Decoding Toxicity of Small Molecules by Selecting Appropriate Molecular Descriptors

Döring K[1], **Grüger LM**[1], **Grüning BA**[2], **Günther S**[1]
kersten.doering@pharmazie.uni-freiburg.de

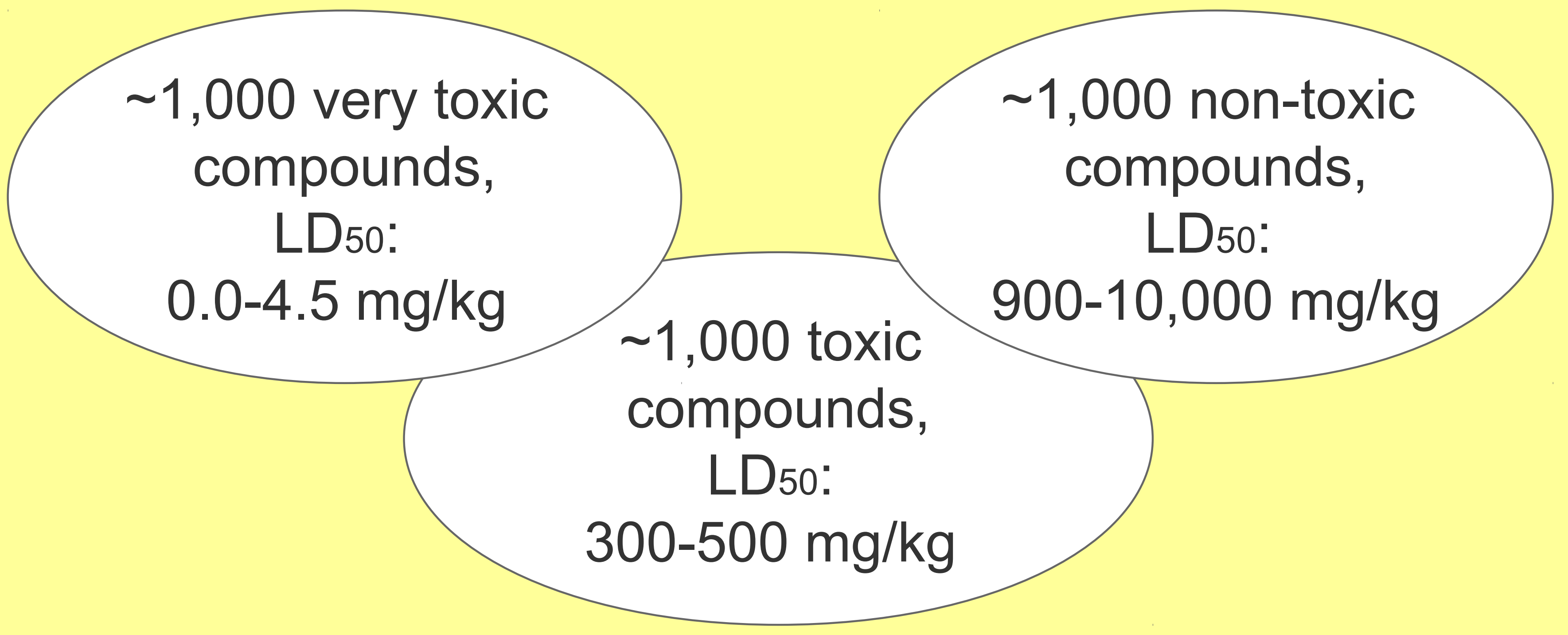
[1] Pharmaceutical Bioinformatics, Institute of Pharmaceutical Sciences, University of Freiburg
[2] Bioinformatics, Department of Computer Science, University of Freiburg

Introduction

An alternative approach to the common way of animal tests is the use of *in silico* methods for detecting toxicological effects [1]. We combine different machine learning algorithms with the generation of molecular descriptors for the prediction of a compound's toxicity. A molecular descriptor can be a value referring to a physicochemical property (e.g. molecular weight or octanol-water partition coefficient), number of functional groups (e.g. hydroxyl or carboxyl groups), or other predefined substructures.

