Marie Curie Initial Training Network
Environmental Chemoinformatics (ECO)

Project report
13 August 2012

Read across approach for assessment of toxicological properties and substance persistence

Duration of Short Term fellowship:
October 2011 – 12th August 2012

Early stage researcher:
Tine Ringsted

Project supervisor:
Prof. Roberto Todeschini

Research Institution:
University of Milano-Bicocca

1. Introduction

Compounds which need a long time to degrade are said to be persistent. Persistent compounds are problematic because they can accumulate in the environment and possibly in organisms. They have the potential of spreading by water and air to different parts of the world where they can act on many species and ecosystems. Since persistent compounds will stay for a longer period in our surroundings it is possible that their damage will not be shown immediately but can immerse after a longer period of time.[1]

The European chemical regulation (REACH) requires information on the PBT (Persistent, Bioaccumulating and Toxic) properties of substances to identify the harm that can be posed to the environment or human health.[2] A substance is regarded as persistent in the REACH regulation when any of the following half-life's are exceeded:

- The degradation half-life in marine water is higher than 60 days;
- The degradation half-life in fresh or estuarine water is higher than 40 days;
- The degradation half-life in marine sediment is higher than 180 days;
- The degradation half-life in fresh or estuarine water sediment is higher than 120 days;
- The degradation half-life in soil is higher than 120 days.

Besides chemical half-life's, other types of data can be used within REACH as indicators of persistency such as ready biodegradation which is a screening test for the assessment of biodegradability. In fact, REACH requires that all organic molecules produced or imported in more than one ton per year needs information on ready biodegradation. The PBT assessment should be performed in a weight of evidence approach considering all available information including alternatives to animal testing as for example *in-vitro* and *in-silico* results.[3] Quantitative structure-activity relationship (QSAR) models are mentioned in REACH as a possible *in-silico* method which can be used in the assessment of chemicals (Annex XI in REACH).[2] In order to use results from QSAR models within REACH, a model should comply with four conditions:

(I) Results are derived from a (Q)SAR model whose scientific validity has been established;
(II) The substance falls within the applicability domain of the (Q)SAR model;
(III) Results are adequate for the purpose of classification and labeling and/or risk assessment;
(IV) Adequate and reliable documentation of the applied method is provided.

In order to fulfill condition (I), the following five OECD principles of QSAR validation should be followed[4]:

(1) A defined endpoint;
(2) An unambiguous algorithm;
(3) A defined domain of applicability;
(4) Appropriate measures of goodness of fit, robustness and predictivity;
(5) A mechanistic interpretation if possible.

The focus of this research project was mainly on ready biodegradability. The final aim was to build a predictive QSAR model for ready biodegradation. In the first part of the project, a review of existing models on biodegradability was carried out in order to evaluate the classification performances of existing models and the type of descriptors used to model biodegradability. It was important that the experimental data used in QSAR modeling was in compliance with REACH and the data was therefore screened prior to modeling. The QSAR models were developed using different molecular descriptors and fingerprints together with the modeling method *k*-nearest neighbors (*k*-NN). It was investigated if the selection of few important molecular descriptors could improve the model statistics compared to models built with no selection of molecular descriptors. In the last part of the project, QSAR models were interpreted with knowledge from the literature on biodegradation.

2. State of the art

Several QSAR models have been developed in order to predict biodegradation (see Table 1). Different types of data have been used in modeling (e.g. biodegradation half-life, expert judgment and biodegradation screening tests). Both fingerprints (i.e., vectorial descriptors with ones and zeros stating the presence and absence of selected structural features, respectively) and molecular descriptors have been used in modeling.

Some structural features have been seen to increase the degradation time (e.g. halogens, chain branching, nitro groups, polycyclic residues, heterocyclic residues and aliphatic ether bonds).[1] Other structural features have been found to decrease the time for biodegradation including esters, amides, hydroxyl groups, aldehyde groups, carboxylic acid groups, unbranched linear alkyne chains and phenyl rings.[1] A number of physical-chemical properties have also been found to correlate with biodegradation. Water soluble molecules tend to be more easily biodegradable compared to non-soluble molecules.[5] Also molecular weight has been found to be connected with biodegradability because molecules with a molecular weight above 500 cannot be transported across a bacterial cell membrane.[6]

Table 1: QSAR models on ready biodegradation with the non-error rate (NER) found in the literature.

1st Author	Year	Endpoint	Method[a]	Training set[b]	External test set[b]	Fitting[a,b] (NER%)	Cross-validation[a,b] (NER%)	Test set validation[a,b] (NER%)
Geating[7]	1981	Biodegradation in water	S-DA, RR	349		88.5		
Niemi[8]	1987	5-days BOD	PCA, DA	287		92.0		
Boethling[9]	1989	Biodegradation in water	MLR	46	23 (a) 17 (b)	80.4 (A) 89.1 (B)		96.2 (A+a) 82.0 (B+b)
Howard[10]	1991	Aerobic biodegradation	MLR	235		89.0 - 94.0		
Howard[11]	1992	Aerobic biodegradation	MLR, LR	264	27	91.0 (MLR) 90.0 (LR)		82.0 (MLR) 89.0 (LR)
Klopman[5]	1993	Aerobic biodegradation	MLR	283 (I) 153 (II)	27		74.0 (I)	74.0 (I) 86.0 (I+II)
Boethling[12]	1994	Aerobic biodegradation	MLR, LR	295		89.5 (MLR) 93.2 (LR)		
Boethling[12]	1994	Ultimate and primary biodegradation	MLR	200		82.5 (pri. bio) 83.5 (ult. bio)		
Gamberger[13]	1996	Aerobic biodegradation	Expert rules	45 (I) 146 (II)	40			87.5 (I) 97.5 (II)
Loonen[14]	1999	Ready biodegradation	PLS-DA	894			81.0 – 84.0	
Tunkel[15]	2000	Ready biodegradation	MLR, LR	589	295	82.0 (MLR) 82.7 (LR)		81.4 (MLR) 80.7 (LG)
Huuskonen[16]	2001	Ultimate and primary biodegradation	MLR, ANN	172	12 57			84.0 (MLR) 86.0 (ANN)
Jaworska[17]	2002	Ready biodegradation	CATABOL	532			(Q^2 = 0.88)	
Alikhanidi[18]	2003	Biodegradation half-life	DT	315	105	86.3		78.1
Sakuratatani[19]	2005	Ready biodegradation	CATABOL	743	338 (a) 1123 (b)	83		80.0 (a) 81.0 (b)
Sedykh[20]	2007	Ready biodegradation	CG	1190			(r^2 = 0.69)	
Cheng[21]	2012	Ready biodegradation	SVM, *k*-NN, DT, NB	1440	164	100 (*k*-NN) 87.2 (SVM)		84.2 (*k*-NN) 81.1 (SVM)

[a]: S-DA: Stepwise discriminant analysis, DA: Discriminant analysis, RR: Ridge regression, PCA: Principal component analysis, MLR: Multiple linear regression, LR: Logistic regression, Expert rules: Two rules were generated with 3 statements in each, PLS-DA: Partial least squares-Discriminant analysis, ANN: Artificial neural networks, CATABOL: The CATABOL model used biodegration or degration transformations, DT: Decision tree CG: Conjugate gradient method by minimizing the standard sum of squared errors, SVM: support vector machines, *k*-NN: *k*-nearest neighbors, NB: Naive Bayes.

