A comparison of nine machine learning mutagenicity models and their application for predicting pyrrolizidine alkaloids

Christoph Helma* [1], Verena Schöning[5], Jürgen Drewe[2,4], and Philipp Boss[3]

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

[2]Max Zeller Söhne 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

[4]Clinical Pharmacology, Department of Pharmaceutical Sciences, University Hospital Basel, University of Basel, Petersgraben 4, 4031 Basel, Switzerland

[5]Clinical Pharmacology and Toxicology, Department of General Internal Medicine, University Hospital Bern, University of Bern, Inselspital, 3010 Bern, Switzerland

* 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 8290 unique chemical structures utilizing MolPrint2D and Chemistry Development Kit (CDK) descriptors. Crossvalidation accuracies of all investigated models ranged from 80-85% which is comparable with the interlaboratory variability of the *Salmonella* mutagenicity assay. Pyrrolizidine alkaloid predictions showed a clear distinction between chemical groups, where Otonecines had the highest proportion of positive mutagenicity predictions and Monoester the lowest.

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 Chemistry Development Kit (*CDK*) 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*, Weininger, Weininger, and Weininger (1989)) 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 8290 unique structures and 8309 individual measurements.

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/tree/mutagenicity/mutagenicity.csv.

Pyrrolizidine alkaloid (PA) dataset

The pyrrolizidine alkaloid dataset was created from five independent, necine base substructure searches in PubChem (https://pubchem.ncbi.nlm.nih.gov/) and compared to the PAs listed in the EFSA publication EFSA (2011) and the book by Mattocks Mattocks (1986), to ensure, that all major PAs were included. PAs mentioned in these publications which were not found in the downloaded substances were searched individually in PubChem and, if available, downloaded separately. Non-PA substances, duplicates, and isomers were removed from the files, but artificial PAs, even if unlikely to occur in nature, were kept. The resulting PA dataset comprised a total of 602 different PAs.

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

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

- Retronecine-type (1,2-unstaturated necine base, 392 compounds)
- Otonecine-type (1,2-unstaturated necine base, 46 compounds)
- Platynecine-type (1,2-saturated necine base, 140 compounds)

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

- N-oxide-type (84 compounds)
- Tertiary-type (PAs which were neither from the N-oxide- nor DHP-type, 495 compounds)
- Dehydropyrrolizidine-type (pyrrolic ester, 23 compounds)

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

- Monoester-type (154 compounds)
- Open-ring diester-type (163 compounds)
- Macrocyclic diester-type (255 compounds)

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 CDK) they are generated dynamically from chemical structures. This has the advantage that they can capture unknown substructures of toxicological relevance that are not included in other descriptors. In addition they allow the efficient calculation of chemical similarities (e.g. Tanimoto indices) with simple set operations.

MolPrint2D fingerprints were calculated with the OpenBabel cheminformatics library (O'Boyle et al. (2011)). They can be obtained from the following locations:

Training data:

- sparse representation (https://git.in-silico.ch/mutagenicity-paper/tree/mutagenicity/mp2d/fingerprints.mp2d)
- descriptor matrix (https://git.in-silico.ch/mutagenicity-paper/tree/mutagenicity/mp2d/mutagenicity-fingerprints.csv.gz)

Pyrrolizidine alkaloids:

- sparse representation (https://git.in-silico.ch/mutagenicity-paper/tree/pyrrolizidine-alkaloids/mp2d/fingerprints.mp2d)
- descriptor matrix (https://git.in-silico.ch/mutagenicity-paper/tree/pyrrolizidine-alkaloids/mp2d/pa-fingerprints.csv.gz)

Chemistry Development Kit (*CDK*) descriptors

Molecular 1D and 2D descriptors were calculated with the PaDEL-Descriptors program (http://www.yapcwsoft.com version 2.21, Yap (2011)). PaDEL uses the Chemistry Development Kit (*CDK*, https://cdk.github.io/index.html) library for descriptor calculations.

As the training dataset contained 8290 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 1442 descriptors for 8083 compounds.

CDK training data can be obtained from https://git.in-silico.ch/mutagenicity-paper/tree/mutagenicity/cdk/mutagenicity-mod-2.new.csv.

The same procedure was applied for the pyrrolizidine dataset yielding descriptors for compounds. CDK features for pyrrolizidine alkaloids are available at https://git.in-silico.ch/mutagenicity-paper/tree/pyrrolizidine-alkaloids/cdk/PA-Padel-2D_m2.csv.

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.

Feature preprocessing

MolPrint2D features were used without preprocessing. Near zero variance and strongly correlated CDK descriptors were removed and the remaining descriptor values were centered and scaled. Preprocessing was performed with the R `caret` preProcess function using the methods “nzv”,“corr”,“center” and “scale” with default settings.

Neighbour identification

Utilizing this modularity, similarity calculations were based both on MolPrint2D fingerprints and on CDK 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 CDK 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 (MP2D/Tanimoto) or 0.9 (CDK/Cosine) 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 (MP2D/Tanimoto) or 0.7 (CDK/Cosine) and the prediction is flagged with a warning that it might be out of the applicability domain of the training data (*low confidence*).

- These Similarity thresholds are the default values chosen by 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.

Validation

10-fold cross validation was performed for model evaluation.

Pyrrolizidine alkaloid predictions

For the prediction of pyrrolizidine alkaloids models were generated with the MP2D and CDK training datasets. The complete feature set was used for MP2D predictions, for CDK predictions the intersection between training and pyrrolizidine alkaloid features was used.

