A comparison of nine 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>

Random forest, support vector machine, logistic regression, neural networks and k-nearest neighbor (`lazar`) algorithms, were applied to new *Salmonella* mutagenicity dataset with 8309 unique chemical structures. The best prediction accuracies in 10-fold-crossvalidation were obtained with `lazar` models and MolPrint2D descriptors, that gave accuracies (84%) similar to the interlaboratory variability of the Ames test.

TODO: PA results

Introduction

TODO: rationale for investigation

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
- 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.

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.

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

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, descriptors 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 dependent variable (experimental mutagenicity), 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.

PaDEL descriptors were used in global (RF, SVM, LR, NN) and local (`lazar`) 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 arbitrary algorithms 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)

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

$\sum \mathrm{sim}_n$ 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 (*RF*)

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 (*SVM*)

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

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

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.

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.

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

Applicability domain

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.

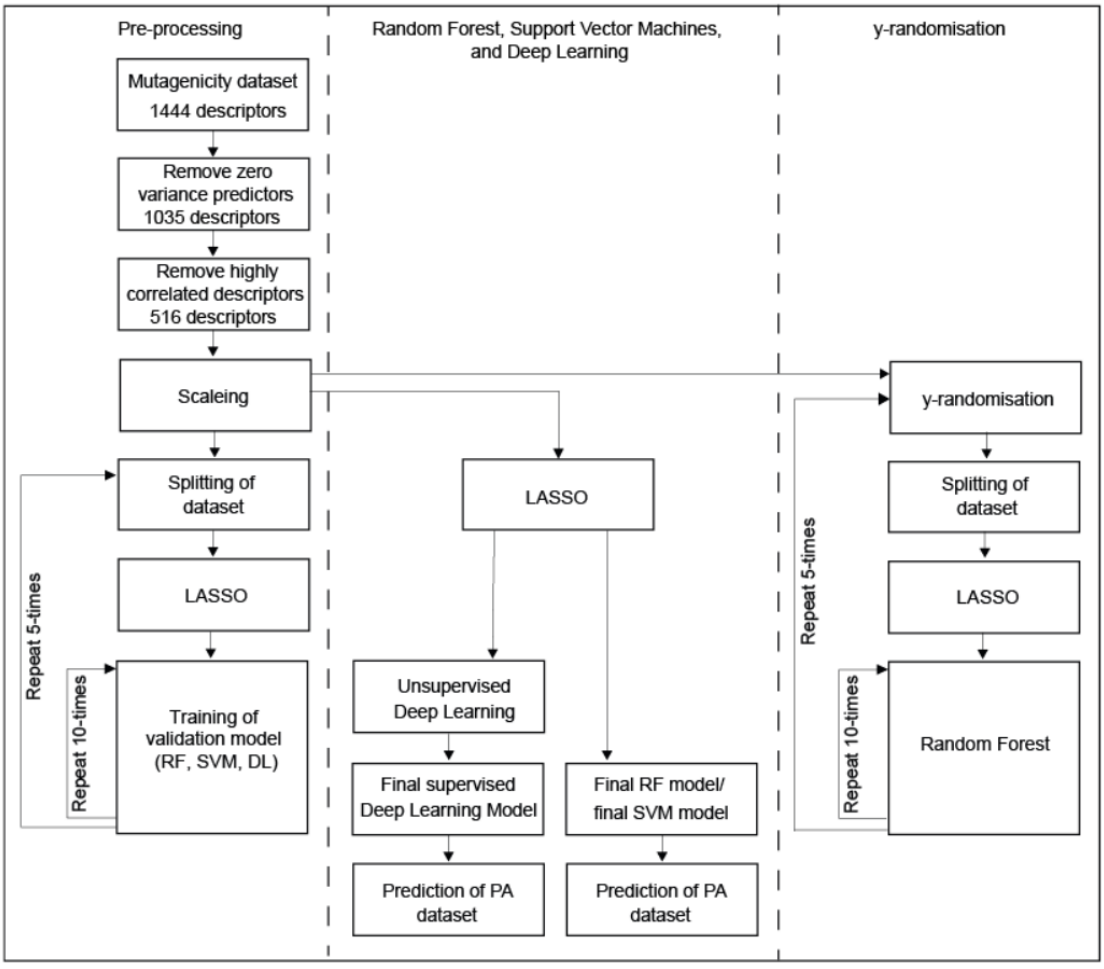


Figure 1: Flowchart of the generation and validation of the models generated in R-project

Tensorflow 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** Ich hab die alten Ergebnisse mit feature selection weggelassen, ist das ok? Dann muesste auch dieser Absatz gestrichen werden, oder?

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

Random forests (*RF*)

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

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

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

Neural Nets (*NN*)

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.

