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[4], 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
[4]Clinical Pharmacology and Toxicology, Department of General Internal Medicine, Bern University Hospital, 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.

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 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*) 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.

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 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.

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

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

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

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

Applicability domain

The applicability domain (AD) of `lazar` models is determined by the structural diversity of the training data. If no similar compounds are found in the training data no predictions will be generated. Warnings are issued if the similarity threshold had to be lowered from 0.5 to 0.2 in order to enable predictions. Predictions without warnings can be considered as close to the applicability domain (*high confidence*) and predictions with warnings as more distant from the applicability domain (*low confidence*). Quantitative applicability domain information can be obtained from the similarities of individual neighbours.

Availability

- 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

TODO: **Philipp** Kannst Du bitte die folgenden Absaetze ergaenzen und die Vorgangsweise fuer MP2D/CDK bzw CV/PA Vorhersagen beschreiben.

Random forests (*RF*)

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

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

Neural Nets (*NN*)

Support vector machines (*SVM*)

Validation

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

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	88	85	78	83	83	82	83
True negative rate	78	79	82	84	85	85	86
Positive predictive value	82	80	81	84	84	84	85
Negative predictive value	85	84	80	84	84	83	84
Nr. predictions	6300	7777	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	52	52	84	79	80	85	82
True positive rate	90	90	81	81	80	85	82
True negative rate	14	14	86	78	80	85	82
Positive predictive value	52	52	85	79	80	85	82
Negative predictive value	56	56	82	80	80	85	82
Nr. predictions	811	811	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/.

With exception of lazar/CDK 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 from all investigated models for 602 pyrrolizidine alkaloids (PAs) can be downloaded from https://git.in-silico.ch/mutagenicity-paper/tree/pyrrolizidine-alkaloids/pa-predictions.csv. A visual representation of all PA predictions

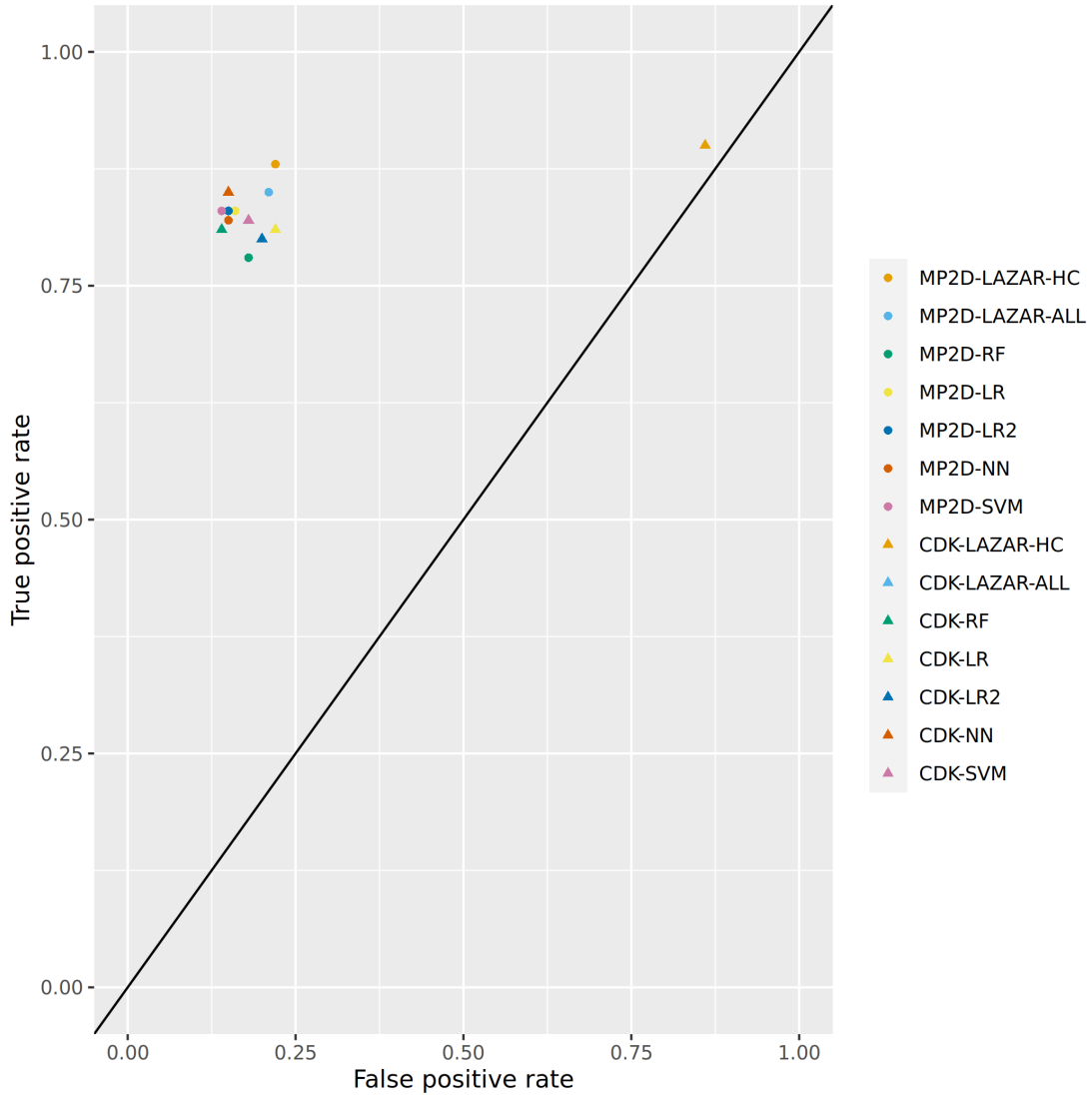

Figure 1: ROC plot of crossvalidation results (lazar-HC: lazar with high confidence, lazar-all: all lazar predictions, RF: random forests, LR-sgd: logistic regression (stochastic gradient descent), LR-scikit: logistic regression (scikit), NN: neural networks, SVM: support vector machines).

can be found at https://git.in-silico.ch/mutagenicity-paper/tree/pyrrolizidine-alkaloids/pa-predictions.pdf.

Table 3 and Table 4 summarise the outcome of pyrrolizidine alkaloid predictions from all models with MolPrint2D and CDK descriptors.