Availability

- Source code for this manuscript (GPL3): https://git.in-silico.ch/lazar/tree/?h=mutagenicity-paper
- Crossvalidation experiments (GPL3): https://git.in-silico.ch/lazar/tree/models/?h=mutagenicity-paper
- Pyrrolizidine alkaloid predictions (GPL3): https://git.in-silico.ch/lazar/tree/predictions/?h=mutagenicity-paper
- Public web interface: https://lazar.in-silico.ch

Tensorflow models

Feature Preprocessing

For preprocessing of the CDK features we used a quantile transformation to a uniform distribution. MP2D features were not preprocessed.

Random forests (*RF*)

For the random forest classifier we used the parameters n_estimators=1000and max_leaf_nodes=200. For the other parameters we used the scikit-learn default values.

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

For the logistic regression we used an ensemble of five trained models. For each model we used a batch size of 64 and trained for 50 epoch. As an optimizer ADAM was chosen. For the other parameters we used the tensorflow default values.

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

For the logistic regression we used as parameters the scikit-learn default values.

Neural Nets (*NN*)

For the neural network we used an ensemble of five trained models. For each model we used a batch size of 64 and trained for 50 epoch. As an optimizer ADAM was chosen. The neural network had 4 hidden layers with 64 nodes each and a ReLu activation function. For the other parameters we used the tensorflow default values.

Support vector machines (*SVM*)

We used the SVM implemented in scikit-learn. We used the parameters kernel='rbf', gamma='scale'. For the other parameters we used the scikit-learn default values.

Validation

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

Pyrrolizidine alkaloid predictions

For the prediction of pyrrolizidine alkaloids we trained the model described above on the training data. For training and prediction only the features were used that were in the intersection of features from the training data and the pyrrolizidine alkaloids.

Availability

Jupyter notebooks for these experiments can be found at the following locations

Crossvalidation:

- MolPrint2D fingerprints: https://git.in-silico.ch/mutagenicity-paper/tree/crossvalidations/mp2d/tensorflow
- CDK descriptors: https://git.in-silico.ch/mutagenicity-paper/tree/crossvalidations/cdk/tensorflow

Pyrrolizidine alkaloids:

- MolPrint2D fingerprints: https://git.in-silico.ch/mutagenicity-paper/tree/pyrrolizidine-alkaloids/mp2d/tensorflow
- CDK descriptors: https://git.in-silico.ch/mutagenicity-paper/tree/pyrrolizidine-alkaloids/cdk/tensorflow
- CDK desc

Results

10-fold crossvalidations

Crossvalidation results are summarized in the following tables: Table 1 shows results with MolPrint2D descriptors and Table 2 with CDK descriptors.

Table 1: Summary of crossvalidation results with MolPrint2D descriptors (lazar-HC: lazar with high confidence, lazar-all: all lazar predictions, RF: random forests, LR-sgd: logistic regression (stochastic gradient descent), LR-scikit: logistic regression (scikit), NN: neural networks, SVM: support vector machines)

	lazar-HC	lazar-all	RF	LR-sgd	LR-scikit	NN	SVM
Accuracy	84	82	80	84	84	84	84
True positive rate	89	85	78	83	83	82	83
True negative rate	78	78	82	84	85	85	86
Positive predictive value	83	80	81	84	84	84	85
Negative predictive value	86	84	80	84	84	83	84
Nr. predictions	5864	7782	8303	8303	8303	8303	8303

Table 2: Summary of crossvalidation results with CDK descriptors (lazar-HC: lazar with high confidence, lazar-all: all lazar predictions, RF: random forests, LR-sgd: logistic regression (stochastic gradient descent), LR-scikit: logistic regression (scikit), NN: neural networks, SVM: support vector machines)

	lazar-HC	lazar-all	RF	LR-sgd	LR-scikit	NN	SVM
Accuracy	85	82	84	79	80	85	82
True positive rate	87	84	81	81	80	85	82
True negative rate	82	80	86	78	80	85	82
Positive predictive value	85	81	85	79	80	85	82
Negative predictive value	85	82	82	80	80	85	82

	lazar-HC	lazar-all	RF	LR-sgd	LR-scikit	NN	SVM
Nr. predictions	4872	7353	8077	8077	8077	8077	8077

Figure 1 depicts the position of all crossvalidation results in receiver operating characteristic (ROC) space.

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

All investigated algorithm/descriptor combinations give accuracies between (80 and 85%) which is equivalent to the experimental variability of the *Salmonella typhimurium* mutagenicity bioassay (80-85%, Benigni and Giuliani (1988)). Sensitivities and specificities are balanced in all of these models.

Pyrrolizidine alkaloid mutagenicity predictions

Mutagenicity predictions of 602 pyrrolizidine alkaloids (PAs) from all investigated models can be downloaded from https://git.in-silico.ch/mutagenicity-paper/tree/pyrrolizidine-alkaloids/pa-predictions.csv. A visual representation of all PA predictions can be found at https://git.in-silico.ch/mutagenicity-paper/tree/pyrrolizidine-alkaloids/pa-predictions.pdf.

Figure 2 displays the proportion of positive mutagenicity predictions from all models for the different pyrrolizidine alkaloid groups. Tensorflow models predicted all 602 pyrrolizidine alkaloids, `lazar` MP2D models predicted 560 compounds (301 with high confidence) and `lazar` CDK models 500 compounds (246 with high confidence).