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`

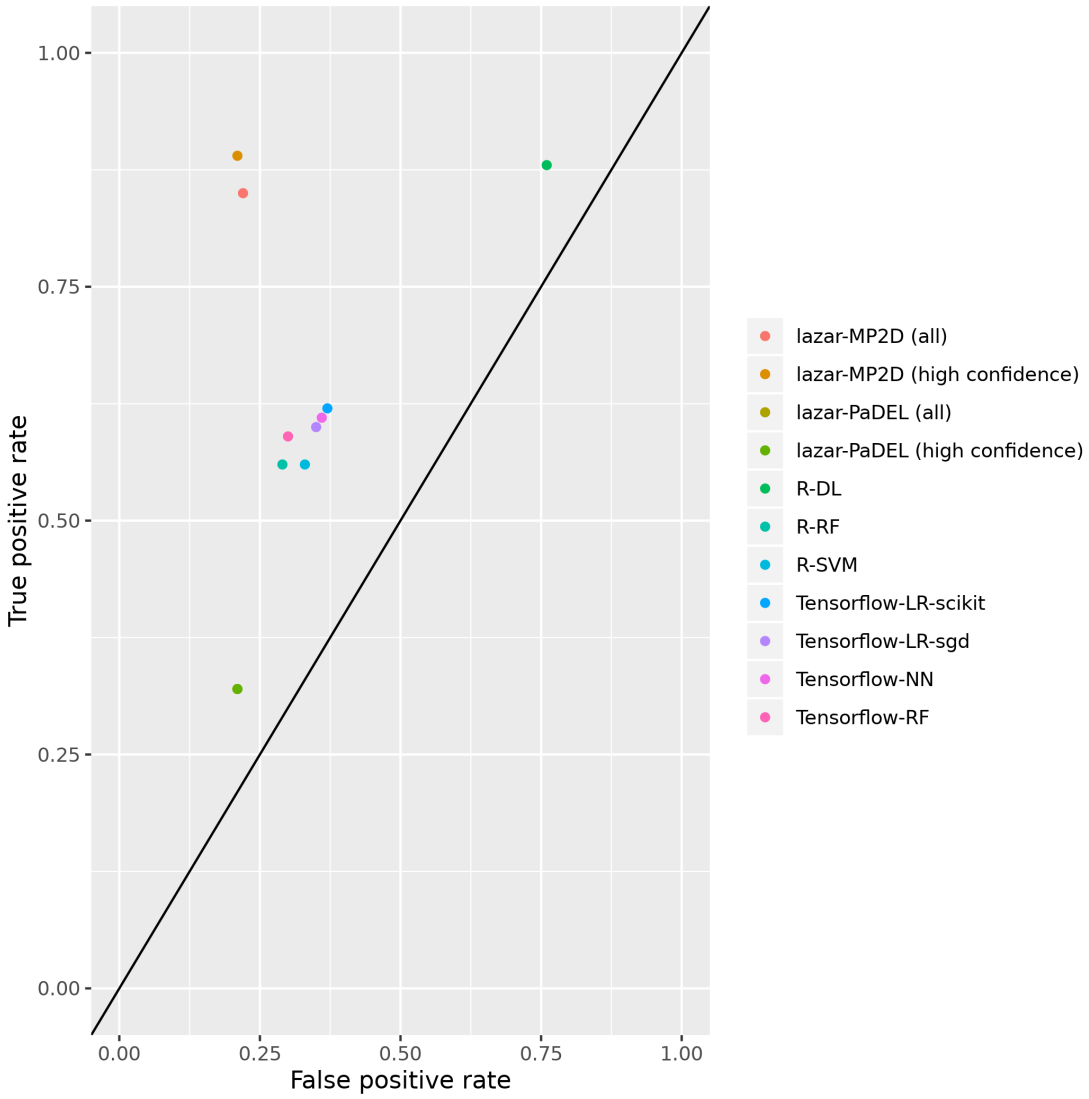


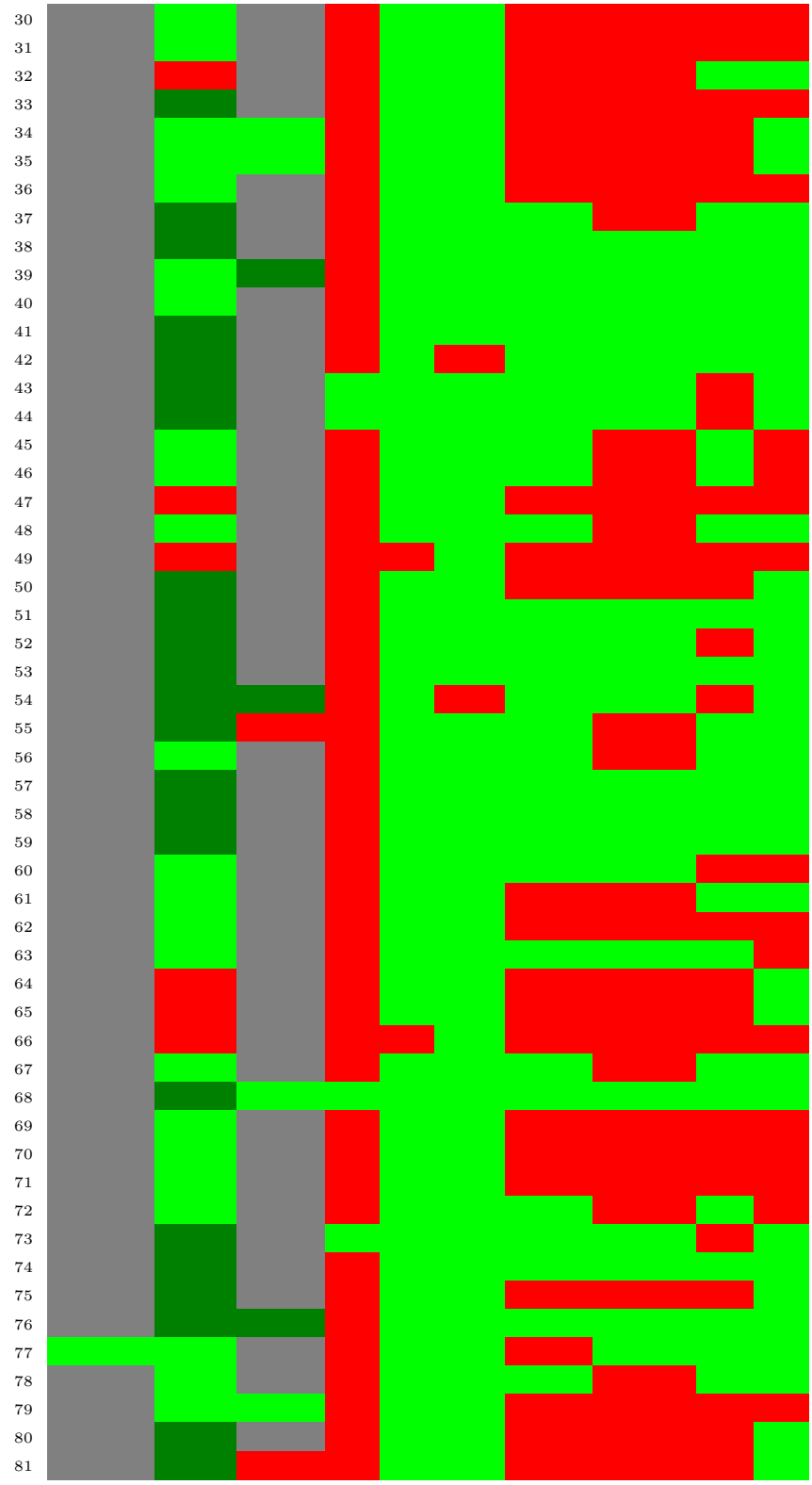
Figure 2: ROC plot of crossvalidation results.