Table 3: Summary of MolPrint2D pyrrolizidine alkaloid predictions

Model	mutagenic	non-mutagenic	Nr. predictions
lazar-all	20% (111)	80% (449)	93% (560)
lazar-HC	25% (76)	75% (225)	50% (301)
RF	5% (28)	95% (574)	100% (602)
LR-sgd	21% (127)	79% (475)	100% (602)
LR-scikit	20% (118)	80% (484)	100% (602)
NN	21% (124)	79% (478)	100% (602)
SVM	14% (82)	86% (520)	100% (602)

Table 4: Summary of CDK pyrrolizidine alkaloid predictions

Model	mutagenic	non-mutagenic	Nr. predictions
lazar-all	20% (111)	80% (449)	93% (560)
lazar-HC	25% (76)	75% (225)	50% (301)
RF	2% (10)	98% (592)	100% (602)
LR-sgd	16% (97)	84% (505)	100% (602)
LR-scikit	15% (88)	85% (514)	100% (602)
NN	25% (150)	75% (452)	100% (602)
SVM	3% (19)	97% (583)	100% (602)

Figure 2 - Figure 10 display the proportion of positive mutagenicity predictions from all

models for the different pyrrolizidine alkaloid groups.

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, 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.

Figure 11 shows the t-SNE of pyrrolizidine alkaloids (PA) and the mutagenicity training data in MP2D space (Tanimoto/Jaccard similarity).

Figure 12 shows the t-SNE of pyrrolizidine alkaloids (PA) and the mutagenicity training data in CDK 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 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.