For the visualisation of the position of pyrrolizidine alkaloids in respect to the training data set we have applied t-distributed stochastic neighbor embedding (t-SNE,

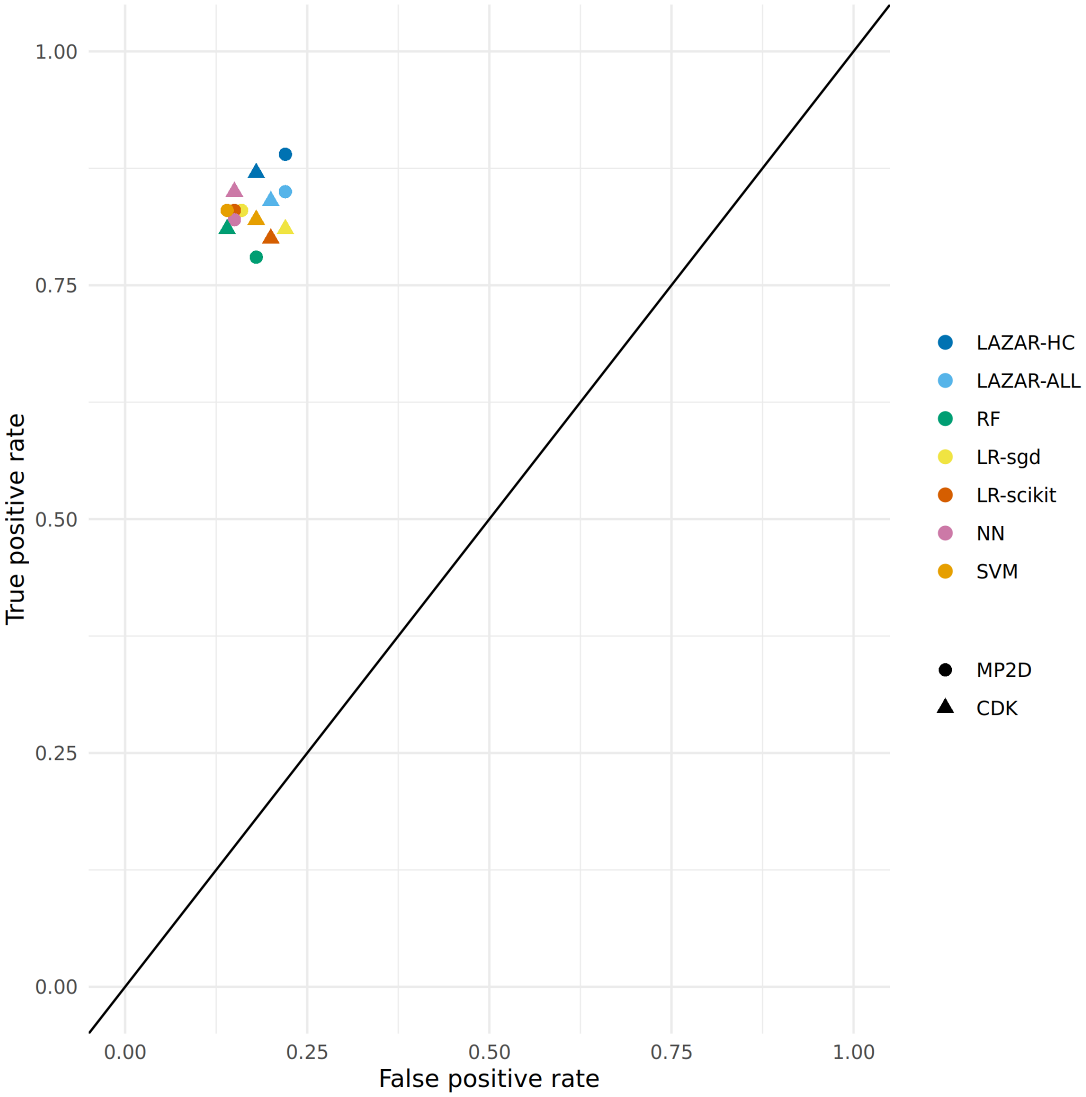


Figure 1: ROC plot of crossvalidation results (lazar-HC: lazar with high confidence, lazar-all: all lazar predictions, RF: random forests, LR-sgd: logistic regression (stochastic gradient descent), LR-scikit: logistic regression (scikit), NN: neural networks, SVM: support vector machines).