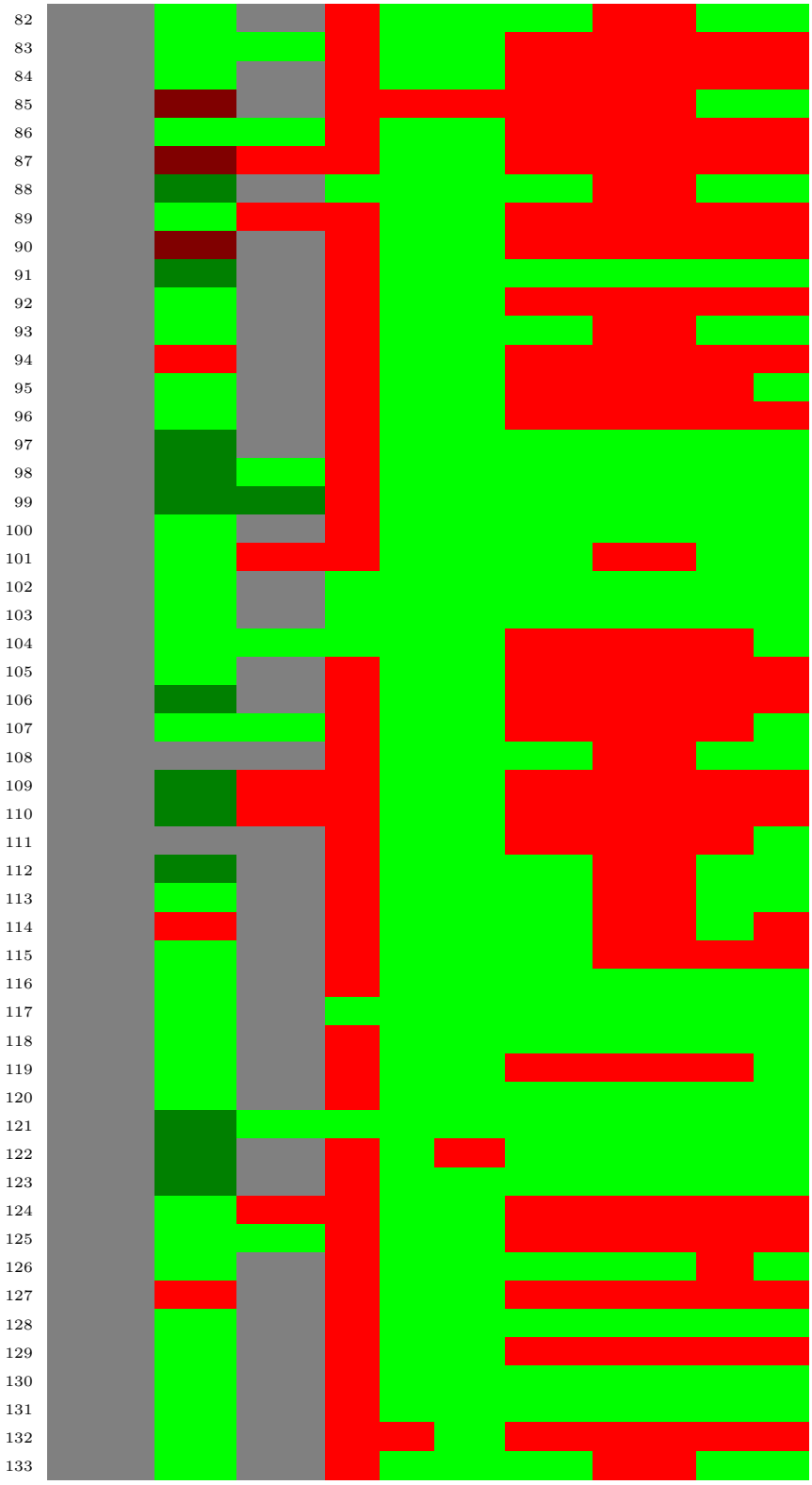
models using MolPrint2D descriptors (0.84 for predictions with high confidence, 0.82 for all predictions). Models utilizing PaDEL descriptors have generally lower accuracies ranging from 0.56 (R deep learning) to 0.64 (R/Tensorflow random forests). Sensitivity and specificity is generally well balanced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