[b]: I: Training set 1, II: Training set 2, a: External test set 1, b: External test set 2, A: model 1, B: model 2, pri. bio: primary biodegradation, ult. bio: ultimate biodegradation.

3. Materials and methods

3.1. Data

Experimental data on ready biodegradation were collected from the webpage of the National Institute of Technology and Evaluation (NITE) of Japan.[22] The data followed the OECD test guideline (301 C) which measures the Biochemical Oxygen Demand (*BOD*) in aerobic aqueous medium for 28 days[23,24] *BOD* is calculated as shown in Eq. 1 and 2 where *ThOD* is the theoretical oxygen demand which is the total amount of oxygen required to oxidize a chemical completely; it is calculated from the molecular formula.

$$BOD = \frac{\text{mg } O_2 \text{ uptake by test substance} - \text{mg } O_2 \text{ uptake by blank}}{\text{mg test substance in vessel}} \quad (1)$$

$$\%\ BOD = \frac{BOD\ (\text{mg } O_2 \text{ oxygen/mg test substance})}{ThOD\ (\text{mg } O_2 \text{ oxygen/mg test substance})} \text{ x } 100 \quad (2)$$

Chemicals with a *BOD* value higher than 60% are considered as ready biodegradable and molecules with a *BOD* less than 60% are regarded as not ready biodegradable.

The experimental *BOD* values and classification judgments were collected for 1309 molecules. The replicated *BOD* values were given for 223 molecules. The test period was 28 days for 882 molecules and for the rest of the molecules the test period was between 14 and 25 days. A screening procedure was performed to make sure that the experimental data was in compliance with the OECD test protocol for ready biodegradation. The screening procedure removed 247 molecules (section 4.1. describes the procedure). In the end, 1062 molecules remained to be used in modeling.

3.2. Molecular descriptors and fingerprints

The molecular descriptors shown in Table 2 were calculated from the software Dragon.[25] Descriptors with missing values, constant and near-constant values were not used in modeling.

Fingerprints were produced from SubMAT and PaDEL.[26,27] From PaDEL the following eight fingerprints were calculated, namely CDK fingerprint (FP), CDK extended fingerprint (ExtFP), Estate fingerprint (EstateFP), CDK graph only fingerprint (GraphFP), MACCS keys (MACCS), PubChem fingerprint (PubChemFP), Substructure fingerprint (SubFP) and Klekota-Roth fingerprint (KRFP). Fingerprints with a defined set of molecular features are named structural keys. It is also possible to make fingerprints from a set of rules, e.g. all paths with a length of up to 8 atoms. This can generate a high number of paths and a hashing function can therefore be used to fix the size of the fingerprint. Hashed fingerprints

are the name of these fingerprints and the three fingerprints from CDK are hashed fingerprints (see Table 3).

Table 2: Molecular descriptors[25]

Group of descriptors	Number	Description
Constitutional indices	32	Chemical composition of a compound
Ring descriptors	25	Information on rings
Topological indices	34	Numbers from a graph representation of a molecule
2D matrix-based descriptors	84	Topological indices applying algebraic operators to a graph-theoretical matrix of a molecule
Functional group counts	94	The number of selected functional groups
Atom centered fragments	90	The number of selected atom types and groups
Atom-type E-state indices	80	The number of atom-types encoding information on electron accessibility
2D atom pairs	508	The number of defined pairs of non-hydrogen atoms with a defined connection

Table 3: Vectorial descriptors/fingerprints

Name	Bits/Keys (Number)	Description
SubMAT	1365	Structural keys[28]
CDK (FP)	1024	Path based, hashed fingerprint[29]
CDK extended (ExtFP)	1024	Extends the FP with additional bits describing ring features[27]
CDK graph only (GraphFP)	1024	Specialized version of the FP which does not take bond orders into account[27]
Estate (EstateFP)	79	Structural keys containing E-state fragments[30]
MACCS (MACCS)	166	Structural keys[31]
PubChem (PubChemFP)	881	Structural keys
Substructure (SubFP)	307	Structural keys[27]
Klekota-Roth (KRFP)	4860	Structural keys[32]

3.3. Modeling method

k-NN was used as the modeling method to find the relationship between the experimental values and the chemical structures and properties. Different distance/similarity measures were used to find the nearest neighbors. The global molecular descriptors we used were the Cityblock metric, the Euclidian distance and the Minkowski metric. For the vectorial binary descriptors, the similarity indices used were the Jaccard-Tanimoto and the Consonni-Todeschini CT4 index.[33]

Principal component analysis (PCA) and Multidimensional scaling (MDS) were performed on molecular descriptors and fingerprints, respectively, in order to see how the molecular descriptors and fingerprints spread out the data and to study the degree of separation between ready and not ready biodegradable molecules.

Two variable selection methods were investigated. The first method used the univariate Wilk's Lamdba to find the molecular descriptors which were the most important descriptors

in respect to ready and not ready biodegradable molecules.[34] The Wilk's Lambda value was calculated for each descriptor and the 50 descriptors with the lowest Wilk's Lambda values were used in modeling. In the second variable selection method, Wilk's lambda was calculated for all the molecular descriptors and the 200 descriptors with the lowest Wilk's lambda values were applied a Genetic Algorithm.[35]

3.4. Model validation

The 1062 molecules were randomly split into 20 % test and 80% training set. The balance between ready and not ready biodegradable molecules was retained in the test and training set (see Table 4).

Table 4: Number of molecules in the training and test set.

	Ready biodegradable	Not ready biodegradable	Total
Total data	356	706	1062
Training set	285	565	850
Test set	71	141	212

Cross-validation was performed by dividing the training set into ten groups and iteratively predicting one group from a model built on the remaining nine groups. The model statistics was estimated by the use of sensitivity (ability to correctly predict positives/ready biodegradable), specificity (ability to correctly predict negatives/not ready biodegradable) and non-error rate. The estimates were defined as the sensitivity = (TP/(TP+FN)) and specificity = (TN/(TN+FP)) where TP, FN, TN and FP were true positives, false negatives, true negatives and false positives, respectively. The non-error rate was calculated as the average of sensitivity and specificity.

4. Results and discussion

4.1. Data screening

It was important that the QSAR model was built on accurate chemical structures and that the experimental data followed the OECD test guideline (301 C). The screening procedure was done in the following way:

1) ***Molecular structure***

Simplified molecular-input line-entry system (SMILES) describes the molecular structure of a substance and these were gathered for all the molecules. The OECD QSARtoolbox[36] was used to collect the SMILES strings from the CAS number of each molecule. Only SMILES from high or medium quality information were collected. SMILES can be written in different ways and they were therefore canonicalized. The data set was then searched for duplicates but no duplicates were found.

Sometimes two CAS numbers were given to one experimental result, namely "CAS" and "Biodegradation: CAS Registry No." and in those cases "Biodegradation: CAS Registry No" was used. For some experimental results several names were specified and in those situations "Chemical Name in the Official Bulletin" was used unless "Biodegradation: Name of chemical tested" was present. For 81 molecules no defined SMILES string could be found and those molecules were removed.

2) ***Molecules with more than 20% difference between BOD replicates***

Most molecules with replicated *BOD* values had three values. If one of the three values was deviating from the two other *BOD* values then the deviating value was removed if it was an outlier according to Dixon's Q test with a 90% confidence limit.[37] After the removal of outliers, if a molecule still had a difference between replicates of more than 20% and the replicate values classified the molecule into different categories then the molecule was removed. This was the case for 24 molecules which were removed. For the remaining molecules with replicate values the average *BOD* was used.