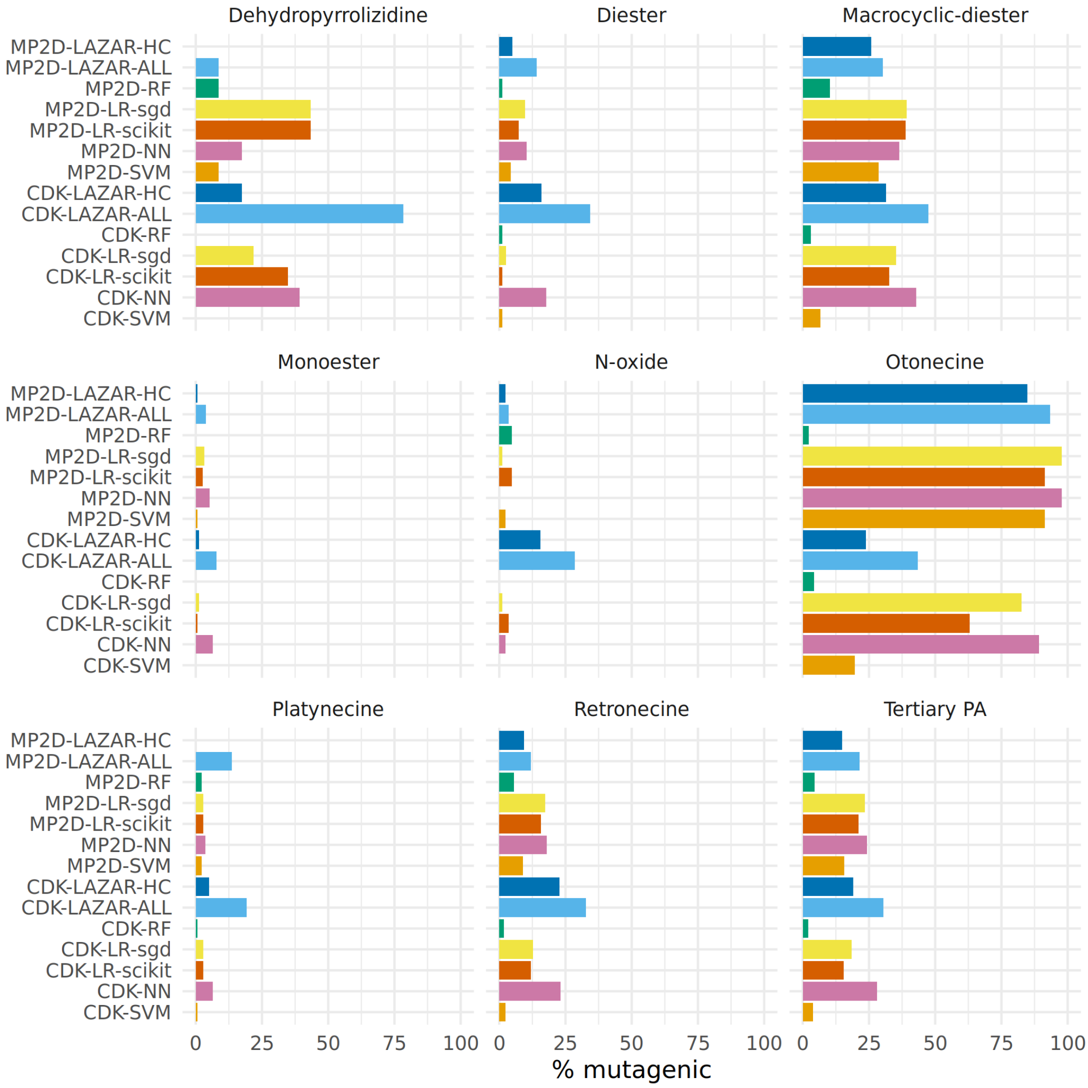


Figure 2: Summary of pyrrolizidine alkaloid predictions

Maaten and Hinton (2008)) for MolPrint2D and CDK descriptors. t-SNE maps each high-dimensional object (chemical) to a two-dimensional point, maintaining the high-dimensional distances of the objects. Similar objects are represented by nearby points and dissimilar objects are represented by distant points. t-SNE coordinates were calculated with the R `Rtsne` package using the default settings (perplexity = 30, theta = 0.5, max_iter = 1000).

Figure 3 shows the t-SNE of pyrrolizidine alkaloids (PA) and the mutagenicity training data in MP2D space (Tanimoto/Jaccard similarity), which resembles basically the structural diversity of the investigated compounds.

Figure 4 shows the t-SNE of pyrrolizidine alkaloids (PA) and the mutagenicity training data in CDK space (Euclidean similarity), which resembles basically the physical-chemical properties of the investigated compounds.

Figure 5 and Figure 6 depict two example pyrrolizidine alkaloid mutagenicity predictions in the context of training data. t-SNE visualisations of all investigated models can be downloaded from https://git.in-silico.ch/mutagenicity-paper/figures.

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 8290 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/tree/mutagenicity/mutagenicity.csv.