Pyrrolizidine alkaloid mutagenicity predictions

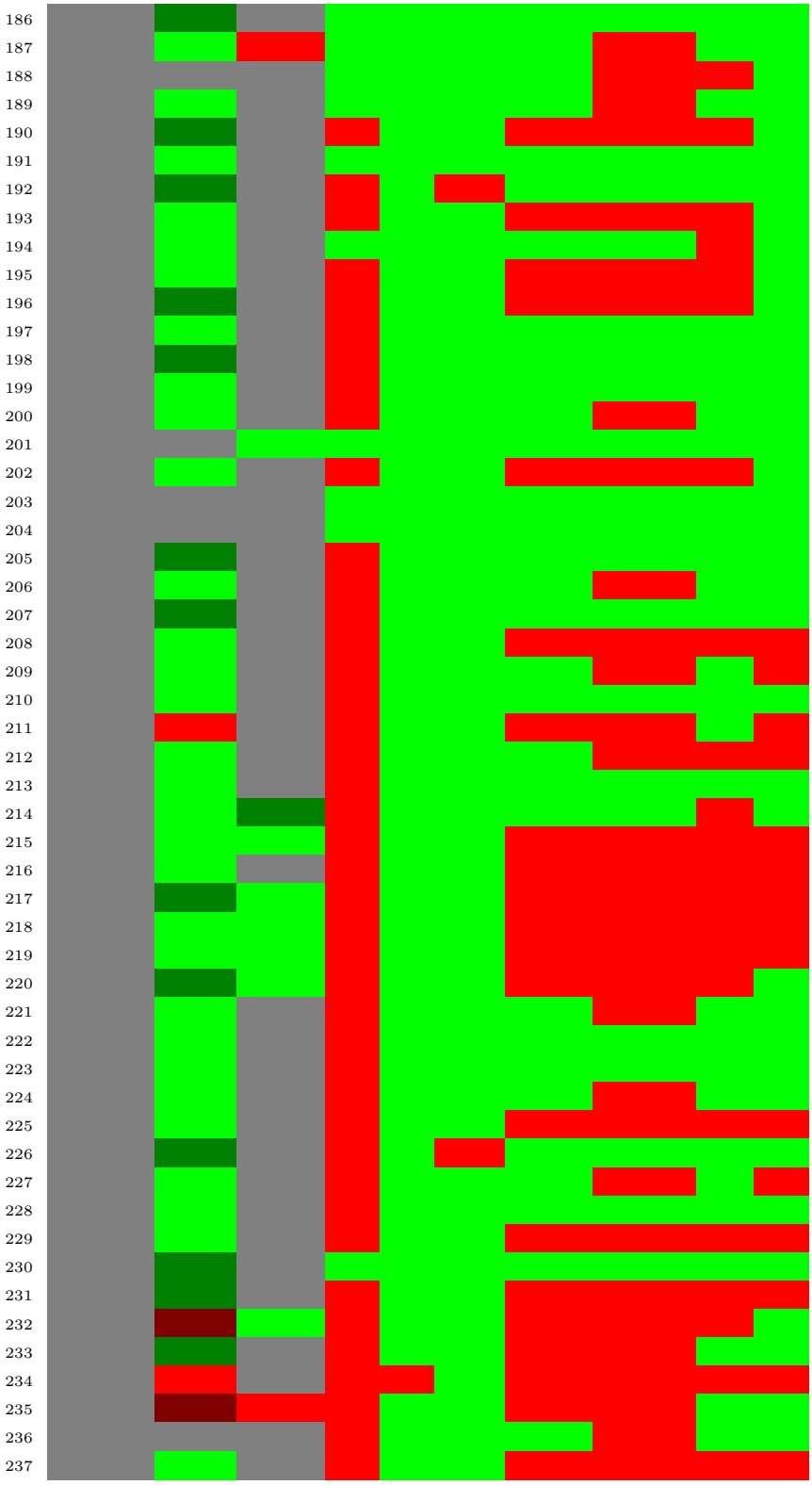
Mutagenicity predictions from all investigated models for 602 pyrrolizidine alkaloids are summarized in Table 4.

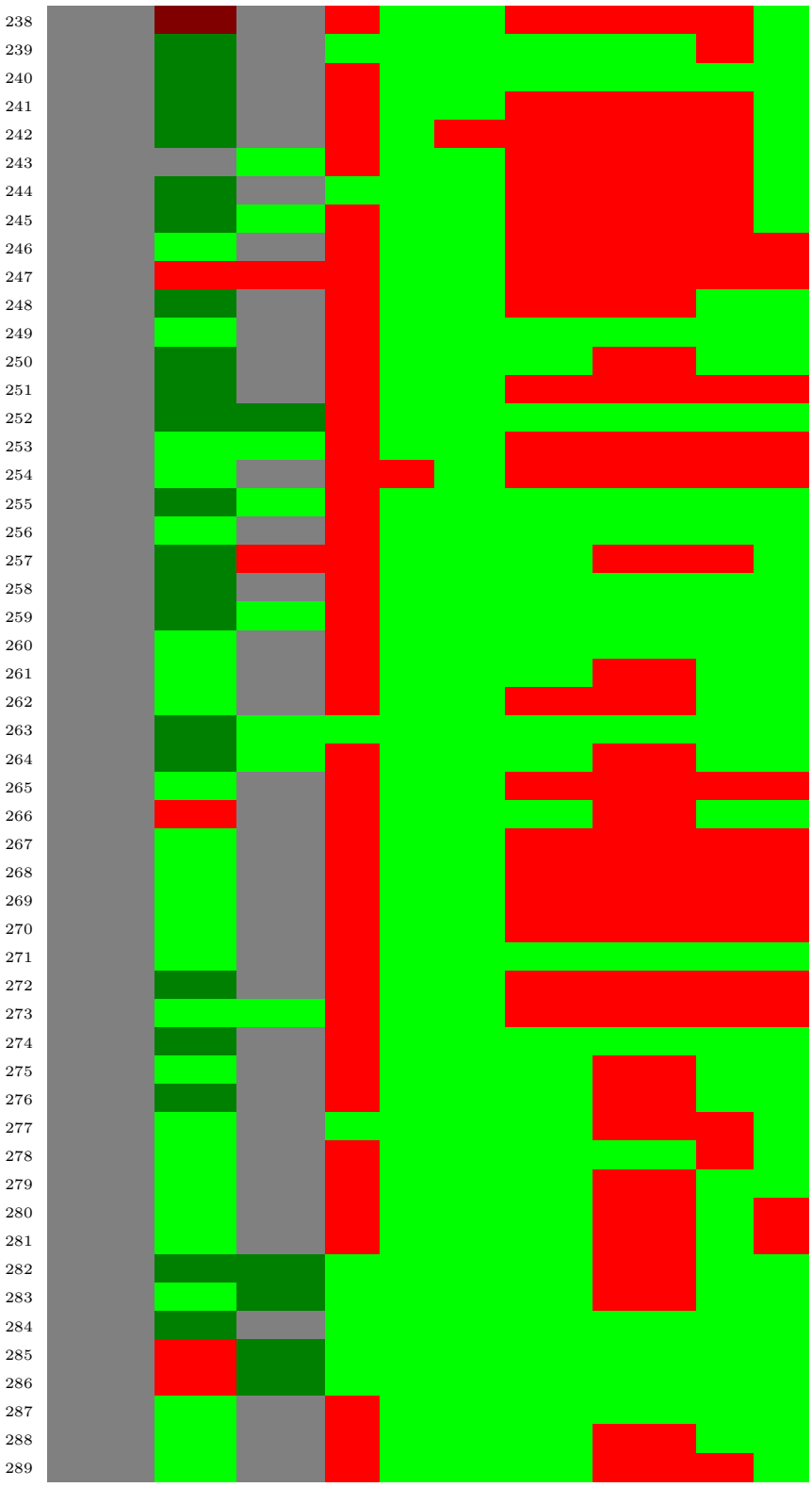
Table 4: Summary of pyrrolizidine alkaloid predictions: red: mutagen, green: non-mutagen, grey: no prediction, dark red/green: low confidence