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

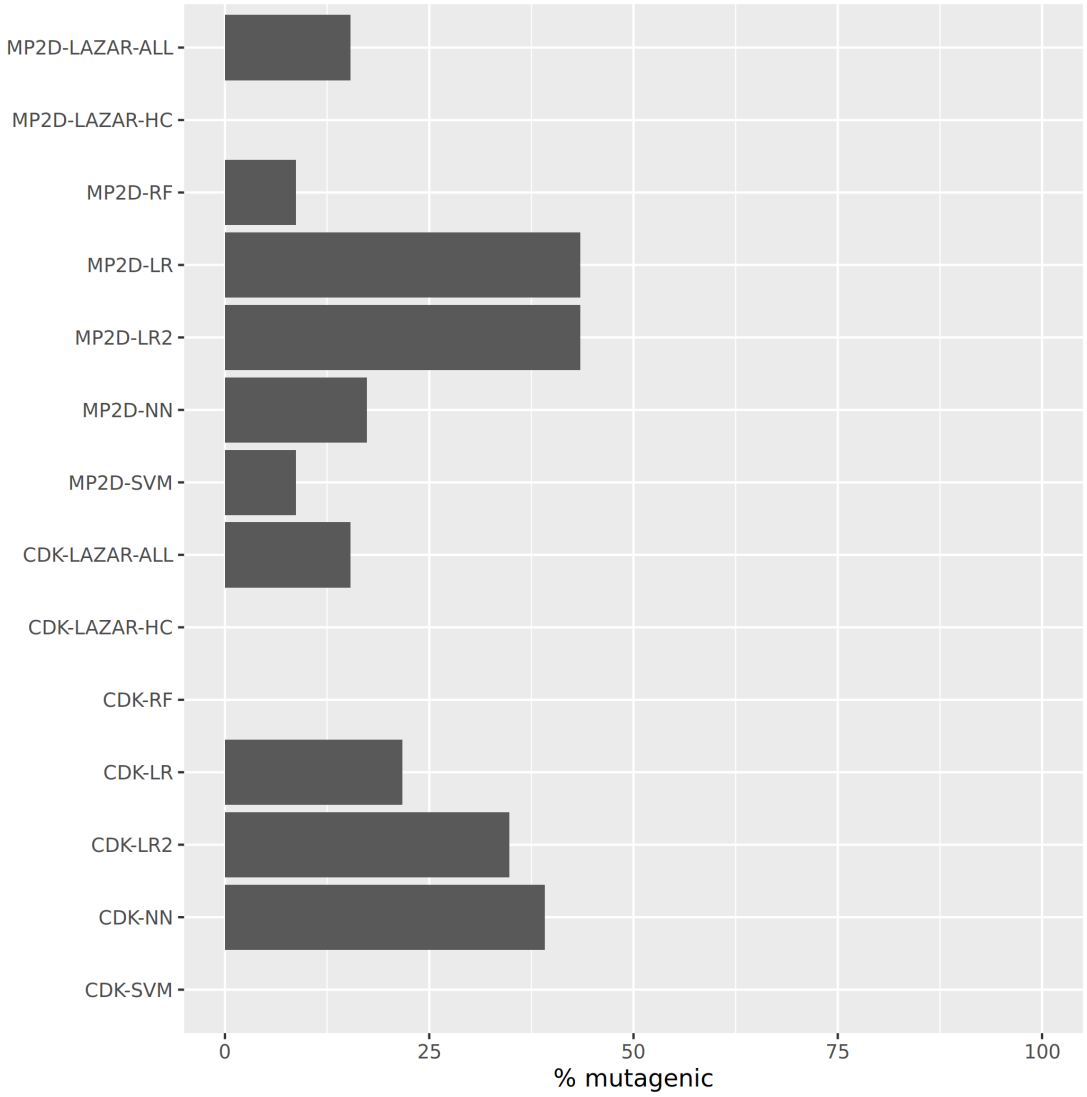

Figure 2: Summary of Dehydropyrrolizidine predictions

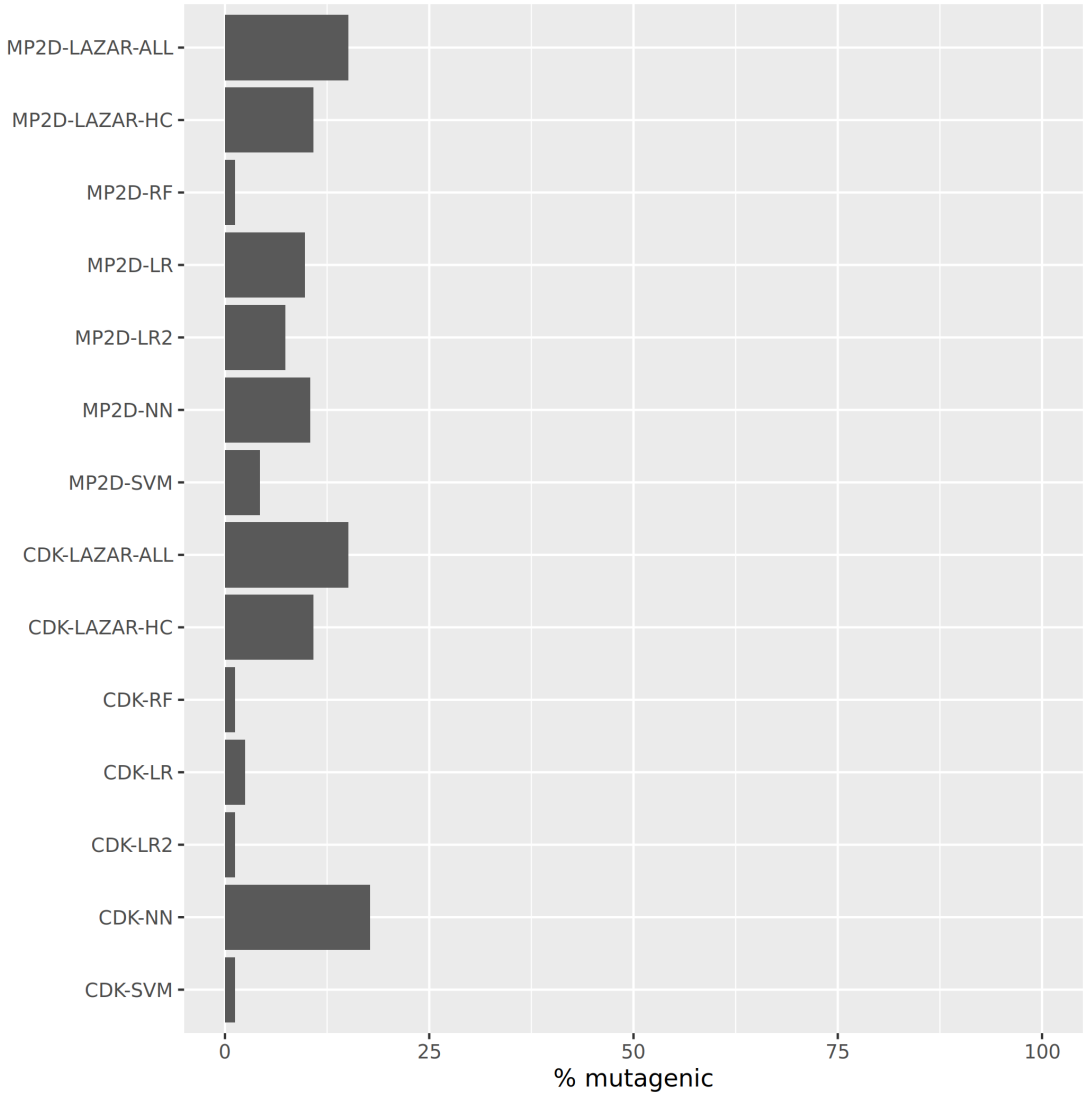

Figure 3: Summary of Diester predictions

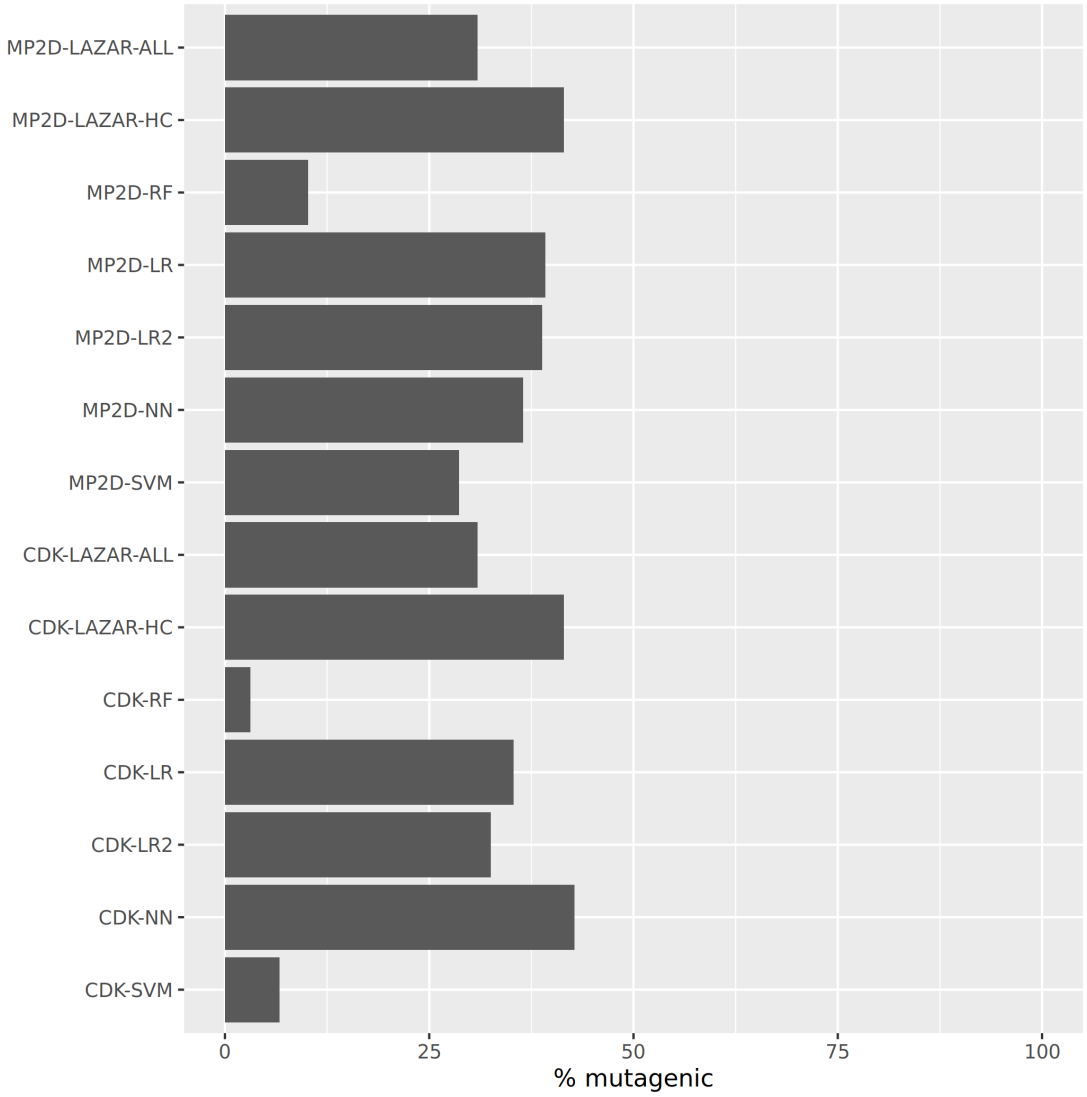