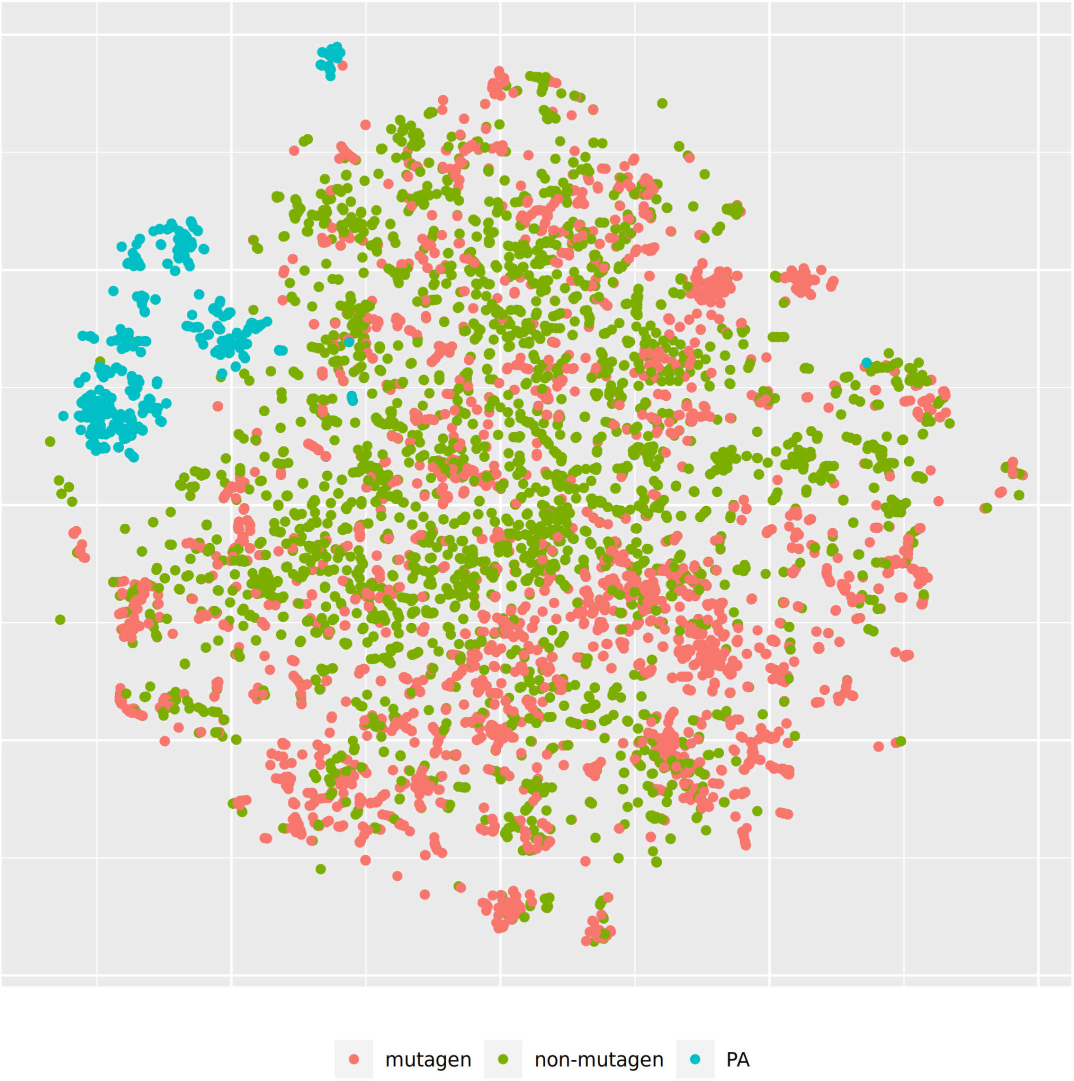


Figure 3: t-SNE visualisation of mutagenicity training data and pyrrolizidine alkaloids (PA) in MP2D space

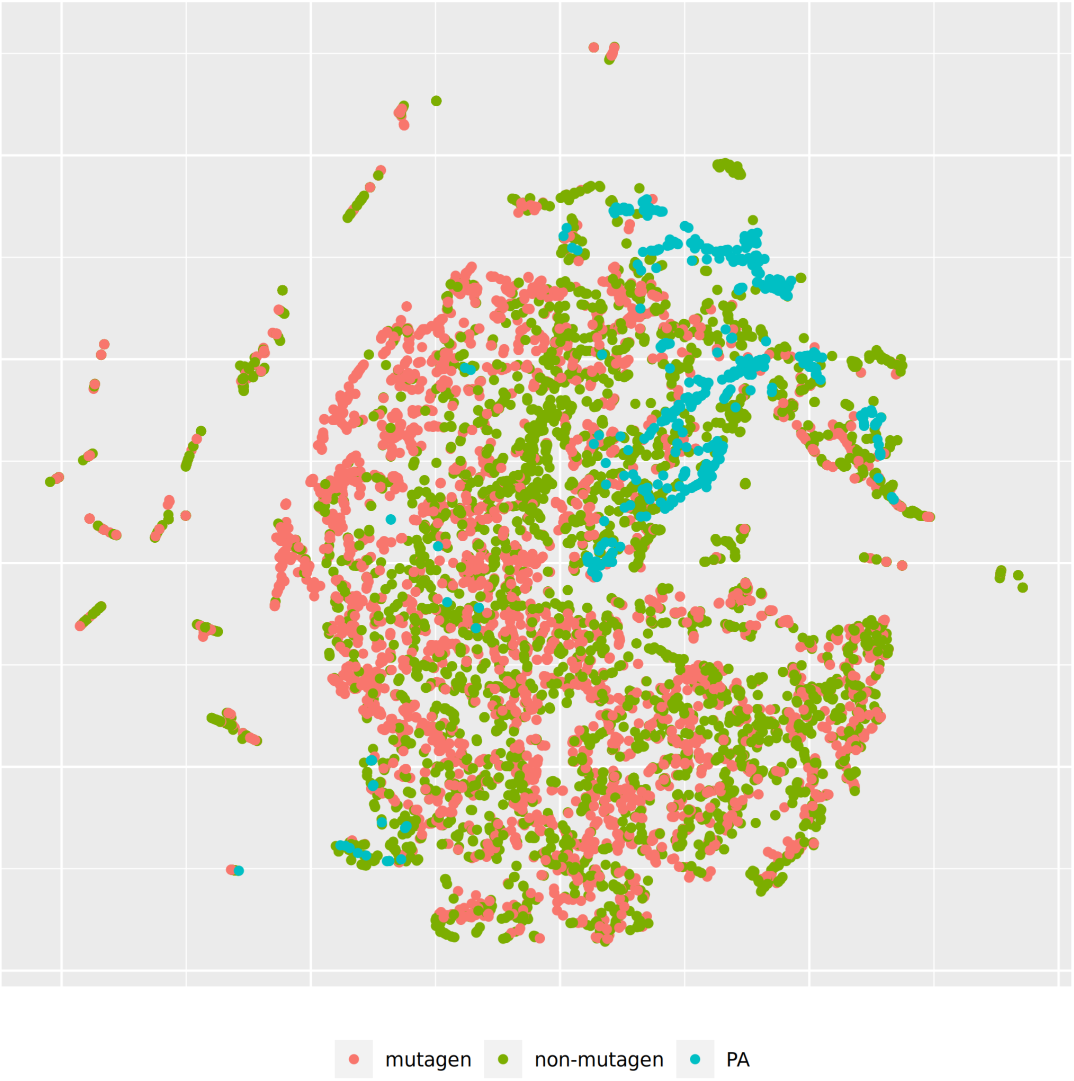


Figure 4: t-SNE visualisation of mutagenicity training data and pyrrolizidine alkaloids (PA) in CDK space