> ***"... in silico methods as an alternative to animal testing."***

Data set



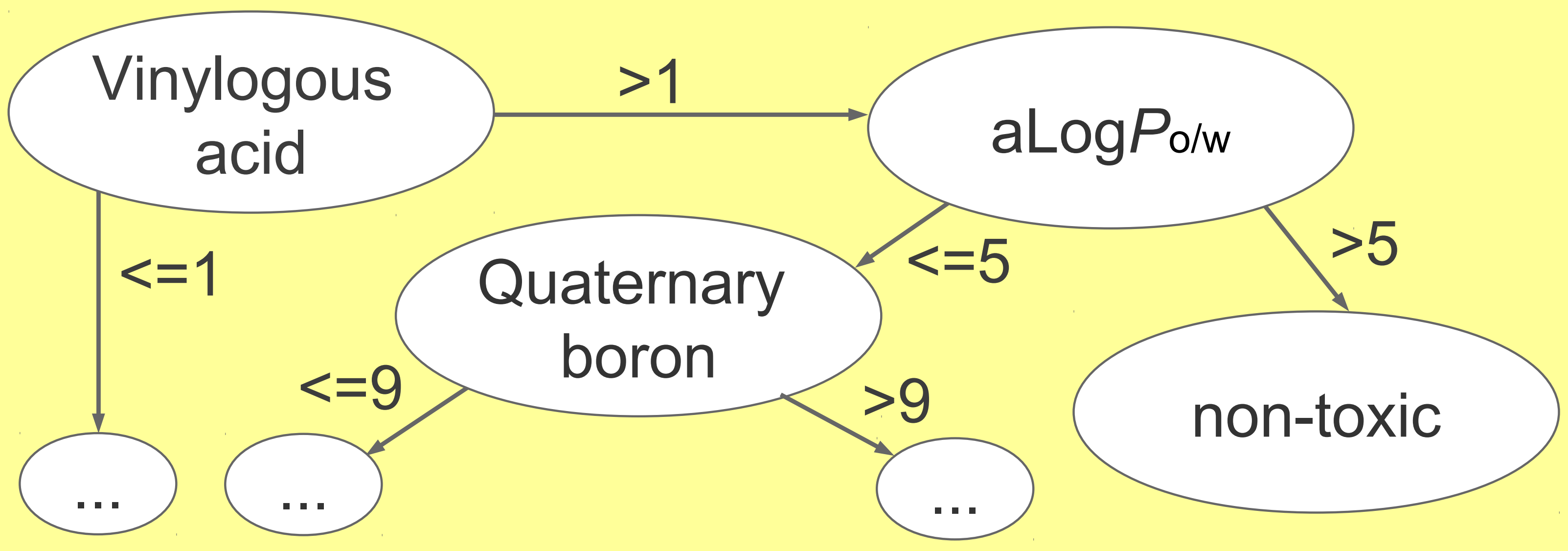
~1,000 very toxic compounds, LD_{50}: 0.0-4.5 mg/kg

~1,000 toxic compounds, LD_{50}: 300-500 mg/kg

~1,000 non-toxic compounds, LD_{50}: 900-10,000 mg/kg

The data set consists of a subset of an in-house database of around 20,000 chemical compounds, tested intravenously on mice (LD_{50}).

Decision Tree



Vinylogous acid

>1

aLog$P_{o/w}$

<=1

<=5

>5

Quaternary boron

non-toxic

<=9

>9

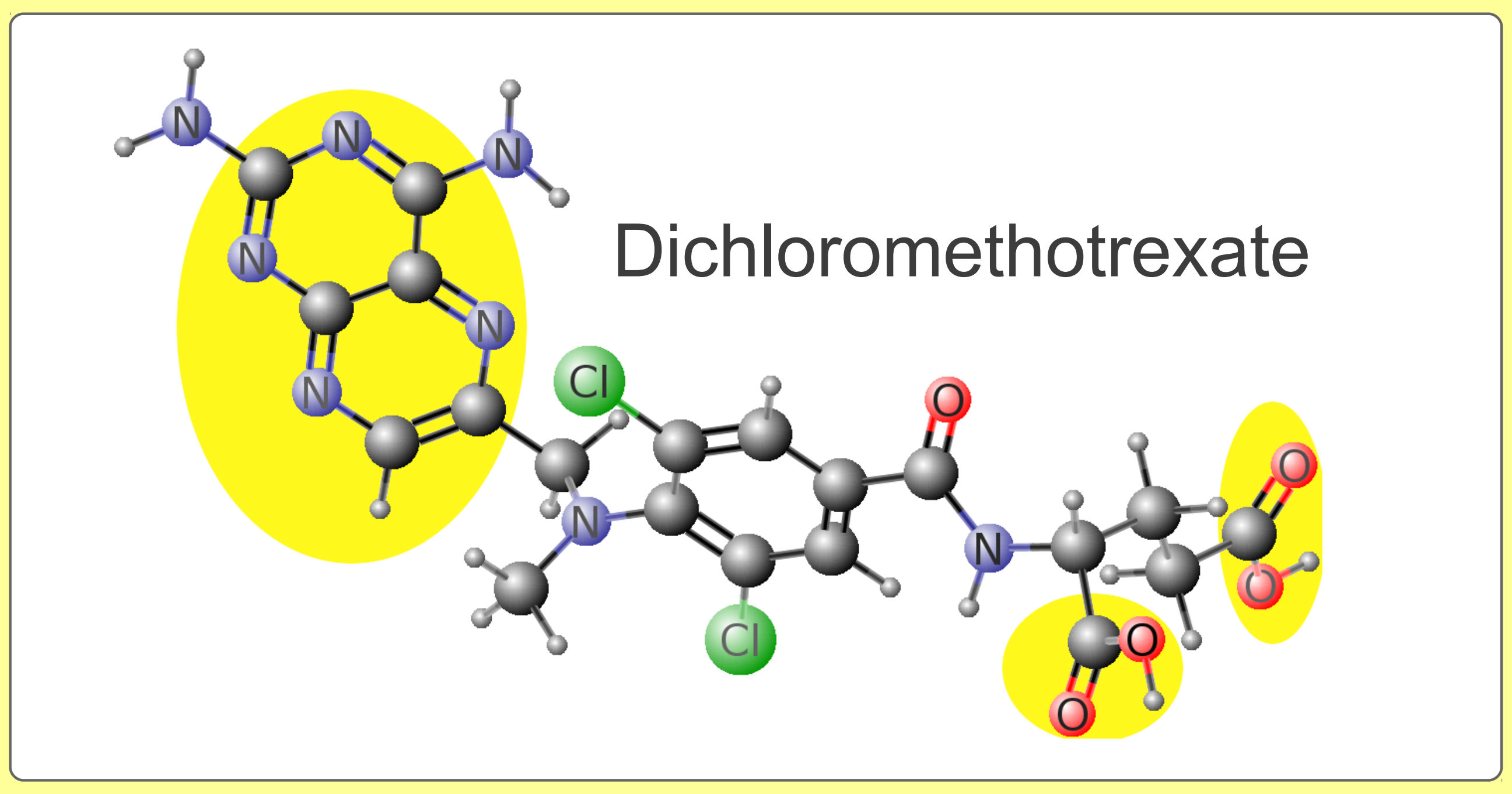
...