3) ***Transforming <28 days test results into 28 days***

The test period was not the same in all the experimental results since the original OECD protocol used a 14 days test period (Figure 1).[15] The results with <28 days test period was transformed into 28 days test results as was done in Sedykh & Klopman (2007).[20] This means that the *BOD* % was multiplied with 1+((28-x)/28) where x was the number of days of the test. This extrapolation could over or under estimate the *BOD* values. However, experimental data was only used if the *BOD* value and the judgment by NITE (step 5) classified in the same way. It was therefore assumed that classification errors due the extrapolation could be neglected.

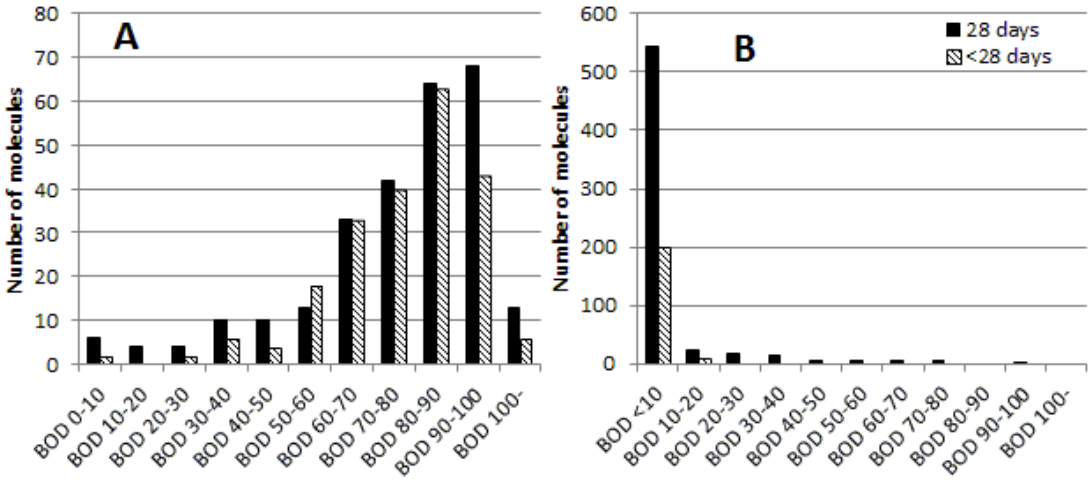

Figure 1: Distribution of *BOD* values for test periods of 28 days and <28 days. (A) Molecules judged by NITE as ready biodegradable. (B) Molecules judged by NITE as not ready biodegradable.

4) Molecules where the classification changed if nitrification was taken into account

If a molecule contains nitrogen then there is a possibility for nitrification in the ready biodegradation test.[23] Nitrification involves the consumption of oxygen and it is therefore necessary to exclude this consumption from the *BOD* value since the *BOD* should only measure the oxygen used by microorganisms. From the collected data it was not possible to know the extent of nitrification. Molecules which differed in their classification depending on the assumption of complete or no nitrification was therefore removed and this was the case for 4 molecules.

5) Molecules where the experimental value did not agree with the classification on NITE

Some molecules had *BOD* values which classified them in one way and a judgment by NITE which classified them in another way. The classification was done according to the *BOD* %. This meant that *BOD* > 60% was classified as ready biodegradable and *BOD* < 60% was classified as not ready biodegradable. The molecules which were classified and judged differently were removed and this was the case for 54 molecules.

6) Disconnected structures

Some compounds contained disconnected structures, e.g. salts, mixtures, isomer mixtures as for instance Cresol and polymers as for example paraformaldehyde. Salts were removed because the ion concentration has an influence on the solubility of a molecule and the solubility is correlated with biodegradation. Mixtures, isomer mixtures and polymers were removed because they contained several structures. In total 84 disconnected molecules were removed.

Table 5: The reasons for removing molecules from the data set.

Reason for removal of molecules from the data set	Removed molecules (number)
No defined SMILES string	81
Replicate values had more than 20% difference and classified differently	24
The classification would change if nitrification was taken into account	4
The experimental value did not agree with the classification on NITE	54
Disconnected structures	84

Table 5 summarizes the removal of molecules during the data screening. 1062 molecules remained after the data screening with 356 ready biodegradable and 706 not ready biodegradable molecules.

4.2. Modeling

It was investigated how the fingerprints and molecular descriptors spread out the ready and not ready biodegradable molecules. As seen from Figure 2, the ready biodegradable molecules had a tendency to form a group inside the group of the not ready biodegradable molecules. This means that it might be difficult to separate the two classes in a linear model since they contain some of the same chemical information. A non-linear model like *k*-NN could therefore improve the classification.

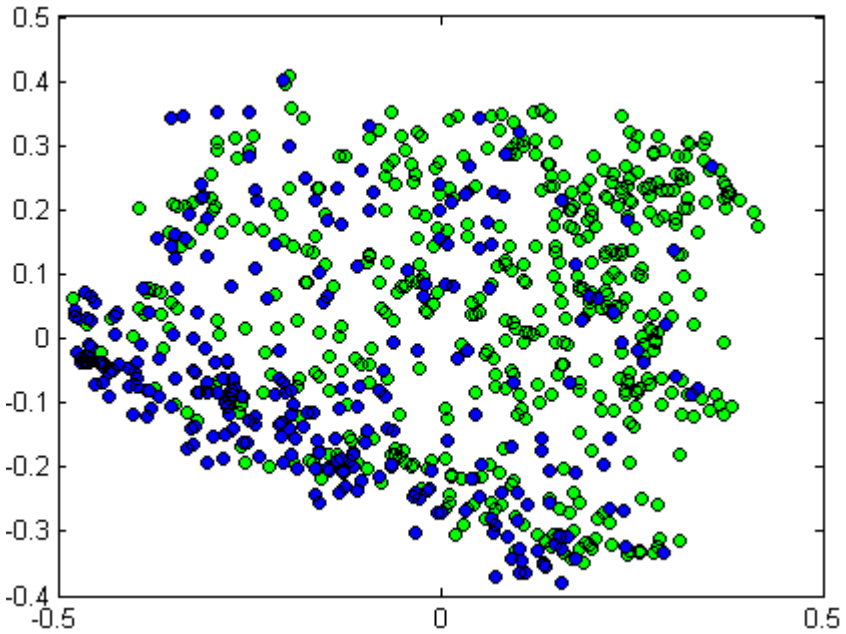

Figure 2: The training set of 850 ready and not ready biodegradable molecules shown as an MDS plot where the distance measure Jaccard-Tanimoto and the MACCS fingerprint was used. (Green) not ready biodegradable, (blue) ready biodegradable.

k-NN models were developed using different fingerprints and groups of molecular descriptors. From the fingerprints, SubMAT and MACCS produced the highest non-error rate of 84% in cross-validation (Table 6). The highest non-error rate for the groups of molecular descriptors from DRAGON was found to be 83% using functional group counts or atom-type E-state indices (Table 7).