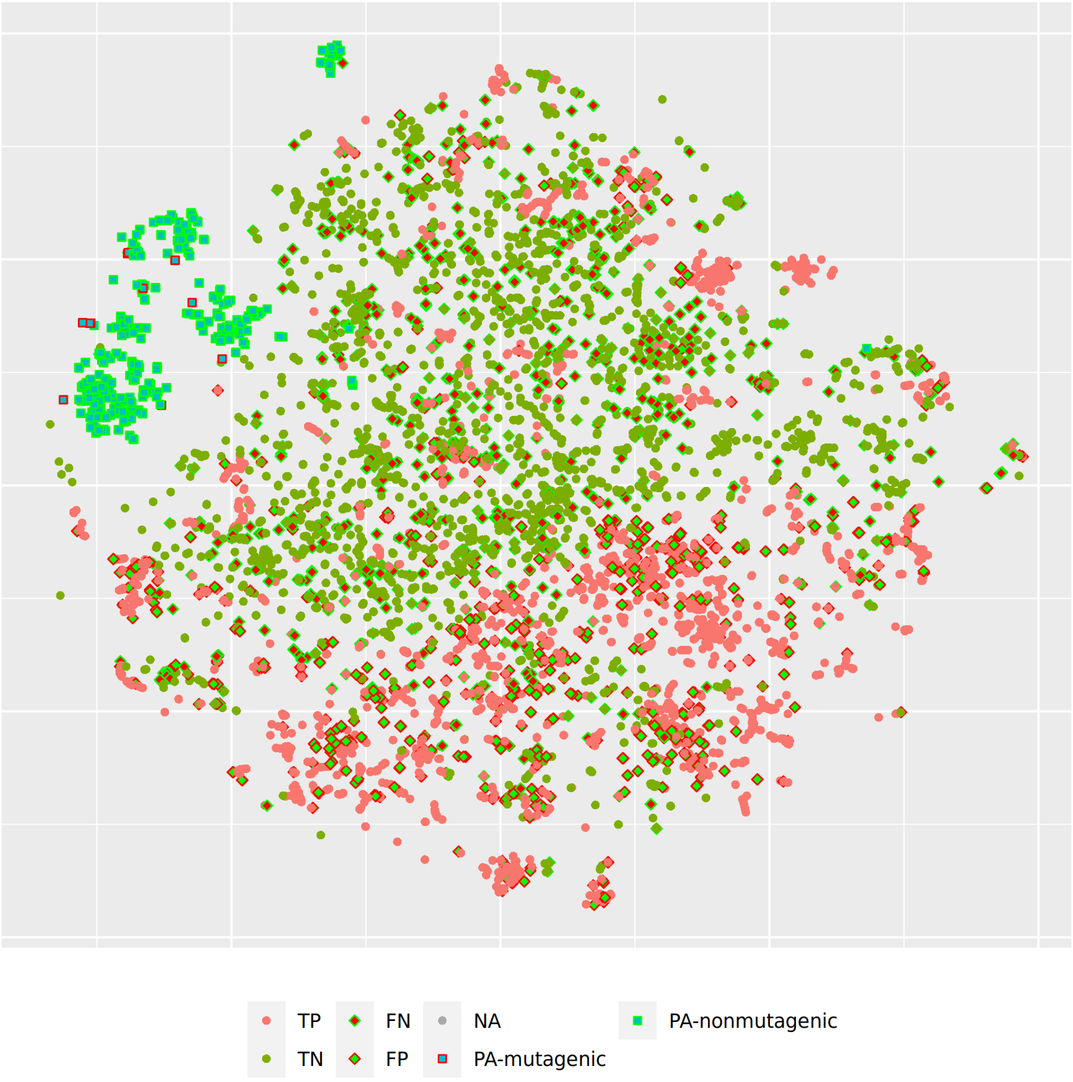


Figure 5: t-SNE visualisation of MP2D random forest predictions

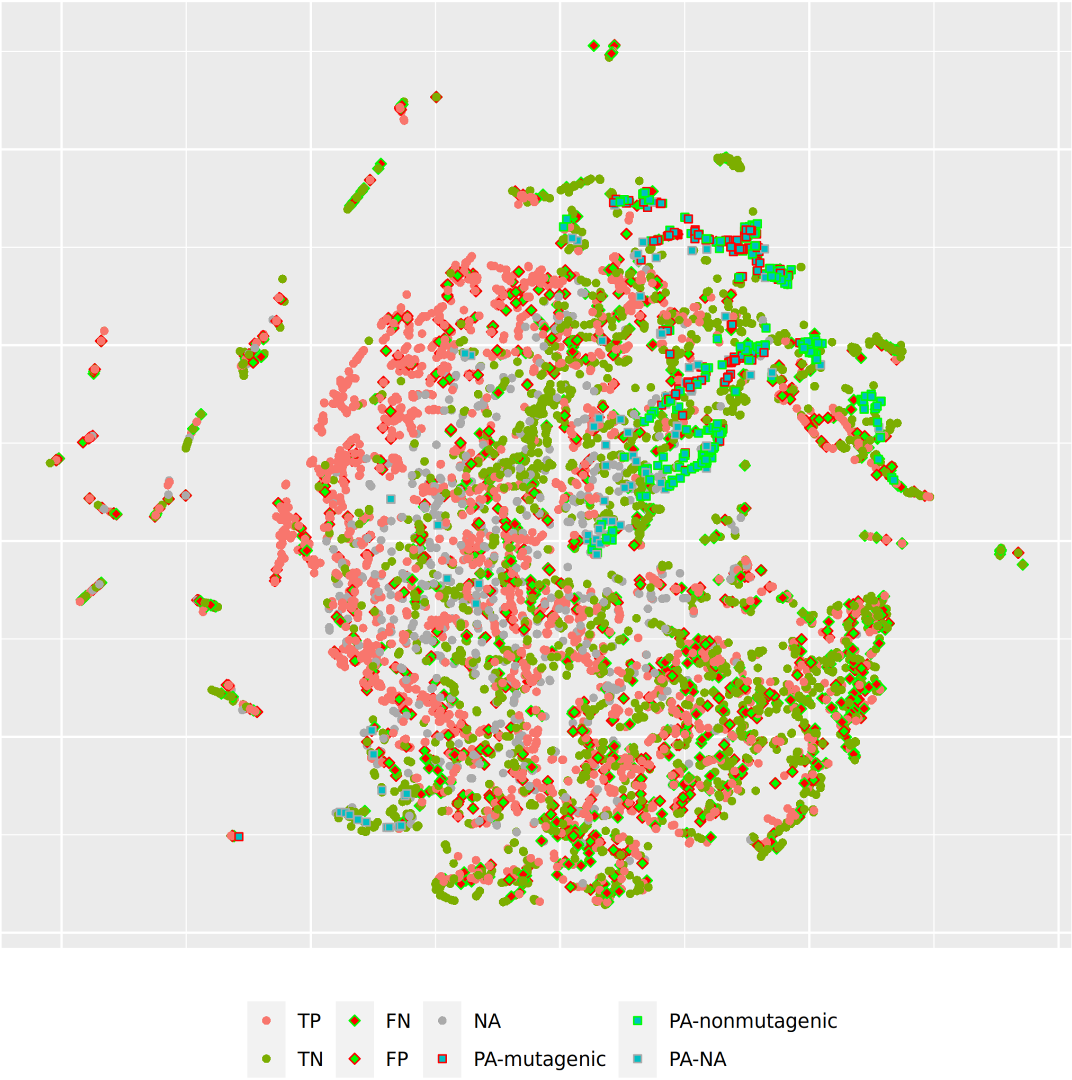


Figure 6: t-SNE visualisation of all CDK lazar predictions

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. The investigated 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.

Table 1, Table 2 and Figure 1 show that the crossvalidation accuracies of all models are comparable to the experimental variability of the *Salmonella typhimurium* mutagenicity bioassay (80-85% according to Benigni and Giuliani (1988)). All of these models have balanced sensitivity (true position rate) and specificity (true negative rate) and provide highly significant concordance with experimental data (as determined by McNemar's Test). This is a clear indication that *in-silico* predictions can be as reliable as the bioassays. Given that the variability of experimental data is similar to model variability it is impossible to decide which model gives the most accurate predictions, as models with higher accuracies might just approximate experimental errors better than more robust models.

