Comparative Modelling of Biodegradability Data

Susann Vorberg, Igor V. Tetko
Institute of Structural Biology, Helmholtzzentrum Munich

HelmholtzZentrum münchen
German Research Center for Environmental Health

Motivation

Biodegradability describes the capacity of substances to be mineralized by free living bacteria. It is a crucial property to estimate a compounds long-term impact on the environment. In order to substitute laborious experimental testing, it is of interest to reliably predict biodegradability. Yet, this endpoint is difficult to model due to inconsistency of available experimental data. Our approach makes use of OCHEM's [2] rich supply of machine learning methods and descriptor sets to build classification models for biodegradability.

Data Set

The initial dataset has been arranged from several sources:

1. ~**1400** measurements extracted from the CHRIP (Chemical Risk Information Platform) and ECHA (European Chemical Agency) database.

2. ~**1500** measurements assembled by Cheng at al.[1] from the Japanese NITE database and the BIOWIN data set [3].

3. ~**60** measurements of fragrances gathered from various online resources.

Data Validation

The preliminary dataset has been manually cleaned, excluding compounds:

- listed as oligomers or mixtures
- having ambiguous structures
- with inconsistent biodegradabilty values
- with duplicates
- causing descriptor calculation errors

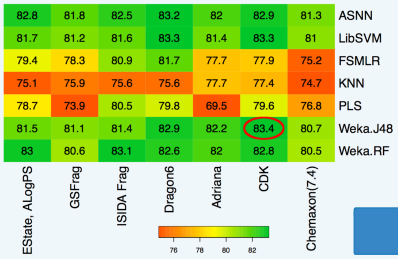
The final dataset consists of **1971** compounds out of which **732** are readily biodegradable and **1239** are not readily biodegradable and is freely available in OCHEM.

Methods

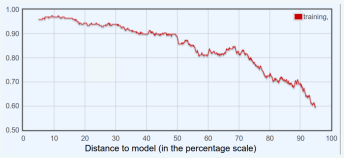
Model Validation: Stratified Bagging Validation has been applied. The optimal bagsize was analysed in the range between 64 and 512 for all initial models. 64 bags emerged to yield the best performing models. It was used for validation for all further models.
Parameter Optimization: For all machine learning methods, parameter optimization was performed on the initial models using stratified bagging validation with 64 bags.
Model Performance: Balanced accuracy is defined as the arithmetic mean of sensitivity and specificity. It was used as a performance measure.

Initial Models

Models were generated for all combinations of machine learning methods and descriptors.
best Ø perf. descriptor: Dragon6 (**83.4%**)
best Ø perf. method: ASNN (**82.4%**)



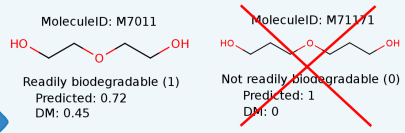
Williams Plot for best model (Weka J48 + CDK) with Bagging-STD as dist. to model.



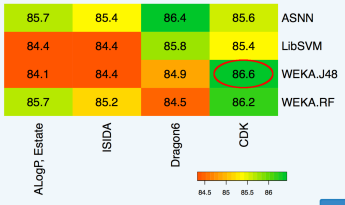
Exclusion Models

The distance to model (Bagging-STD) measure was used to exclude 51 compounds that were incorrectly predicted with high confidence in the previously best models.

Balanced Accuracy of Weka-J48 + CDK model:
initial model (51 outliers left out): 85.7 %
exclusion model (1920 cpds.): 86.6 %



best Ø perf. descriptor: CDK (**86%**)
best Ø perf. method: ASNN (**85.8%**)

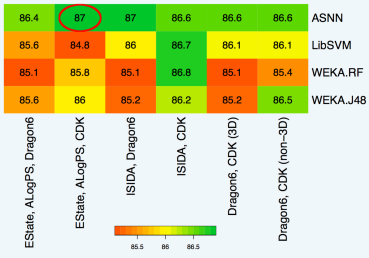


Subsets and Combinations

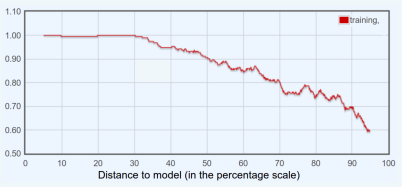
Subsets (3D and non-3D) of the Dragon6 and CDK descriptor sets were analysed, with non-3D descriptors yielding better performing models.

	3D	non-3D
Dragon6	84.2%	86.2%
CDK	82.3%	85.6%

The previously determined best working descriptor sets have been combined:



Williams plot for best model (ASNN + EState, AlogPs, CDK) with Bagging-STD as distance to model.



Regression model

For less than half of the dataset (767 compounds) concrete BOD values were available. From initial regression models, 34 outliers were excluded (p-values < 0.001).

Method	Descriptor	Q^2
ASNN	ISIDA,Dragon	0.66
ASNN	non-3D Dragon6	0.66
ASNN	Dragon6	0.65

Outlook

Ongoing work is concerned with analysis of the relationship between characteristic compound properties and biodegradation. Special focus is set on outlier analysis and identification of primarily responsible descriptors.

References

[1] Cheng, Feixiong et al. In Silico Assessment of Chemical Biodegradability In *Journal of Chemical Information and Modeling,52:3,655-669, 2012*

[2] Sushko, Iurii et al. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. In *Journal of Computer-Aided Molecular Design, 25:6, 533 - 554, 2011*

[3] Tunkel, J. et al. Predicting Ready Biodegradability in the Japanese Ministry of International Trade and Industry In *Test. Environ. Toxicol. Chem., 19, 2478-2485, 2000*