Table 6: Statistical results for *k*-NN models with the use of different vectorial descriptors/fingerprints

Fingerprint.[a]	*k*[b]	Dis./sim.[c]	Cross-validation[d]			Test set validation[d]		
			NER (%)	Spe. (%)	Sen. (%)	NER (%)	Spe. (%)	Sen. (%)
SubMAT	5	JT	84	89	79	76	83	69
FP	3	CT	80	84	76	73	82	65
ExtFP	3	CT	79	85	74	75	83	66
GraphFP	10	JT	76	79	74	70	78	62
EstateFP	6	JT	80	78	83	76	74	77
MACCS	8	JT	84	91	78	83	89	77
PubChemFP	6	JT	80	83	76	75	83	68
SubFP	3	CT	80	79	82	76	79	73
KRFP	8	CT	82	86	78	80	85	75

[a]: KRFP: Klekota-Roth fingerprint, SubFP: Substructure fingerprint, EstateFP: E-state fingerprint, PubChemFP: PubChem fingerprint, FP: CDK fingerprint, ExtFP: CDK extended fingerprint, GraphFP: CDK graph only fingerprint.

[b]: *k*: the number of *k*-nearest neighbors.

[c]: Dis./sim.: Distance or similarity measurement method, JT: Jaccard Tanimoto, CT: Consonni-Todeschini.

[d]: NER: Non-error rate, Sen.: Sensitivity (correctly predicted ready biodegradable), Spe.: Specificity (correctly predicted not ready biodegradable).

Wilk's Lambda was used to select 50 variables among the DRAGON descriptors from Functional group counts, Atom centered fragments, Atom-type E-state indices and 2D atom pairs. This approach did not improve the non-error rate compared to the cross-validation results from the different groups of descriptors (Table 7).

Wilk's lambda was used together with a genetic algorithm as a different method to select the most important molecular descriptors. This method was performed on the same data set but the SMILES strings were collected in a slightly different way. The SMILES strings were gathered by a colleague (Kamel Mansouri) from two other databases, namely Chemspider and Cactus.[39,40] The SMILES strings used in the beginning of the result section are almost the same as the ones collected by Kamel Mansouri since only 36 SMILES were differing. From the 36 SMILES strings, 13 molecules were different because they were represented as a salt and a non-salt, 8 molecules were tautomers and one molecule was represented as neutral and charged.

The Wilk's Lambda used together with the genetic algorithm resulted in *k*-NN models with higher non-error rates (85% and 83% in cross-validation) compared to the other *k*-NN models made from single blocks of DRAGON descriptors (Table 7).

Table 7: Statistical results for eight *k*-NN models, one PLS-DA model and one SVM model with the use of different DRAGON molecular descriptors

Descriptors[a]	*k*[b]	Dis./sim.[c]	Cross-validation[d]			Test set validation[d]		
			NER (%)	Spe. (%)	Sen. (%)	NER (%)	Spe. (%)	Sen. (%)
Func. group	4	City	83	86	79	78	82	75
Atom cen.	6	Euclid	82	81	83	78	82	75
Estate	5	City	83	79	87	78	79	76
Atom pairs	6	City	80	85	75	77	82	72
50 Wilk's L.	3	Euclid	80	86	73	81	89	73
19 GA	5	Euclid	85	88	82	84	90	78
8 GA	7	Euclid	83	91	76	83	92	74

[a]: Func. group: Functional group counts, Atom cen.: Atom centered fragments, Estate: Atom-type E-state indices, Atom pairs: 2D atom pairs, 50 Wilk's L.: 50 DRAGON descriptors chosen by Wilk's Lambda, 19 GA: 19 DRAGON descriptors selected by Wilk's Lambda and a genetic algorithm. 8 GA: 8 DRAGON descriptors selected by Wilk's Lambda and a genetic algorithm.

[b]: *k*: the number of *k*-nearest neighbors.

[c]: Dis./sim.: Distance or similarity measurement method, City: Cityblock metric, Euclid: Euclidian distance.

[d]: NER: Non-error rate, Sen.: Sensitivity (correctly predicted ready biodegradable), Spe.: Specificity (correctly predicted not ready biodegradable).

From all the model statistics it was seen that there was a tendency for higher specificity compared to sensitivity. This was expected since other models on ready biodegradation have had the same trend.[14,15,21]

4.3 Model interpretation

Within the ECO project, additional QSAR models were developed by another ECO fellow (Kamel Mansouri) on the basis of the data gathered in this project. To be more specific, three different models based on *k*-NN, Partial Least Squares Discriminant Analysis (PLS-DA) and Support Vector Machines (SVM) were calculated and retained as acceptable ones. Afterwards, an analysis on the selected molecular descriptors in each of these models was carried out in order to see if the information in the models were in accordance with current knowledge on biodegradation.

The *k*-NN model had a high non-error rate of 86% in the cross-validation (results not shown). The 12 molecular descriptors selected for the *k*-NN model can be seen in Table 8. The descriptors were used in a PCA model to study how they would spread out the data.

Table 8: 12 molecular descriptors selected for a *k*-NN model built on ready and not ready biodegradation.

Symbol	Description[a]
SpMax_L	Leading eigenvalue from Laplace matrix
J_Dz(e)	Balaban-like index from Barysz matrix weighted by Sanderson electronegativity
nHM	Number of heavy atoms
F01[N-N]	Frequency of N-N at topological distance 1
F04[C-N]	Frequency of C-N at topological distance 4
NssssC	Number of atoms of type ssssC
nCb-	Number of substituted benzene C(sp2)
C%	Percentage of C atoms
nCp	Number of terminal primary C(sp3)
nO	Number of oxygen atoms
F03[C-N]	Frequency of C-N at topological distance 3
SdssC	Sum of dssC E-states

[a]: ssssC: Carbon with four single bonds, dssC: Carbon with one double bond and two single bonds.

From Figure 3A it was seen that most of the ready biodegradable molecules had negative values on PC1. The descriptors which were responsible for higher values on PC1 contained information on substituted benzene and nitrogen which fits with the knowledge that non-biodegradable molecules contain more cyclic groups and nitro groups (see Figure 3B and the descriptors nCb-, F01[N-N], F04[C-N], and F03[C-N]). PC1 also showed information on branching by having high values for quaternary carbon and carbon bound to three terminal atoms and low values for carbon with two single bonds and one double bond (Figure 3B and the descriptors NssssC, nCp and SdssC). It is known from the literature on biodegradation that branching decreases biodegradation and PC1 is in accordance with this. On PC3 there seemed to be a tendency for ready biodegradable molecules to have a smaller variation around the center compared to not ready biodegradable molecules. The descriptor which had the highest value on PC3 was the percentage of carbon and the descriptors which had the lowest values contained information on oxygen and nitrogen (descriptors C%, nO, F01[N-N], F04[C-N], and F03[C-N]). The descriptors with the highest influence on PC3 therefore seemed to describe the percentage of carbon vs. the percentage of oxygen and nitrogen. This might be the reason why a clear separation between the classes was not seen since the percentage of carbon vs. the percentage of oxygen and nitrogen is not directly correlated with biodegradation. PC3 also had a high value for heavy atoms which could be the reason why not so many biodegradable molecules had high values (descriptor nHM).

