity Searching Using Atom Environments, Information-Based Feature Selec-  
343 tion, and a Naïve Bayesian Classifier.” *Journal of Chemical Information and Computer*  
344 *Sciences* 44 (1): 170–78. <https://doi.org/10.1021/ci034207y>.

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

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

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

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

359 Kazius, J., R. McGuire, and R. Bursi. 2005. “Derivation and Validation of Toxicophores

- 360 for Mutagenicity Prediction.” *J Med Chem*, no. 48: 312–20.
- 361 Maaten, L. J. P. van der, and G. E. Hinton. 2008. “Visualizing Data Using T-Sne.”  
362 *Journal of Machine Learning Research*, no. 9: 2579–2605.
- 363 O’Boyle, Noel, Michael Banck, Craig James, Chris Morley, Tim Vandermeersch, and  
364 Geoffrey Hutchison. 2011. “Open Babel: An open chemical toolbox.” *J. Cheminf.* 3 (1):  
365 33. <https://doi.org/doi:10.1186/1758-2946-3-33>.
- 366 Yap, CW. 2011. “PaDEL-Descriptor: An Open Source Software to Calculate Molecular  
367 Descriptors and Fingerprints.” *Journal of Computational Chemistry*, no. 32: 1466–74.