Figure 4: Summary of Macrocyclic-diester predictions

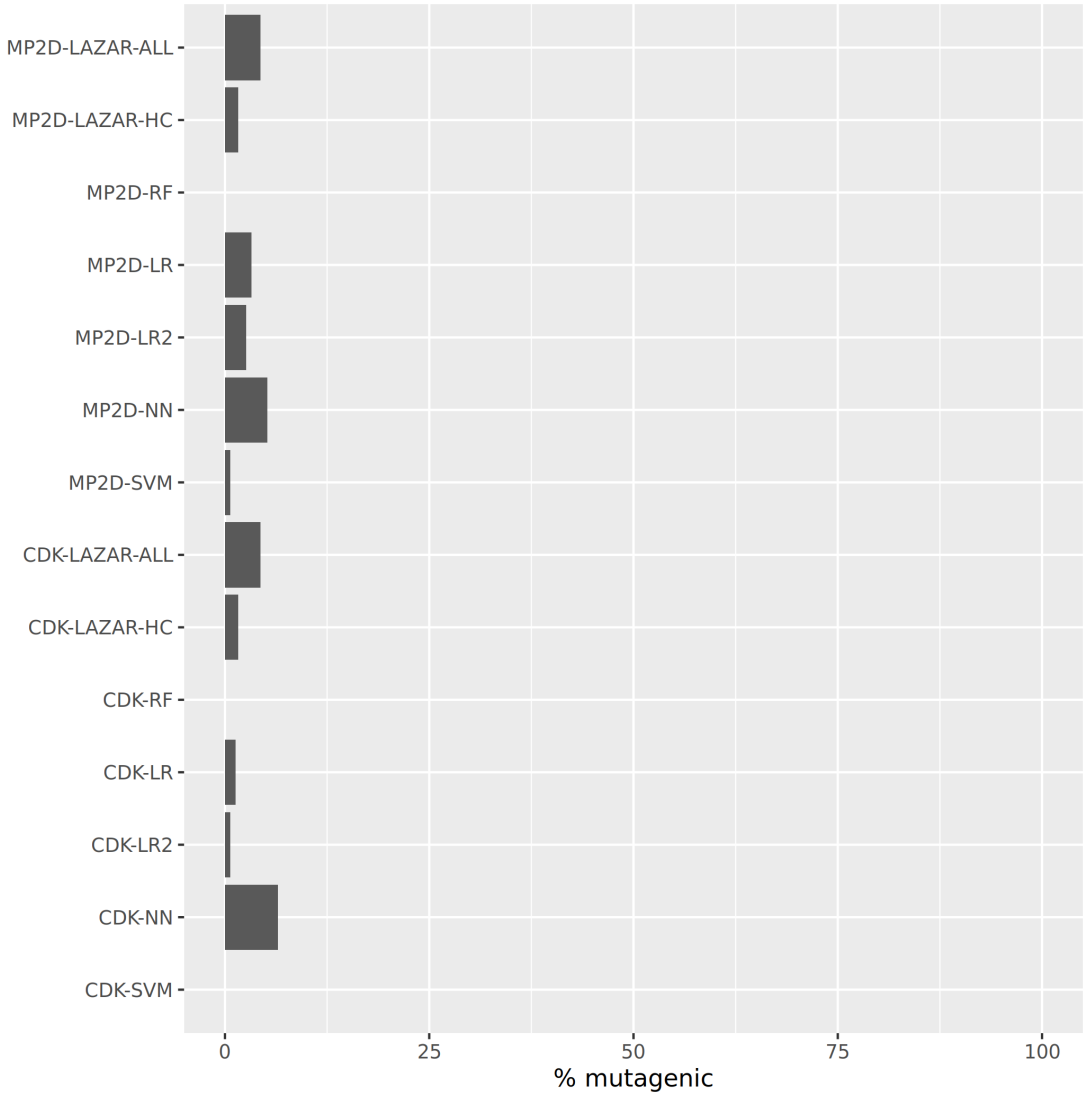

Figure 5: Summary of Monoester predictions

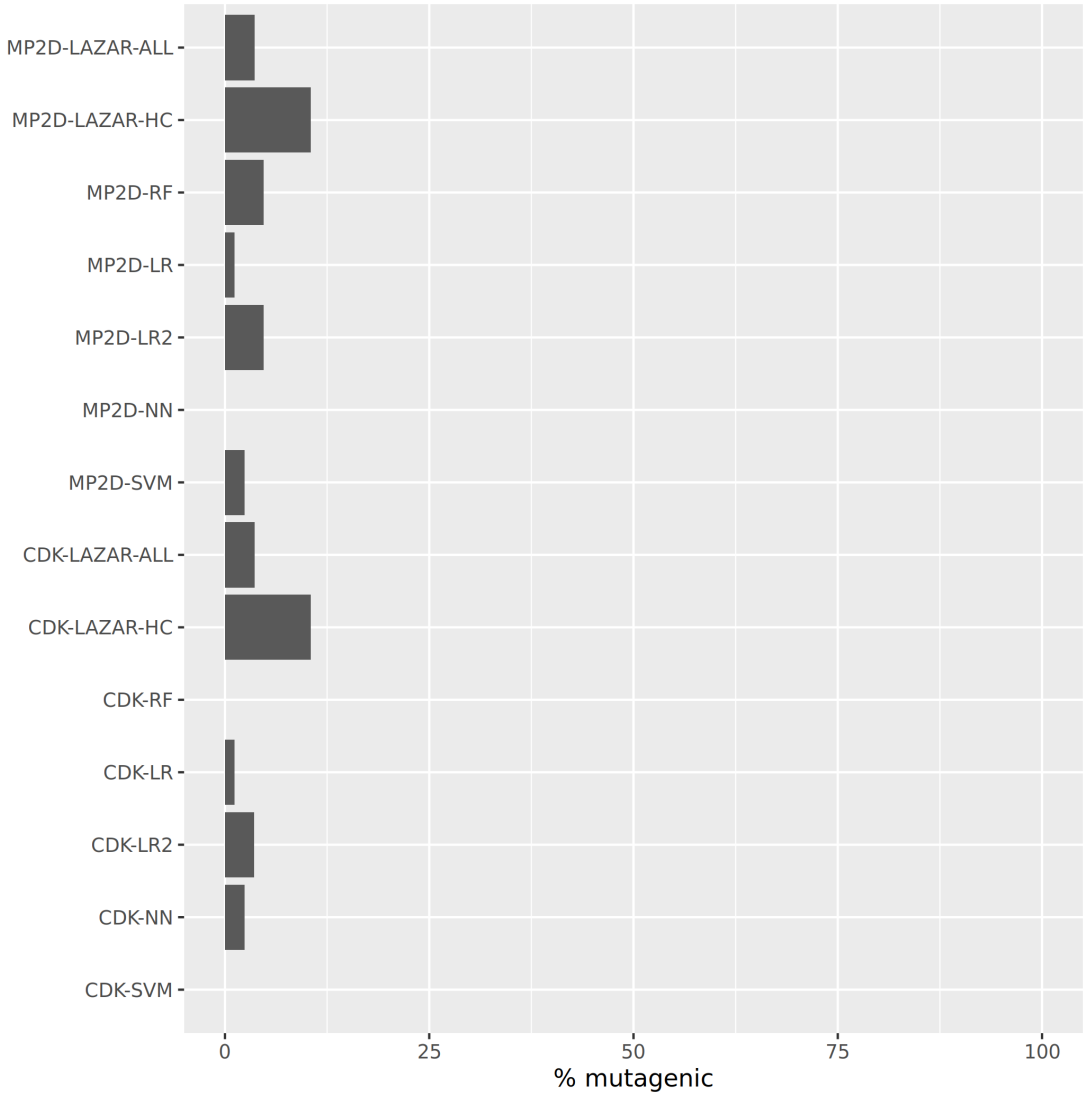

Figure 6: Summary of N-oxide predictions

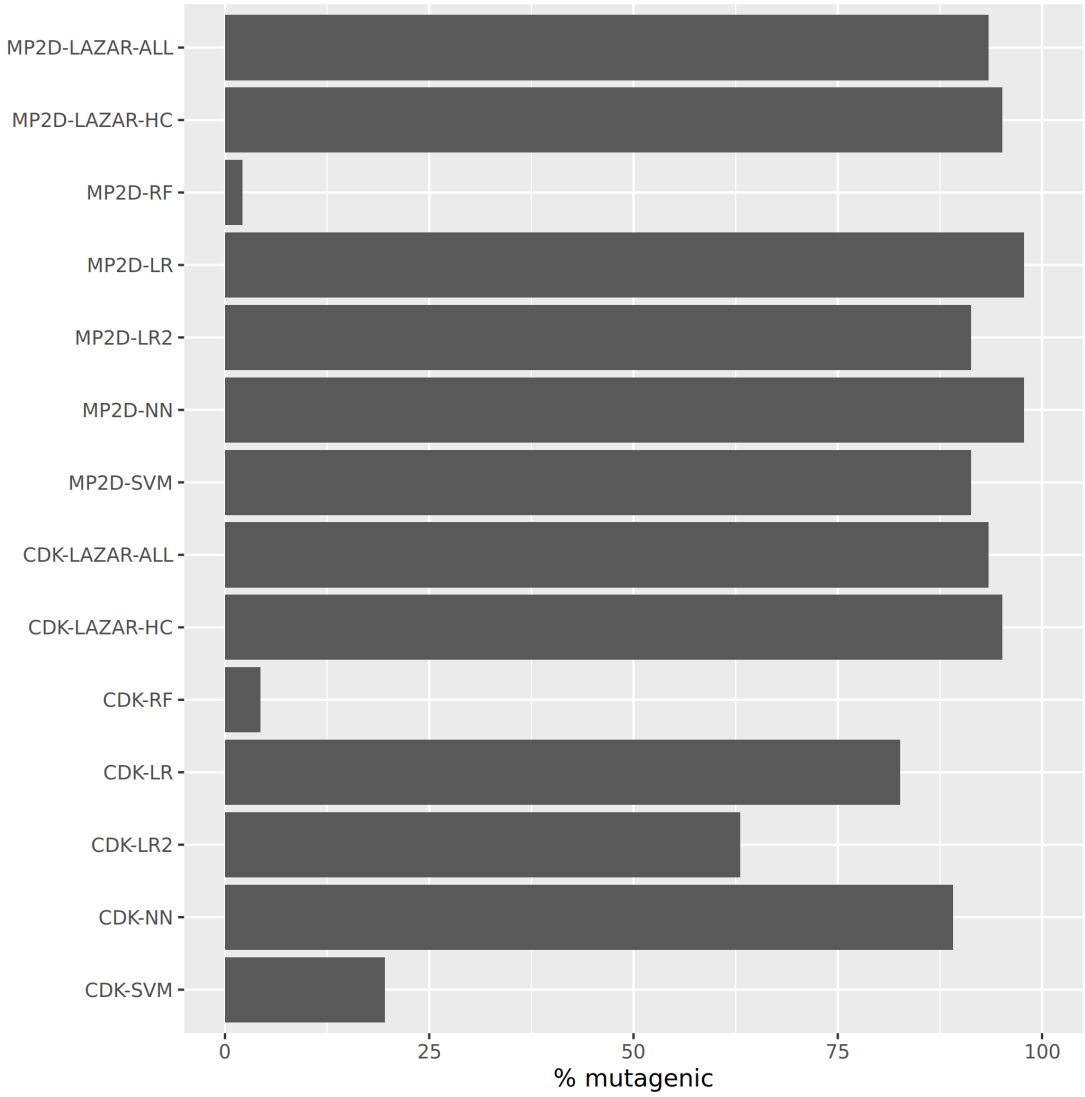