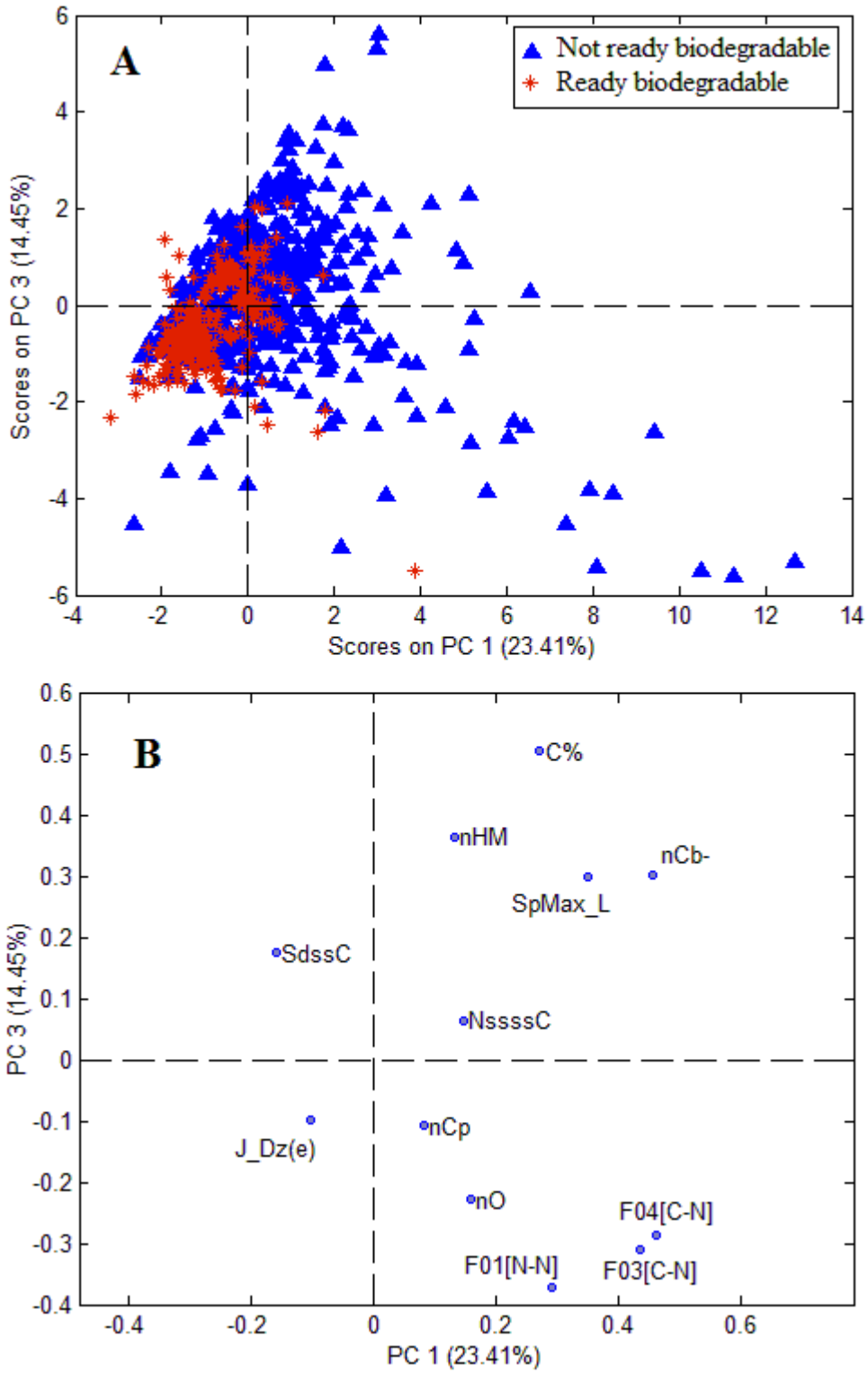

Figure 3: PCA model on ready and not ready biodegradable molecules built on 12 descriptors. The description of the descriptors can be seen in Table 8. (A) PC1 and PC3 showing the score values of ready and not ready biodegradable molecules. (B) PC1 and PC3 with the 12 molecular descriptors.

The statistical performance of the PLS-DA model was a high non-error rate of 86% in the cross-validation (results not shown). In the PLS-DA model the 23 molecular descriptors in Table 9 were used.

Table 9: 23 molecular descriptors selected for a PLS-DA model built on ready and not ready biodegradation.

Symbol	Description[a]
SpMax_L	Leading eigenvalue from Laplace matrix
HyWi_B(m)	Hyper-Wiener-like index (log function) from Burden matrix weighted by mass
LOC	Lopping centric index
nO	Number of Oxygen atoms
SM6_L	Spectral moment of order 6 from Laplace matrix
F03[C-O]	Frequency of C - O at topological distance 3
Me	Mean atomic Sanderson electronegativity (scaled on Carbon atom)
Mi	Mean first ionization potential (scaled on Carbon atom)
nN-N	Number of N hydrazines
nArNO2	Number of nitro groups (aromatic)
nCRX3	Number of CRX3
SpPosA_B(p)	Normalized spectral positive sum from Burden matrix weighted by polarizability
nCIR	Number of circuits
B01[C-Br]	Presence/absence of C - Br at topological distance 1
B03[C-Cl]	Presence/absence of C - Cl at topological distance 3
F04[C-N]	Frequency of C - N at topological distance 4
N-073	Ar2NH / Ar3N / Ar2N-Al / R..N..R
SpMax_A	Leading eigenvalue from adjacency matrix (Lovasz-Pelikan index)
Psi_i_1d	Intrinsic state pseudoconnectivity index - type 1d
B04[C-Br]	Presence/absence of C - Br at topological distance 4
C%	percentage of C atoms
SdO	Sum of dO E-states
TI2_L	Second Mohar index from Laplace matrix

[a]: CRX3: Carbon bound to three halogens, Al: Aliphatic, Ar: Aromatic, ..: Represents aromatic single bonds, R: Represents any group linked through carbon, dO: double bond to oxygen.

Latent variable (LV) 1 in the PLS-DA model had a tendency for lower values for not ready biodegradable molecules compared to ready biodegradable molecules (see Figure 4A). Descriptors with high values on LV1 had information on cycles, halogens and nitrogens which is in alignment with the knowledge on biodegradation since non-biodegradable molecules in general contain these entities (see Figure 4B and the descriptors nCIR, B03[C-Cl], F04[C-N], B04[C-Br], B01[C-Br], N-073, nCRX3). LV2 had high values for oxygen containing compounds and low values for nitrogen and halogen containing compounds. This fits with the literature on biodegradation since biodegradable compounds contain more esters, amides, hydroxyl groups, aldehyde groups and carboxylic acid groups compared to non-biodegradable molecules which on the other hand tend to have more nitrogen and halogens.