30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81

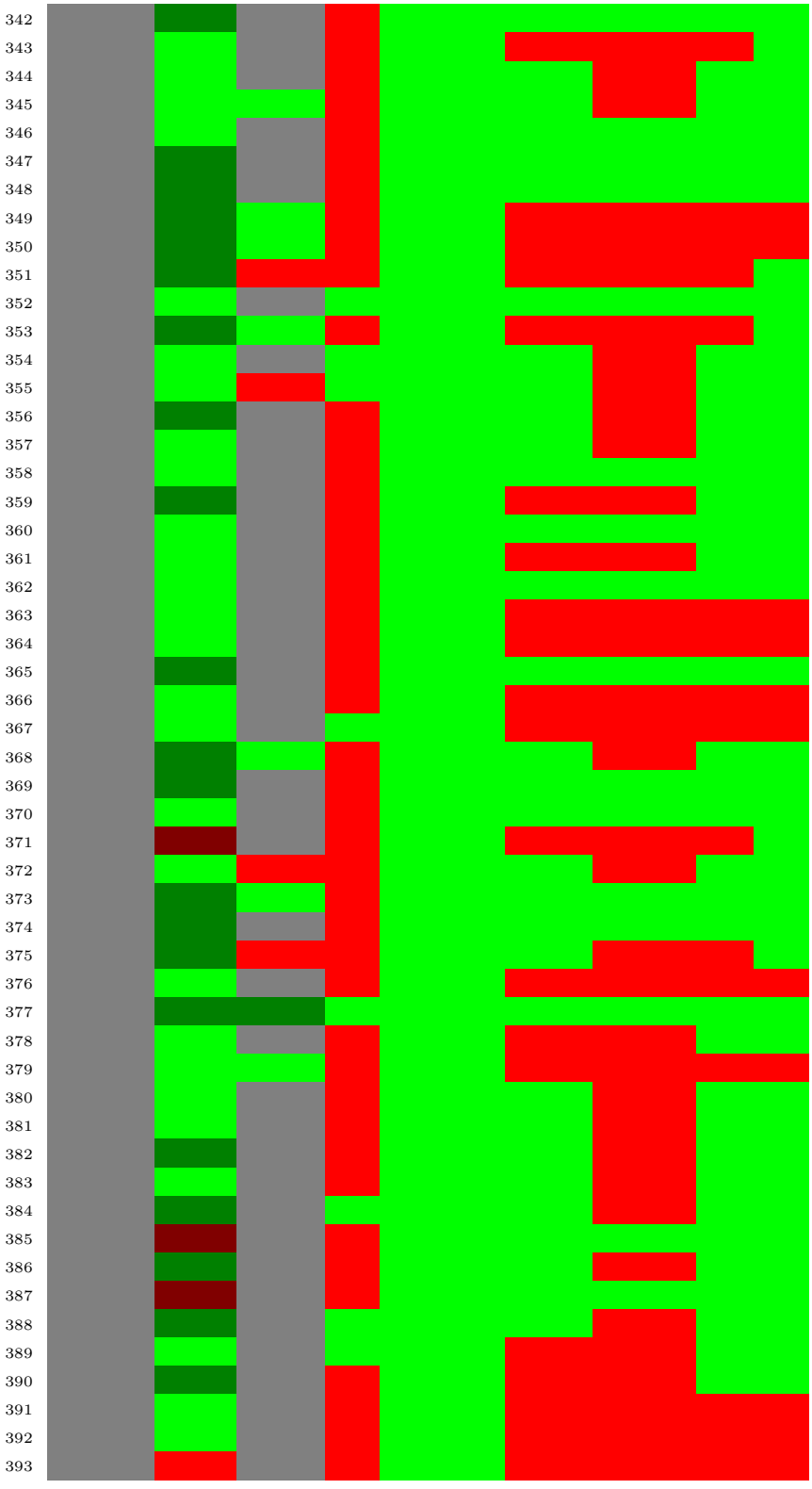

82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133

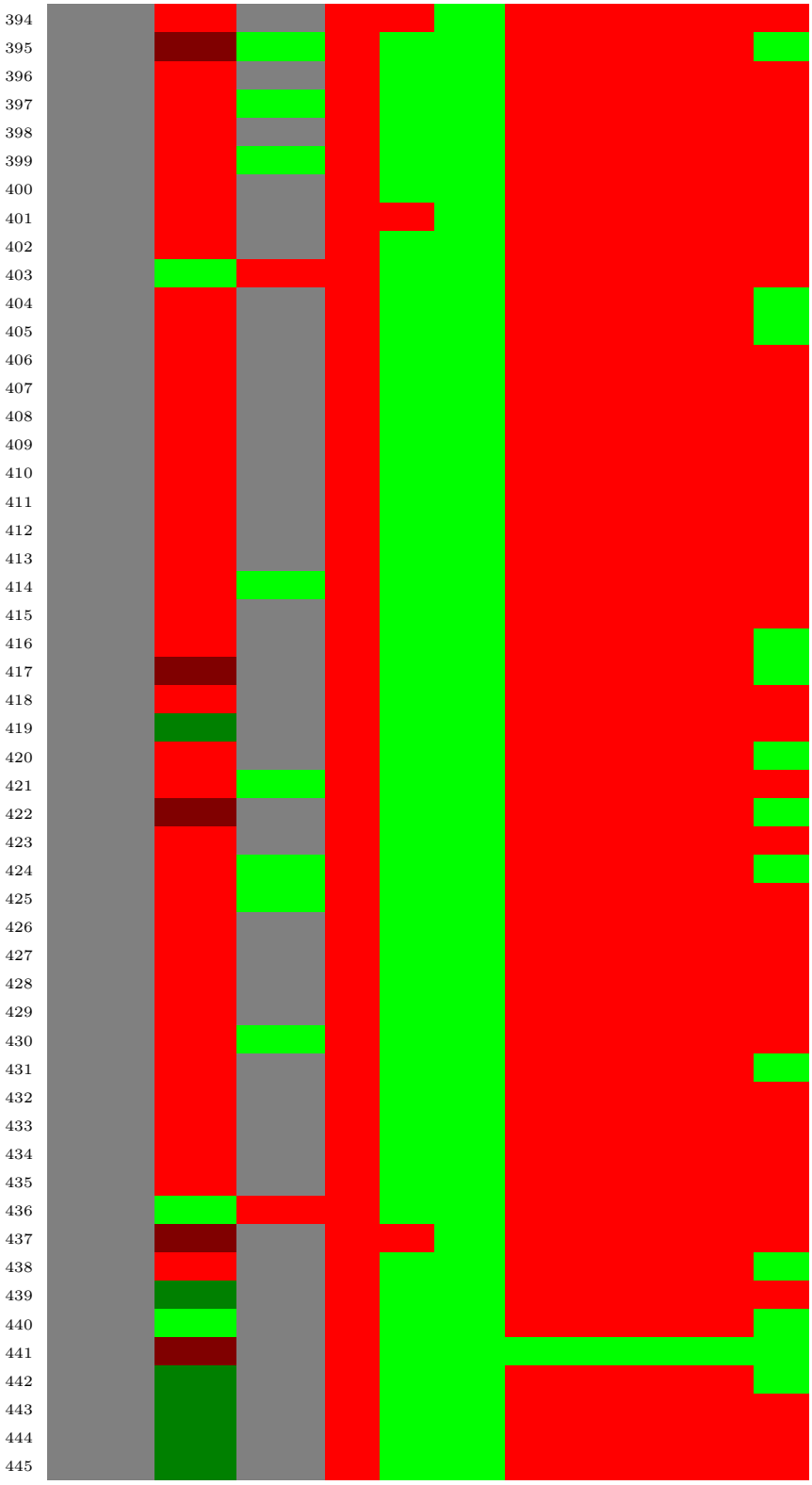

134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185

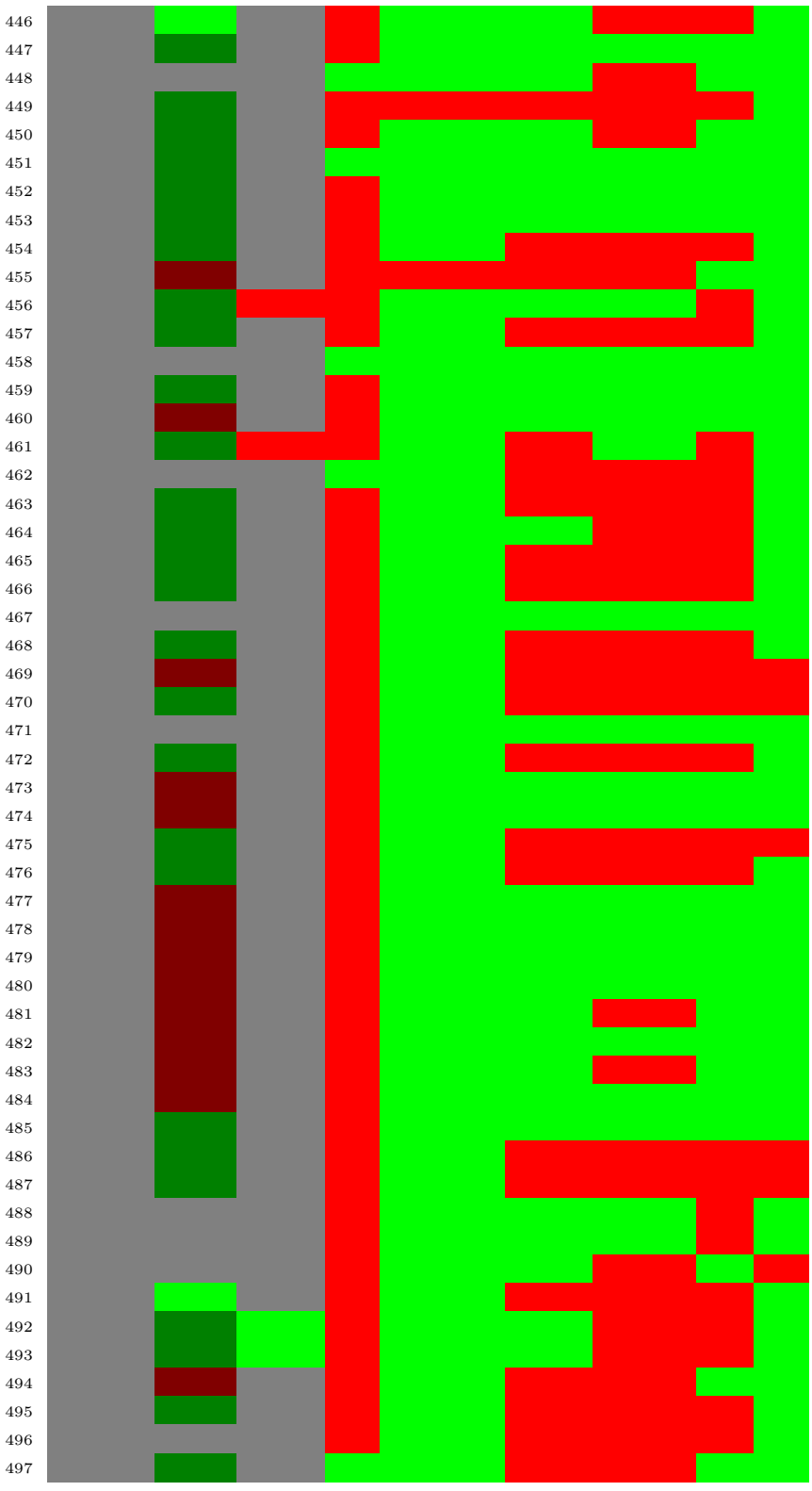

186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237

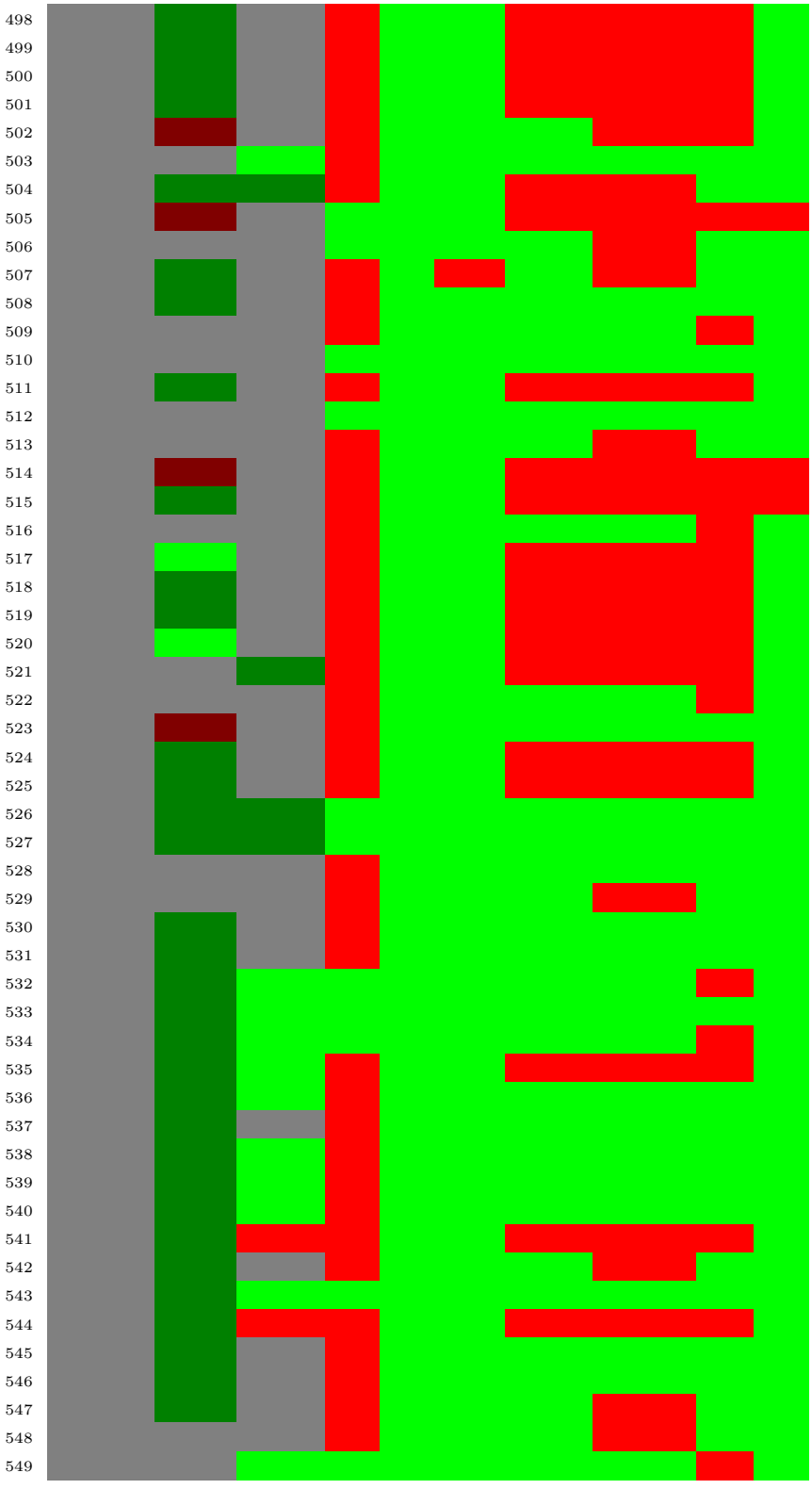

238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289

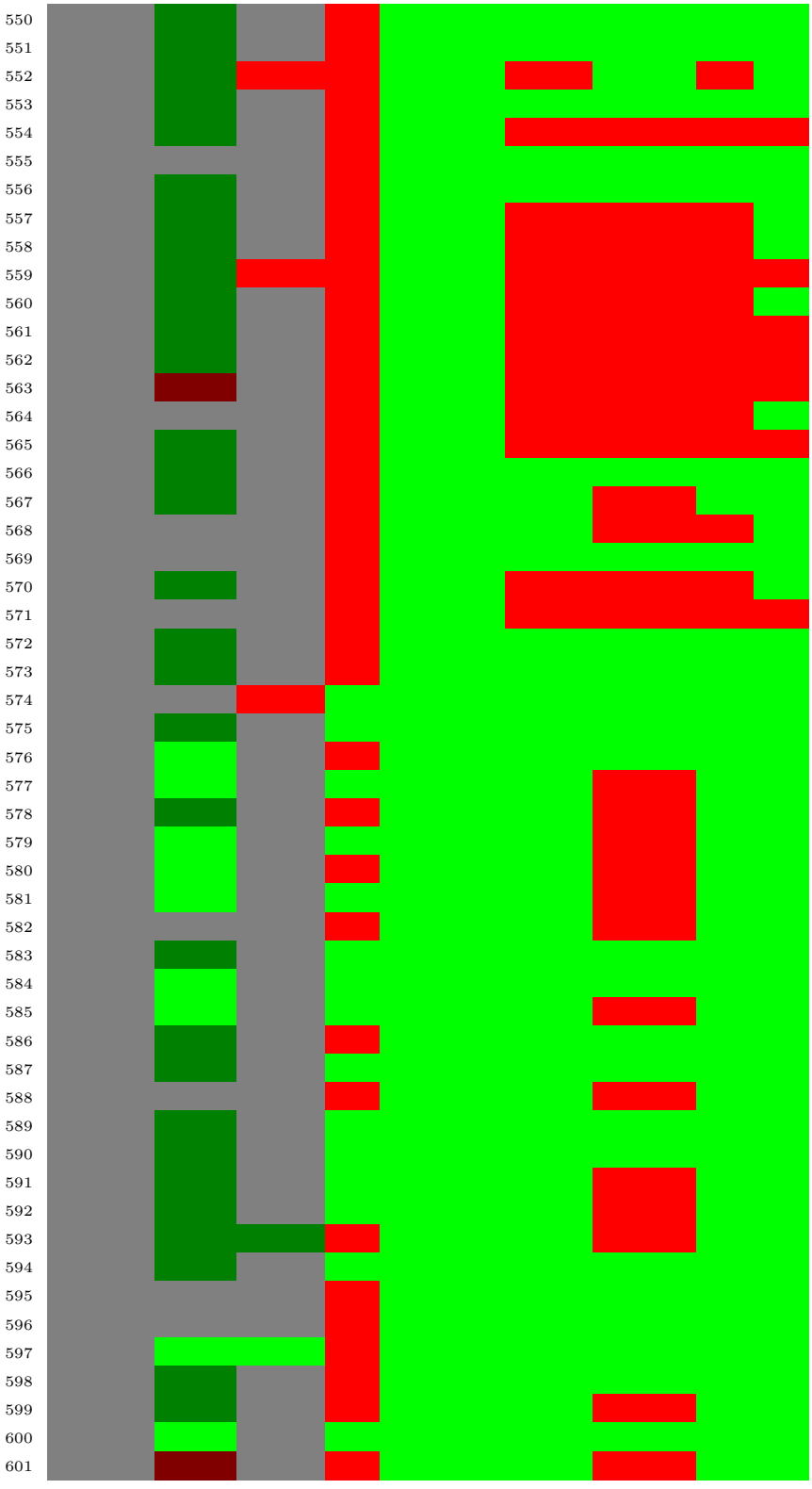