Our results do not support the assumption that local models are superior to global models for classification purposes. For regression models (lowest observed effect level) we have found however that local models may outperform global models (Helma et al. (2018)) with accuracies similar to experimental variability.

As all investigated algorithms give similar accuracies the selection will depend more on practical considerations than on intrinsic properties. Nearest neighbor algorithms like `lazar` have the practical advantage that the rationales for individual predictions can be presented in a straightforward manner that is understandable without a background in statistics or machine learning (Figure 7). This allows a critical examination of individual predictions and prevents blind trust in models that are intransparent to users with a toxicological background.

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 fingerprints (e.g. MACCS, FP2-4).

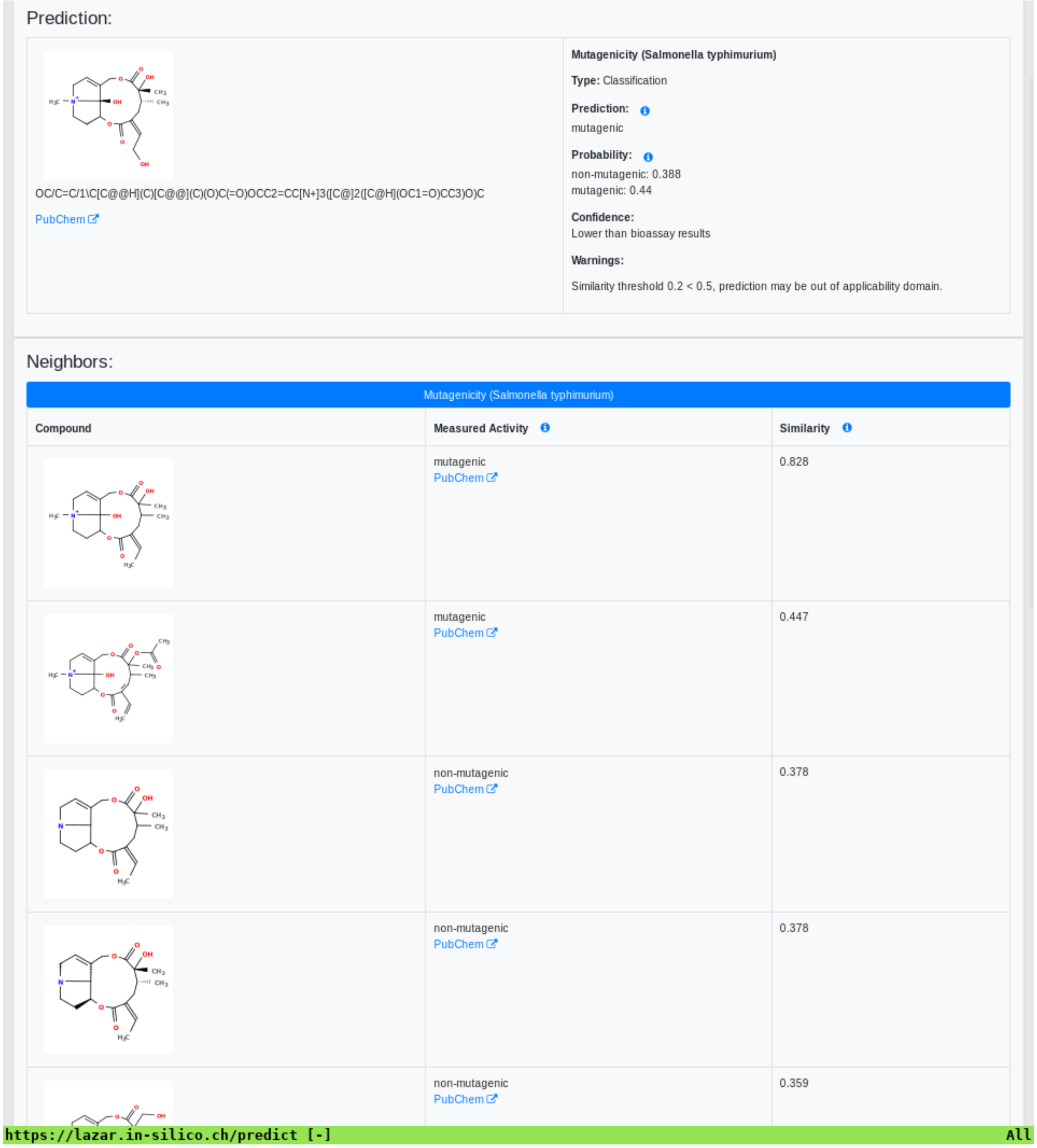
Chemistry Development Kit (CDK, Willighagen, Mayfield, and Alvarsson (2017)) descriptors were calculated with the PaDEL graphical interface (Yap (2011)). They include 1D and 2D topological descriptors as well as physical-chemical properties.