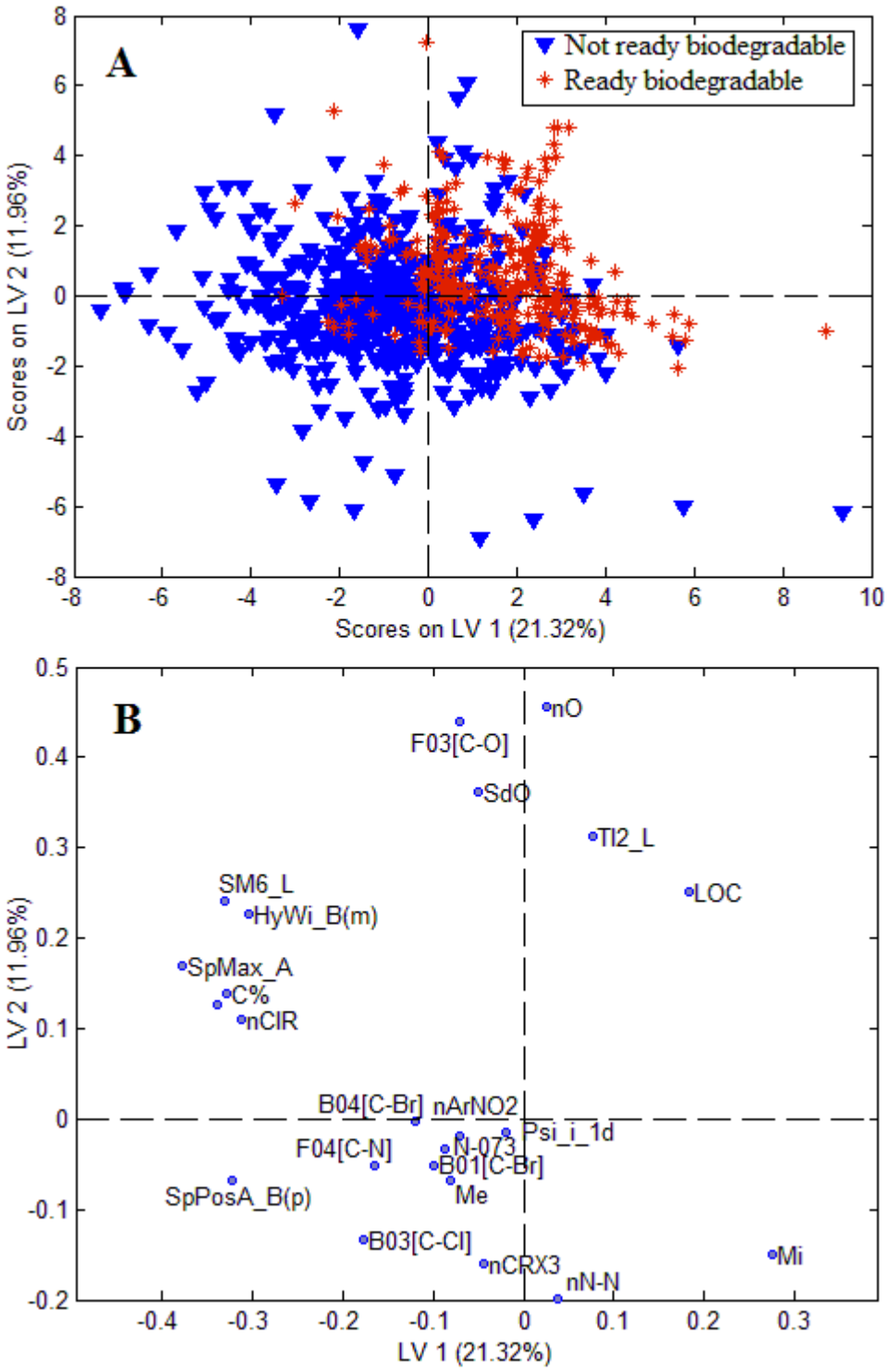

Figure 4: PLS-DA model on ready and not ready biodegradable molecules built on 23 descriptors. The description of the descriptors can be seen in Table 9. (A) LV1 and LV2 showing the score values of ready and not ready biodegradable molecules. (B) LV1 and LV2 with the 23 molecular descriptors.

A SVM model which used 14 descriptors (Table 10) was developed and it had a non-error rate of 86% in the cross-validation (results not shown). The same descriptors were used in a PCA model and the score plot can be seen in Figure 5A. On PC1 not ready biodegradable molecules had higher values compared to ready biodegradable molecules. The descriptors which resulted in higher values on PC1 had information on aromatic groups with electronegative atoms, halogens, nitrogens and quaternary carbon (see Figure 5B and the descriptors C-026, nCb-, nX, NssssC, F02[C-N] and nN). This is in accordance with the literature since non-biodegradable molecules in general have more nitrogen groups and aromatic groups with halogens compared to biodegradable molecules. The descriptors which resulted in high values on PC2 had information on nitrogen and the descriptors that caused lower values had information on halogens, quaternary carbon and rings. Since all these features in general are associated with non-biodegradable molecules it was not surprising that ready biodegradable molecules had more centered values on PC2 compared to not ready biodegradable molecules.

Table 10: 14 molecular descriptors selected for a SVM model built on ready and not ready biodegradation.

Symbol	Description[a]
NssssC	Number of atoms of type ssssC
nCb-	Number of substituted benzene C(sp2)
nCrt	Number of ring tertiary C(sp3)
SpMax_L	Leading eigenvalue from Laplace matrix
C-026	R--CX--R
F02[C-N]	Frequency of C - N at topological distance 2
nN-N	Number of N hydrazines
nHDon	Number of donor atoms for H-bonds (N and O)
SpMax_B(m)	Leading eigenvalue from Burden matrix weighted by mass
Psi_i_A	Intrinsic state pseudoconnectivity index - type S average
nN	Number of Nitrogen atoms
SM6_B(m)	Spectral moment of order 6 from Burden matrix weighted by mass
nArCOOR	Number of esters (aromatic)
nX	Number of halogen atoms

[a]: ssssC: Carbon with four single bonds, R--CX--R: An aromatic carbon bound to an electronegative atom (O, N, S, P, Se, halogens).

It was not possible to explain all descriptors in relation to biodegradation. Some descriptors were not easily interpretable (e.g. SpMax_L which is related to molecular branching) and other descriptors did not describe a specific functional group (e.g. percentage of carbon). However, all three models in Figure 3, 4 and 5 contained some interpretable molecular descriptors which on some PC's or LV's followed the current knowledge on biodegradation.