290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341


342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393


394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445


446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497


498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549


550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601

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

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

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

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 1, Table 2, Table 3 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

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

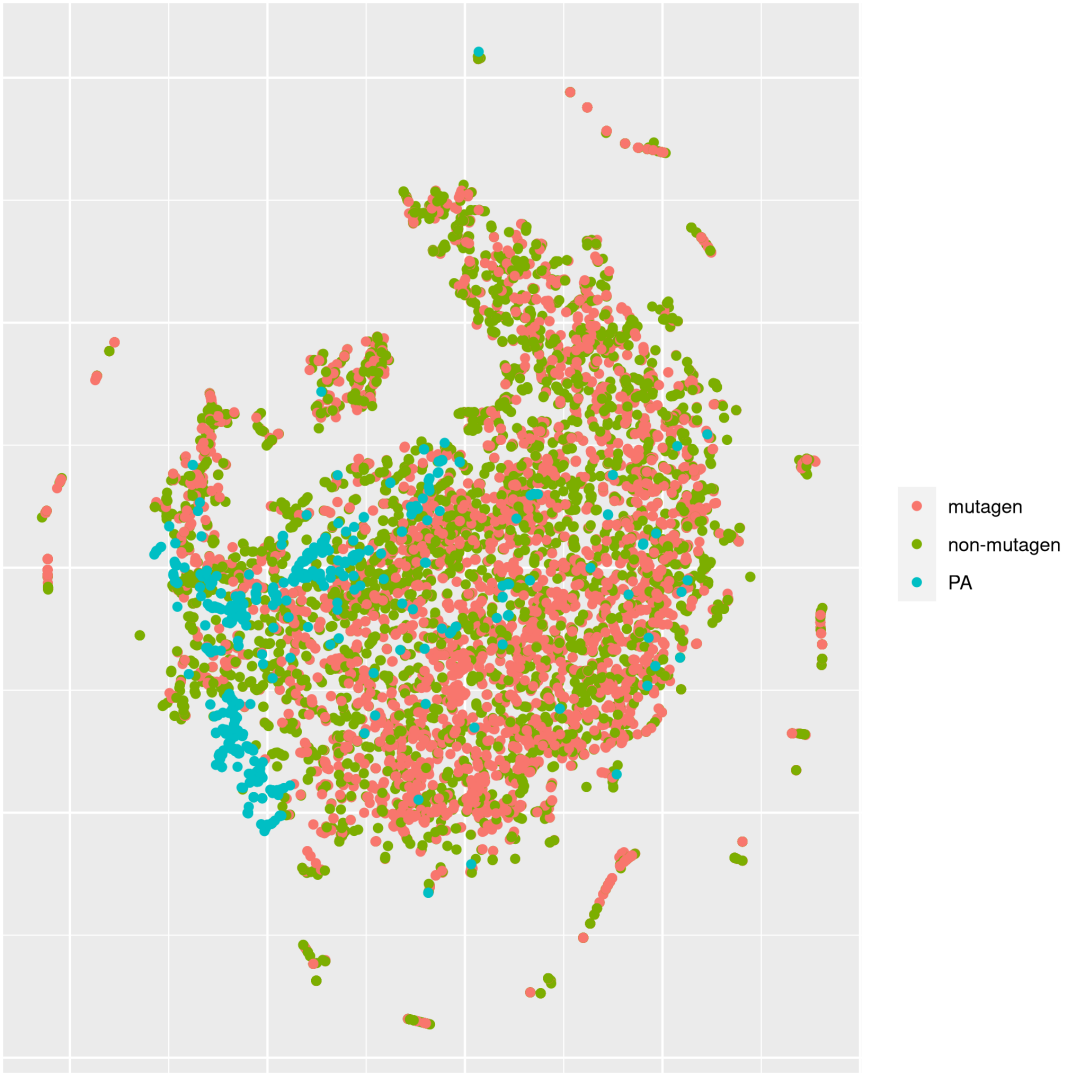


Figure 4: t-SNE visualisation of mutagenicity training data and pyrrolizidine alkaloids (PA)

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: **Philipp** Ich wuerde fuer meinen Teil (generelle Uebersicht, applicability domain) noch die Tensorflow Ergebnisse brauchen.

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 than MolPrint2D models, but only the `lazar` algorithm could use MolPrint2D descriptors.

TODO: PA Vorhersagen

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." *Fron-*

tiers 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.

Maaten, L. J. P. van der, and G. E. Hinton. 2008. "Visualizing Data Using T-Sne." *Journal of Machine Learning Research*, no. 9: 2579–2605.

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.

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