All investigated algorithms obtained models within the experimental variability for both types of descriptors (Table 1, Table 2, Figure 1).

Given that similar predictive accuracies are obtainable from both types of descriptors the choice depends once more on practical considerations:

MolPrint2D fragments can be calculated very efficiently for every well defined chemical structure with OpenBabel (O'Boyle et al. (2011)). CDK descriptor calculations are in contrast much more resource intensive and may fail for a significant number of compounds (from 8290).

MolPrint2D fragments are generated dynamically from chemical structures and can be



Prediction:

OC/C=C/1\C[C@@H](C)[C@@](C)(O)C(=O)OCC2=CC[N+]3([C@]2([C@H](OC1=O)CC3)O)C

PubChem

Mutagenicity (Salmonella typhimurium)

Type: Classification

Prediction:
mutagenic

Probability:
non-mutagenic: 0.388
mutagenic: 0.44

Confidence:
Lower than bioassay results

Warnings:

Similarity threshold 0.2 < 0.5, prediction may be out of applicability domain.

Neighbors:

Mutagenicity (Salmonella typhimurium)

Compound	Measured Activity	Similarity
	mutagenic PubChem	0.828
	mutagenic PubChem	0.447
	non-mutagenic PubChem	0.378
	non-mutagenic PubChem	0.378
	non-mutagenic PubChem	0.359

```
https://lazar.in-silico.ch/predict [-]                                                    All
```

Figure 7: Lazar screenshot of 12,21-Dihydroxy-4-methyl-4,8-secosenecinonan-8,11,16-trione mutagenicity prediction

used to determine if a compound contains structural features that are absent in training data. This feature can be used to determine applicability domains. CDK descriptors contain in contrast a predefined set of descriptors with unknown toxicological relevance.

MolPrint2D fingerprints can be represented very efficiently as sets of features that are present in a given compound which makes similarity calculations very efficient. Due to the large number of substructures present in training compounds, they lead however to large and sparsely populated datasets, if they have to be expanded to a binary matrix (e.g. as input for tensorflow models). CDK descriptors contain in contrast in every case matrices with 1442 columns which can cause substantial computational overhead.

Pyrrolizidine alkaloid mutagenicity predictions

Figure 2 shows a clear differentiation between the different pyrrolizidine alkaloid groups. The largest proportion of mutagenic predictions was observed for Otonecines 65% (407/623), the lowest for Monoesters 2% (52/1889) and N-Oxides 5% (59/1052).

Although most of the models show similar accuracies, sensitivities and specificities in crossvalidation experiments some of the models (MPD-RF, CDK-RF and CDK-SVM) predict a lower number of mutagens (2-5%) than the majority of the models (14-25% (Figure 2). lazar-CDK on the other hand predicts the largest number of mutagens for all groups with exception of Otonecines.

These differences between predictions from different algorithms and descriptors were not expected based on crossvalidation results.

In order to investigate, if any of the investigated models show systematic errors in the vicinity of pyrrolizidine-alkaloids we have performed a detailled t-SNE analysis of all models (see Figure 5 and Figure 6 for two examples, all visualisations can be found at https://git.in-silico.ch/mutagenicity-paper/figures.

Nevertheless none of the models showed obvious deviations from their expected behaviour, so the reason for the disagreement between some of the models remains unclear at the moment. It is however perfectly possible that some systematic errors are covered up by converting high dimensional spaces to two coordinates and are thus invisible in t-SNE visualisations.

Conclusions

A new public *Salmonella* mutagenicity training dataset with 8309 compounds was created and used it to train `lazar` and Tensorflow models with MolPrint2D and CDK descriptors.

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. 2011. "Scientific Opinion on Pyrrolizidine Alkaloids in Food and Feed." *EFSA Journal*, no. 9: 1–134.

———. 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.

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.

Mattocks, AR. 1986. *Chemistry and Toxicology of Pyrrolizidine Alkaloids*. Academic Press.

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.

Schöning, Verena, Felix Hammann, Mark Peinl, and Jürgen Drewe. 2017. "Editor's Highlight: Identification of Any Structure-Specific Hepatotoxic Potential of Different Pyrrolizidine Alkaloids Using Random Forests and Artificial Neural Networks." *Toxicol. Sci.*, no. 160: 361–70.

Weininger, David, Arthur Weininger, and Joseph L. Weininger. 1989. "SMILES. 2. Algorithm for Generation of Unique Smiles Notation." *J. Chem. Inf. Comput. Sci.*, no. 29: 97–101. https://doi.org/https://doi.org/10.1021/ci00062a008.

Willighagen, E. L., J. W. Mayfield, and J. et al. Alvarsson. 2017. "The Chemistry

Development Kit (Cdk) V2.0: Atom Typing, Depiction, Molecular Formulas, and Substructure Searching." *J. Cheminform.*, no. 9(33). https://doi.org/https://doi.org/10.1186/s13321-017-0220-4.

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