Figure 7: Summary of Otonecine predictions

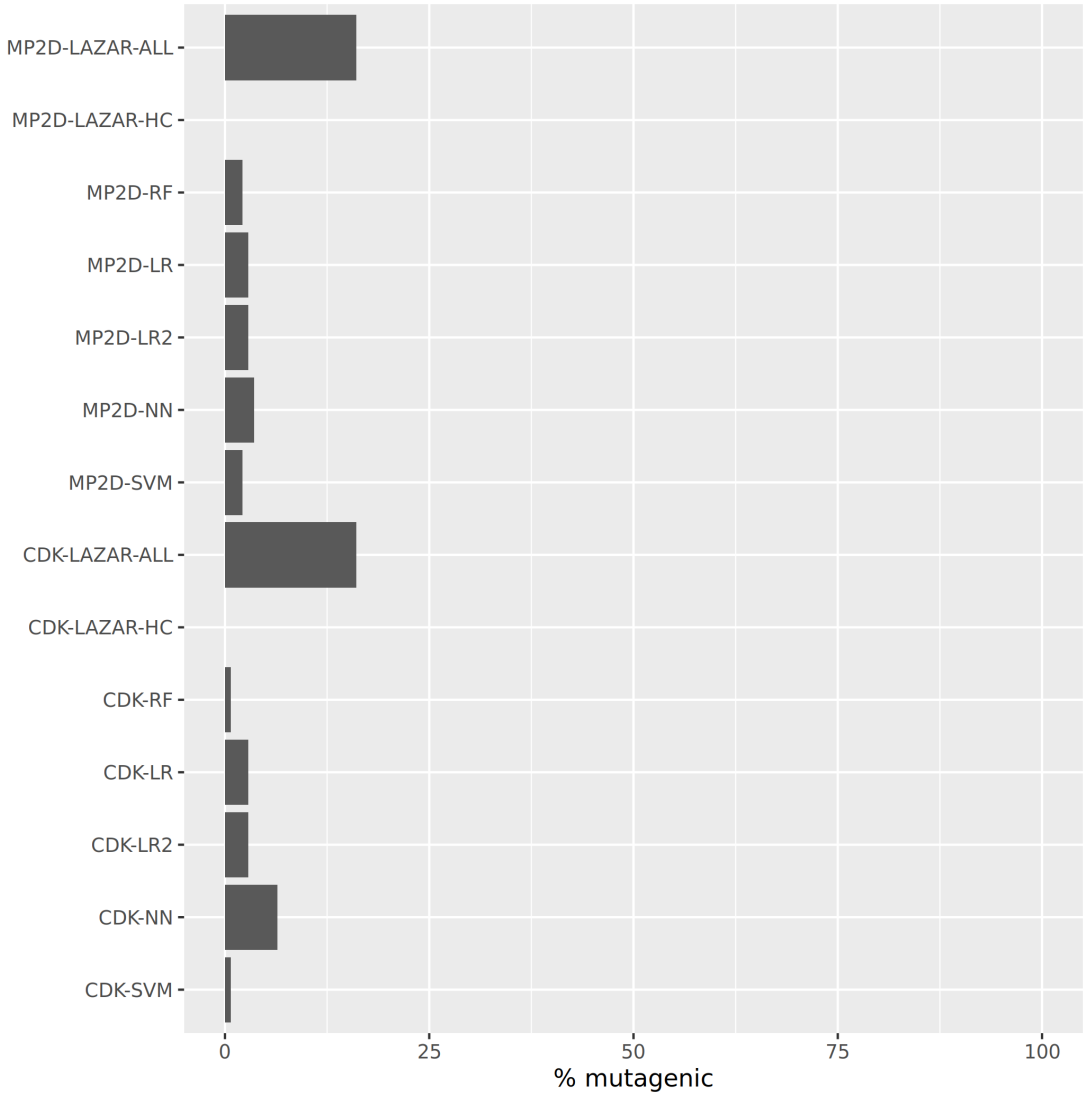

Figure 8: Summary of Platynecine predictions

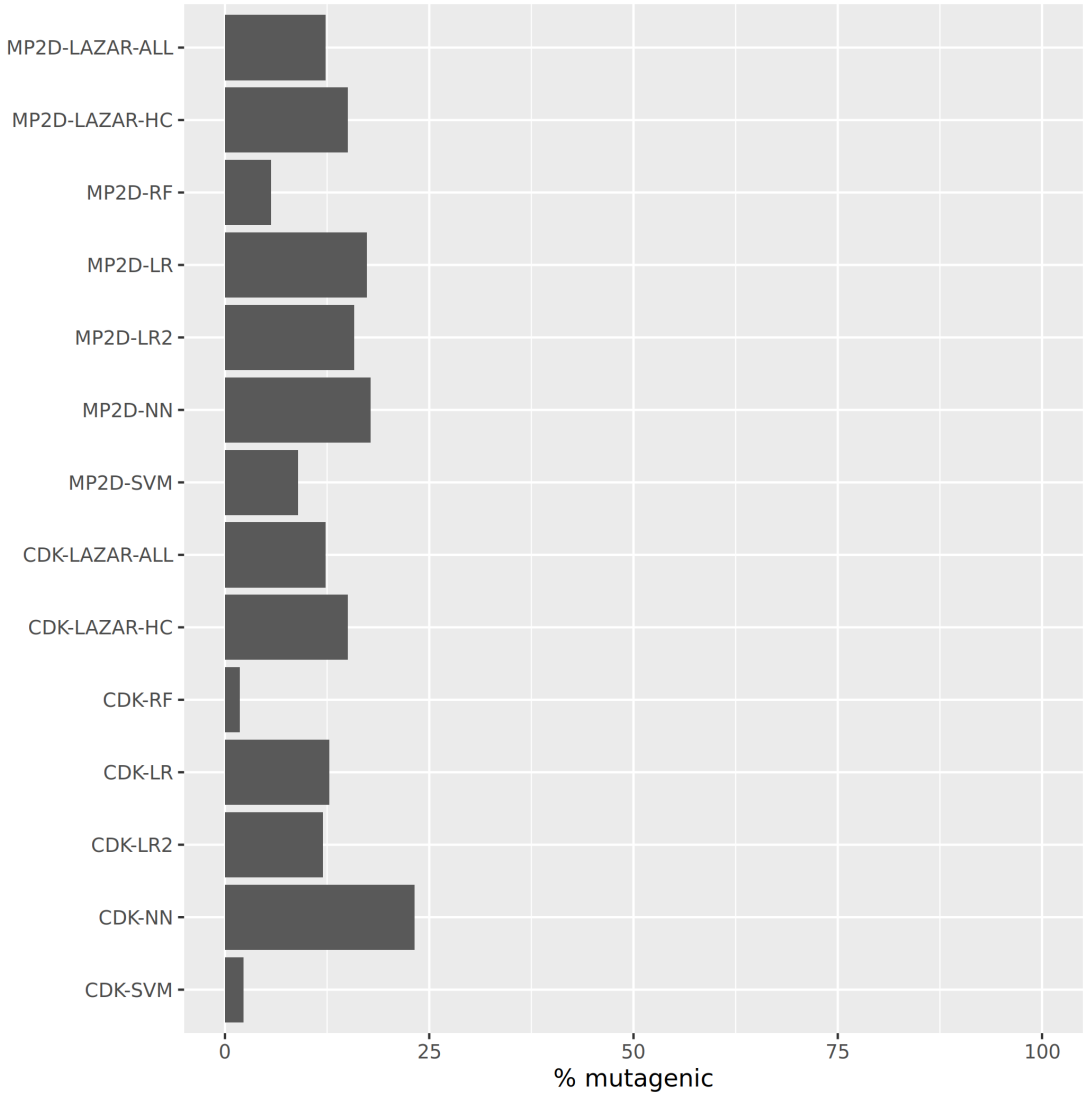

Figure 9: Summary of Retronecine predictions

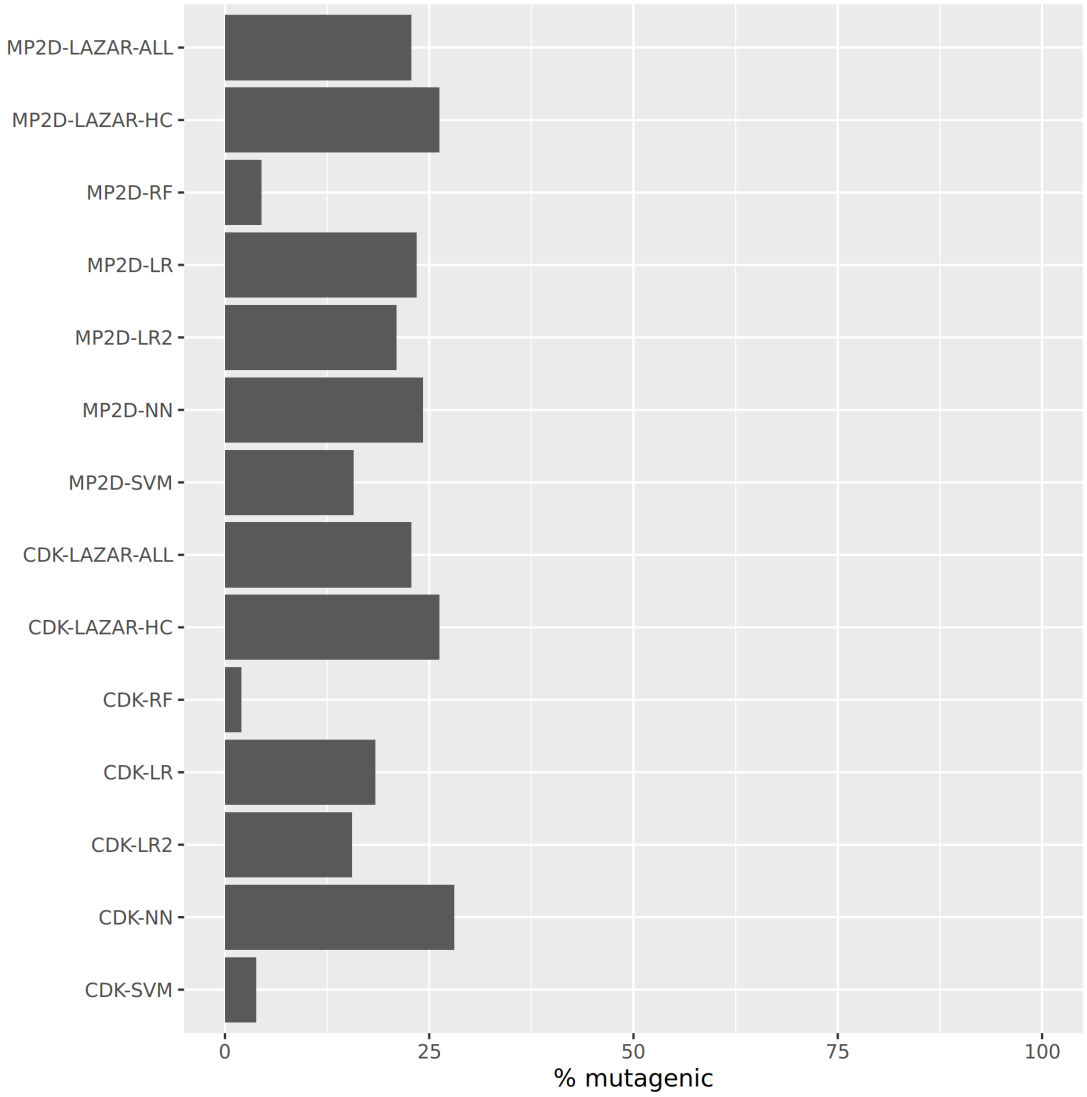

Figure 10: Summary of Tertiary PA predictions

Figure 11: t-SNE visualisation of mutagenicity training data and pyrrolizidine alkaloids (PA)

Figure 12: t-SNE visualisation of mutagenicity training data and pyrrolizidine alkaloids (PA)

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 all models with the exception of lazar-CDK have similar crossvalidation accuracies that 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 (e.g. NN-CDK) might just approximate experimental errors better than more robust models.