...

...

324 molecular descriptors were generated with OpenBabel [2]. WEKA's J48 descision tree selects branches based on information gain [3]. Only three out of many descriptors are shown as an example decision tree from different models.

Example



Dichloromethotrexate

The reported LD_{50} of dichloromethotrexate is 1,021 mg/kg body weight (non-toxic). It contains two times the descriptor *Carboxylic acid* and one annelated ring structure indicated by the circled areas. The outcome non-toxic is based on the simultaneous occurrence of these substructures and the absence of other descriptors.

Results

Class.	Acc. vt/nt	Spec. vt/nt	Sens. vt/nt	Acc. vt/t	Spec. vt/t	Sens. vt/t	Acc. t/nt	Spec. t/nt	Sens. t/nt
ANN	0.82	0.79	0.85	0.61	0.58	0.66	0.56	0.69	0.54
DT	0.91	0.92	0.90	0.83	0.83	0.82	0.77	0.77	0.77
RF	0.91	0.93	0.90	0.87	0.90	0.84	0.79	0.80	0.78
SVM	0.82	0.81	0.84	0.75	0.75	0.76	0.70	0.73	0.68

The classifiers (class.) are artificial neural network (ANN), decision tree (DT), random forest (RF), and support vector machine (SVM). There are three different models (10-fold cross-validation) for each classifier: very toxic vs. non-toxic (vt/nt), very toxic vs. toxic (vt/t), and toxic vs. non-toxic (t/nt). Model vt/nt showed best results and classifier RF performed best for all models. Accuracy, specificity, and sensitivity:

$$Acc = \frac{(\#TP + \#TN)}{(\#TP + \#TN + \#FP + \#FN)} \quad \Big| \quad Spec = \frac{(\#TN)}{(\#TN + \#FP)} \quad \Big| \quad Sens = \frac{(\#TP)}{(\#TP + \#FN)}$$

References

[1] E. Mombelli, 2008. An evaluation of the predictive ability of the QSAR software packages, DEREK, HAZARDEXPERT and TOPKAT, to describe chemically-induced skin irritation. Altern Lab Anim 36:15-24.
[2] http://openbabel.org/wiki/Main_Page
[3] http://www.cs.waikato.ac.nz/ml/weka
[4] C. Steinbeck et al., 2006. Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. Curr Pharm Des. 12:2111-20.
[5] http://www.schrodinger.com

Future Prospects

Although it remains unclear what toxicity is based on in terms of molecular descriptors, first results on this small data set are promising. The set of descriptors will be extended with the those from Chemistry Development Kit [4] and QikProp [5]. Ongoing results will be evaluated with knowledge from literature.

The working group of Pharmaceutical Bioinformatics at the Institute for Pharmaceutical Sciences develops algorithms and software for pharmaceutical research. Our fields of research include the modeling of molecular interactions, prediction of biological effects of molecules, identification of potential new drug agents, analysis of gene expression and methylation data as well as text and data mining. The working group is part of the University of Freiburg's Research Group Program of the Excellence Initiative of the federal and state governments.

http://www.pharmaceutical-bioinformatics.com

This work is part of the CoRS project, which is funded by the German National Research Foundation (DFG, Lis45).

DFG Deutsche Forschungsgemeinschaft

UNI FREIBURG