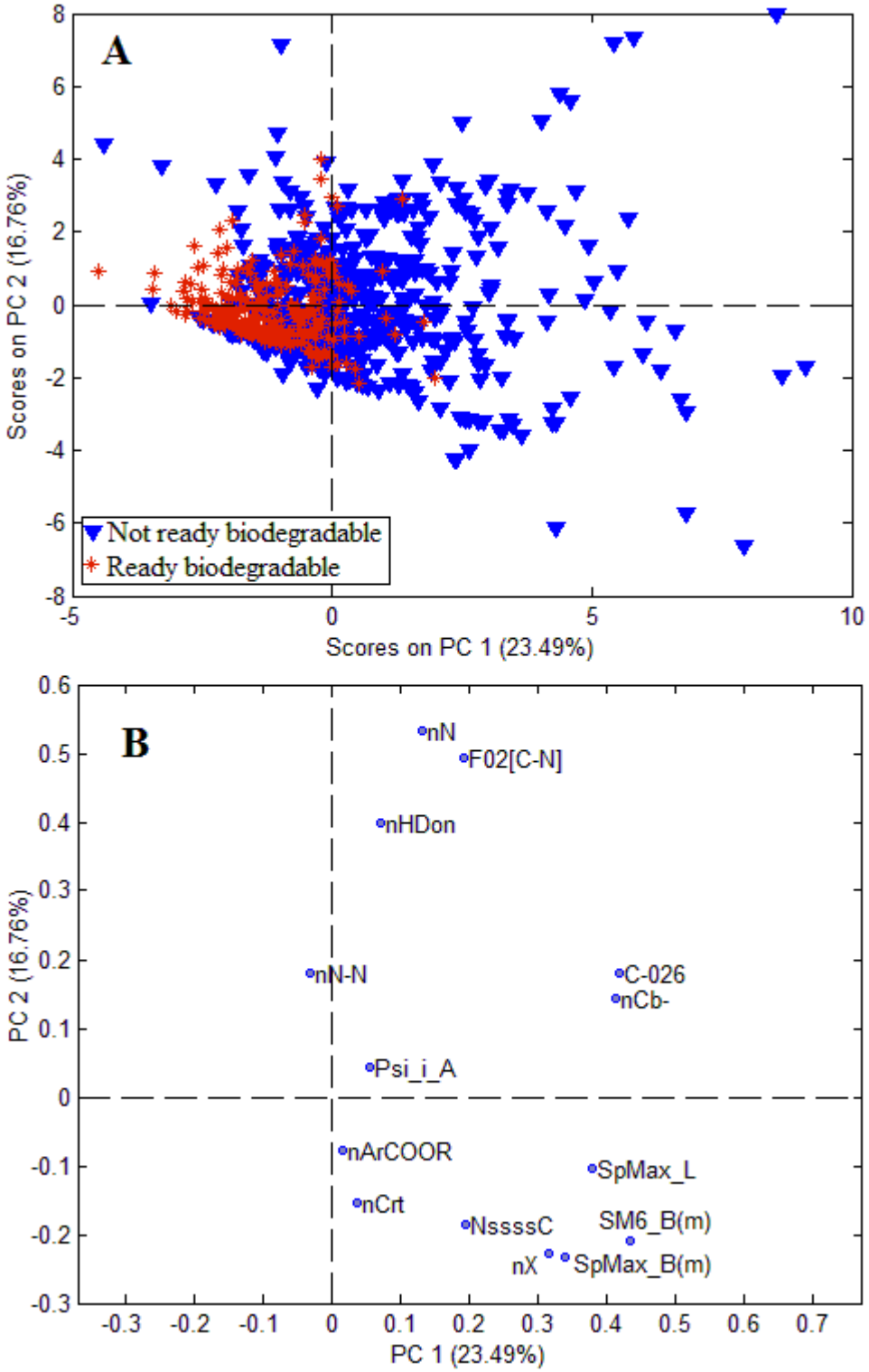

Figure 5: PCA model on ready and not ready biodegradable molecules built on 14 descriptors. The description of the descriptors can be seen in Table 10. (A) PC1 and PC2 showing the score values of ready and not ready biodegradable molecules. (B) PC1 and PC2 with the 14 molecular descriptors.

5. Conclusion

A data set of 1309 ready and not ready biodegradable molecules was collected. The data set was screened in order to obtain correct chemical representations and experimental results which were in accordance with the OECD test guideline (301 C). *k*-NN models were built with different fingerprints and molecular descriptors. The fingerprints SubMAT and MACCS resulted in the highest cross-validation non-error rate of 84% and the molecular descriptors which gave the highest cross-validation result of 83% was functional group counts and atom-type E-state indices. A variable selection technique which used Wilk's Lambda followed by a genetic algorithm resulted in a higher non-error rate in cross-validation of 85%. In the end, it was possible to interpret some of the information used in a *k*-NN, PLS-DA and SVM model built on ready biodegradation.

6. Schools, conferences and training

Web conference: *"Ethoxylates, MS and QSAR"*, University of Milano-Bicocca, 12th October 2011
Presenter: Ian Ken Dimzon

Internal training action: *"Chemoinformatic tools for eco-toxicology"*, University of Milano-Bicocca, 20, 21, 25, 27th October 2011
Presenter: Alberto Manganaro

Web conference: *"Molecular dynamics directed CoMFA studies on carbocyclic neuraminidase inhibitors"*, University of Milano-Bicocca, 2th November 2011
Presenter: Swapnil Chavan

Web conference: *"Prediction of environmental pollutants binding affinities to AhR using MFTA approach"*, University of Milano-Bicocca, 7th December 2011
Presenter: Alexander Safanyaev

Workshop: *"Workshop on High Performance Computing for Proteomics (HPC4P)"*, organized by CINECA, Bologna, 12th December 2011

Web conference: "*Technical aspects of cell-based in vitro assays"*, University of Milano-Bicocca, 18th January 2012
Presenter: Tobias Lammel

External training: *"Scuola di chemiometria",* Department of Chemistry, Pharmaceutical and Nutrition Technologies, University of Genova, 23-26th January 2012

Web conference: *"Computational nanotoxicology",* University of Milano-Bicocca, 15th February 2012
Presenter: Rajesh Rathore

External training: "*2nd Winter School of the Marie Curie Initial Training Network, Chemoinformatics",* Instituto Nacional de Investigacion y Tecnologia Agraria y Alimentaria, Madrid, 27th February – 2nd March 2012

Web conference: *" Species-specific cytotoxicity of nano copper",* University of Milano-Bicocca, 28th March 2012
Presenter: Lan Song

Web conference: *"Pulsed exposure to multiple toxicant: Testing the sequence effect"*, University of Milano-Bicocca, 8th May 2012
Presenter: Isabel O'Connor

External training: *"Use of QSAR in risk assessment: practical use of the VEGA models"*, SETAC Europe, Berlin, Germany, 20th May 2012

External conference: *"SETAC Europe 22nd Annual meeting/6th World meeting"*, SETAC Europe, Berlin, Germany, 21-24th May 2012

External conference: *"2nd summer School 2012of the Marie Curie ITN Environmental Chemoinformatics"*, Milano Chemometrics and QSAR Research Group, Verona, Italy, 11-15th June 2012

External conference: *"15th International workshop on Quantitative Structure-Activity Relationships in Environmental and Health Sciences"*, Tallinn, Estonia, 18-22th June 2012

External training: *"3rd Strasbourg Summer School on Chemoinformatics"*, University of Strasbourg, Strasbourg, France, 25-29th June 2012

External conference: *"Marie Curie Actions Conference 2012"*, The European Commission, Dublin, Ireland, 10-11th July 2012

Web conference: *"Probabilistic risk assessment for the estimation of environmental risk"*, University of Milano-Bicocca, 25th July 2012
Presenter: Pantelis Sopasakis.

7. Publications

Poster at the 2nd Winter School of the Marie Curie Initial Training Network, Chemoinformatics, 2012, Madrid, Spain.
Ringsted, T., Giagloglou, E., Ballabio, D., Mauri, A., Cassotti, M., Consonni, V., Todeschini, R., Read-across methodology in aquatic ecotoxicology and ready biodegradation.

Poster at the 15th International workshop on Quantitative Structure-Activity Relationships in Environmental and Health Sciences, 2012, Tallinn, Estonia.
Ringsted, T., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., Todeschini, R., A (Q)SAR study on ready biodegradability.

Peer reviewed article in preparation for publishing in Journal of Chemical information and modeling.
Ringsted, T., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., Todeschini, R., A (Q)SAR study on ready biodegradability.