`lazar` predictions with CDK descriptors are a notable exception, as it has a much lower overall accuracy () than all other models. `lazar` uses basically a k-nearest-neighbor (with variable k) and it seems that CDK descriptors are not very well suited for chemical similarity calculations. We have confirmed this independently by validating k-nn models from the `R caret` package, which give also sub-par accuracies (data not shown).

Figure 12 is another indication that similarity calculations with CDK descriptors are not as useful as fingerprint based similarities, because it shows a less clearer separation between chemical classes and mutagens/non-mutagens than Figure 11. It seems that more complex models than simple k-nn are required to utilize CDK descriptors efficiently.

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.

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 E. L. (2017)) descriptors were calculated with the PaDEL graphical interface (Yap (2011)). They include 1D and 2D topological descriptors as well as physical-chemical properties.

With exception of `lazar` all investigated algorithms obtained models within the experimental variability for both types of descriptors. As discussed before CDK descriptors seem to be less suitable for chemical similarity calculations than MolPrint2D descriptors.

Given that similar predictive accuracies are obtainable from both types of descriptors the choice depends 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 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.

Pyrrolizidine alkaloid mutagenicity predictions

Figure 2 - Figure 10 show a clear differentiation between the different pyrrolizidine alkaloid groups. The largest proportion of mutagenic predictions was observed for Otonecines 72% (458/634), the lowest for Monoesters 2% (45/1940) and N-Oxides 2% (27/1044).

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% Table 3, Table 4, Figure 2 - Figure 10).

From a practical point we still have to face the question, how to choose model predictions, if no experimental data is available (we found two PAs in the training data, but this number is too low, to draw any general conclusions).

TODO: **Verena** Hier ist ein alter Text von Dir zum Recylen:

Necic acid

The rank order of the necic acid is comparable in the four models considered (LAZAR, RF and DL (R-project and Tensorflow). PAs from the monoester type had the lowest genotoxic potential, followed by PAs from the open-ring diester type. PAs with macro-

cyclic diesters had the highest genotoxic potential. The result fit well with current state of knowledge: in general, PAs, which have a macrocyclic diesters as necic acid, are considered more toxic than those with an open-ring diester or monoester EFSA 2011Fu et al. 2004Ruan et al. 2014b(; ;).

Necine base

The rank order of necine base is comparable in LAZAR, RF, and DL (R-project) models: with platynecine being less or as genotoxic as retronecine, and otonecine being the most genotoxic. In the Tensorflow-generate DL model, platynecine also has the lowest genotoxic probability, but are then followed by the otonecines and last by retronecine. These results partly correspond to earlier published studies. Saturated PAs of the platynecine-type are generally accepted to be less or non-toxic and have been shown in *in vitro* experiments to form no DNA-adducts Xia et al. 2013(). Therefore, it is striking, that 1,2-unsaturated PAs of the retronecine-type should have an almost comparable genotoxic potential in the LAZAR and DL (R-project) model. In literature, otonecine-type PAs were shown to be more toxic than those of the retronecine-type Li et al. 2013().

Modifications of necine base

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

In the LAZAR model, the genotoxic potential of dehydropyrrolizidines (DHP) (using the extended AD) is comparable to that of tertiary PAs. Since, DHP is regarded as the toxic principle in the metabolism of PAs, and known to produce protein- and DNA-adducts Chen et al. 2010(), the LAZAR model did not meet this expectation it predicted the majority of DHP as being not genotoxic. However, the following issues need to be considered. On the one hand, all DHP were outside of the stricter AD of 0.5. This indicates that in general, there might be a problem with the AD. In addition, DHP has

two unsaturated double bounds in its necine base, making it highly reactive. DHP and other comparable molecules have a very short lifespan, and usually cannot be used in *in vitro* experiments. This might explain the absence of suitable neighbours in LAZAR.

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

In the DL (R-project) and RF model, *N*-oxides have a by far more genotoxic potential that tertiary PAs or dehydropyrrolizidines. As PA *N*-oxides are easily conjugated for extraction, they are generally considered as detoxification products, which are *in vivo* quickly renally eliminated Chen et al. 2010(). On the other hand, *N*-oxides can be also back-transformed to the corresponding tertiary PA Wang et al. 2005(). Therefore, it may be questioned, whether *N*-oxides themselves are generally less genotoxic than the corresponding tertiary PAs. However, in the groups of modification of the necine base, dehydropyrrolizidine, the toxic principle of PAs, should have had the highest genotoxic potential. Taken together, the predictions of the modifications of the necine base from the LAZAR, RF and R-generated DL model cannot - in contrast to the Tensorflow DL model - be considered as reliable.

Overall, when comparing the prediction results of the PAs to current published knowledge, it can be concluded that the performance of most models was low to moderate. This might be contributed to the following issues:

1. In the LAZAR model, only 26.6% PAs were within the stricter AD. With the extended AD, 92.3% of the PAs could be included in the prediction. Even though the Jaccard distance between the training dataset and the PA dataset for the RF,

SVM, and DL (R-project and Tensorflow) models was small, suggesting a high similarity, the LAZAR indicated that PAs have only few local neighbours, which might adversely affect the prediction of the mutagenic potential of PAs.

2. All above-mentioned models were used to predict the mutagenicity of PAs. PAs are generally considered to be genotoxic, and the mode of action is also known. Therefore, the fact that some models predict the majority of PAs as not genotoxic seems contradictory. To understand this result, the basis, the training dataset, has to be considered. The mutagenicity of in the training dataset are based on data of mutagenicity in bacteria. There are some studies, which show mutagenicity of PAs in the AMES test Chen et al. 2010(). Also, Rubiolo et al. (1992) examined several different PAs and several different extracts of PA-containing plants in the AMES test. They found that the AMES test was indeed able to detect mutagenicity of PAs, but in general, appeared to have a low sensitivity. The pre-incubation phase for metabolic activation of PAs by microsomal enzymes was the sensitivity-limiting step. This could very well mean that this is also reflected in the QSAR models.

Conclusions

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

Willighagen E. L., Alvarsson J. et al., Mayfield J. W. 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.