8. References

1. Boethling, R.S., 1996. Designing Biodegradable Chemicals. In: DeVito, S.C., Garrett, R.L. (Eds.), Designing Safer Chemicals. American Chemical Society, Washington, DC, USA, Chapter 8.
2. Concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC, REGULATION (EC) No 1907/2006 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL, Official Journal of the European Union, L 136/3, 29.5.2007.
3. Amending Regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) as regards Annex XIII, COMMISSION REGULATION (EU) No 253/2011, Official Journal of the European Union, L 69/7, 16.3.2011.
4. European Chemicals Agency, Guidance on information requirements and chemical safety assessment, Chapter R.6: QSARs and grouping of chemicals, 2008.
5. Klopman, G., Balthasar, D.M., Rosendranz, H.S., Application of the computer-automated structure evaluation (CASE) program to the study of the structure-biodegradation relationships of miscellaneous chemicals, Environmental Toxicology and Chemistry, Vol. 12, 1993, 231–240.
6. Nendza, M., 2004. Prediction of Persistence. In: Cronin, M.T.D., Livingstone, D. (Eds.), Predicting chemical toxicity and fate. CRC Press, Boca Raton, Florida, USA, Chapter 14.
7. Geating, J., Project Summary, Literature Study of the Biodegradability of Chemicals in Water, Vols. 1 and 2, US Environmental Protection Agency, EPA-600/S2-172/176, 1981.
8. Niemi, G.J., Veith, G.D., Regal, R.R., Vaishnav, D.D., Structural features associated with degradable and persistent chemicals, Environmental Toxicology and Chemistry, Vol. 6, 1987, 515–527.
9. Boethling, R.S., Sablijic, A., Screening-level model for aerobic biodegradability based on a survey of expert knowledge, Environmental Science & Technology, Vol. 23, 1989, 672–679.
10. Howard, P.H., Boethling, R.S., Stiteler, W., Meylan, W., Beauman, J., 1991. Development of a predictive model for biodegradability based on BIODEG, the evaluated biodegradation database. In: Hermens, J.L.M., Opperhuizen, A. (Eds.), QSAR in Environmental Toxicology, Vol. IV, Elsevier, New York.
11. Howard, P.H., Boethling, R.S., Stiteler, W.M., Meylan, W.M., Hueber, A.E., Beauman, H.A., Larosche, M.E., Predictive model for aerobic biodegradability developed from a file of evaluated biodegradation data, Environmental Toxicology and Chemistry, Vol. 11, 1992, 593–603.
12. Boethling, R.S., Howard, P.H., Meylan, W., Stiterler, W., Beauman, H., Tirado, N., Group contribution method for predicting probability and rate of aerobic biodegradation, Environmental Science & Technology, Vol. 28, 1994, 459–465.

13. Gamberger D., Hoevaric D., Sekusak S., Sabljic, A., Application of expert's judgements to derive structure-biodegradation relationships, Environmental Science and Pollution Research, Vol. 3, 1996, 224-228.
14. Loonen, H., Lindgren, F., Hansen, B., Karcher, W., Niemelä, J., Hiromatsu, K., Takatsuki, M., Peijnenburg, W., Rorije, E., Struijs, J., Prediction of biodegradability from chemical structure: Modeling of ready biodegradation test data, Environmental Toxicology and Chemistry, Vol. 18, 1999, 1763-1768.
15. Tunkel, J., Howard, P.H., Boethling, R.S., Stiteler, W., Loonen, H., Predicting ready biodegradability in the Japanese ministry of international trade and industry test, Environmental Toxicology and Chemistry, Vol. 19, 2000, 2478-2485.
16. Huuskonen, J. Prediction of biodegradation from the atom-type electrotopolgical state indices. Environmental Toxicology and Chemistry, Vol. 20, 2001, 2152-2157.
17. Jaworska, J., Dimitrov, S., Nikolova, N., Mekenyan, O., Probabilistic assessment of biodegradability based on metabolic pathways: CATABOL System, SAR and QSAR in Environmental Research, Vol. 13, No. 2, 2002, 307-323.
18. Alikhandidi, S., Takahashi, Y., Pesticide Persistence in the Environment – Collected data and structure-Based Analysis, The Journal of Computer Chemistry, Japan, Vol. 3, No. 2, 2004, 59-70.
19. Sakuratani, Y., Yamada, J., Kasai, K., Noguchi, Y., Nishihara, T., External validation of the biodegradability prediction model CATABOL using data sets of existing and new chemicals under the Japanese Chemical Substances Control Law, SAR and QSAR in Environmental Research, Vol. 16, No. 5, 2005, 403-431
20. Sedykh, A. and Klopman, G., Data analysis and alternative modelling of MITI-I aerobic biodegradation, SAR and QSAR in Environmental Research, Vol. 18, No. 7–8, 2007, 693–709.
21. Cheng, F., Ikenaga, Y., Zhou, Y., Yu, Y., Li, W., Shen, J., Du, Z., Chen, L., Xu, C., Liu, G., Lee, P.W., Tang, Y., In silico assesment of chemical biodegradability, Journal of Chemical Information and Modeling, Vol. 52, 2012, 655-669.
22. National Institute of Technology and Evaluation (NITE) of Japan, Chemical Risk Information Platform (CHRIP). Available online at: http://www.safe.nite.go.jp/english/kizon/KIZON_start_hazkizon.html (accessed January 2012).
23. OECD Guideline for testing of chemicals: 301-Ready biodegradability, Organization of Economic Cooperation and Development, Paris, 1992.
24. Tunkel, J., Howard, P.H., Boethling, R.S., Stiteler, W., Loonen, H., Predicting ready biodegradability in the Japanese ministry of international trade and industry test, Environmental Toxicology and Chemistry, Vol. 19, 2478-2485, 2000.
25. Talete srl, DRAGON (Software for Molecular Descriptor Calculations) version 6.0 - 2012 Available online at: http://www.talete.mi.it/
26. Vienna University of Technology, Software SubMat version 3.1. Available online at: http://www.lcm.tuwien.ac.at/w_pr/LCM-sw-SubMat31-info.htm
27. Yap, C.W., PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry, Vol. 32, No. 7, 1466-1474, 2011.

28. Scsibrany, H., Karlovits, M., Demuth, W., Müller, F., Varmuza, K., Clustering and similarity of chemical structures represented by binary substructure descriptors, Chemometrics and Intelligent Laboratory Systems, Vol. 67, 2003, 95-108.
29. Steinbeck, C.; Han, Y.Q.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E.L., The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics, Journal of Chemical Information and Computer Sciences, Vol. 43, No. 2, 2003, 493–500.
30. Hall, L. H., Electropological state indices for atom types: A novel combination of electronic, topological, and valence state information, Journal of chemical information and computer science, Vol. 35, 1995, 1039-1045.
31. Durant,J. L., Leland, B. A., Henry, D. R., Nourse, J. G., Reoptimization of MDL Keys for Use in Drug Discovery, Journal of Chemical Information and Compututer Science,Vol. 42, 2002, 1273-1280.
32. Klekota, J., Roth, F. P., Chemical substructures that enrich for biological activity, Bioinformatics, Vol. 24, No.21, 2008, 2518–2525.
33. Consonni, V., Todeschini, T., New Similarity Coefficients for Binary Data, Communications in mathematical and in computer chemistry, Vol. 68, No. 2, 2012, 581-592.
34. Mardia, K. V., Kent, J. T., Bibby, J. M., 1979. Multivariate Analysis. Academic Press, New York.
35. Leardi, R. and Lupianez, A., Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemometrics and Intelligent Laboratory Systems, Vol. 41, 1998, 195-207.
36. The OECD QSAR Toolbox version 2.2, The Database: Biodegradation OASIS. Available online at: http://www.qsartoolbox.org/download.html (accessed December 2011).
37. Rorabacher, D.B., Statistical treatment for rejection of deviant values: critical values of Dixon's "Q" parameter and related subrange ratios at the 95% confidence level, Analytical chemistry, Vol. 63, No.2, 1991.
38. Sedykh, A. and Klopman, G., Data analysis and alternative modelling of MITI-I aerobic biodegradation, SAR and QSAR in Environmental Research, Vol. 18, No. 7–8, 693–709, 2007.
39. Royal Society of Chemistry, ChemSpider. Available online at: http://www.chemspider.com/ (accessed May 2012).
40. The National Cancer Institute, CADD Group Chemoinformatics Tools and User Services. Available online at: http://cactus.nci.nih.gov/index.html (accessed May 2012).