JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

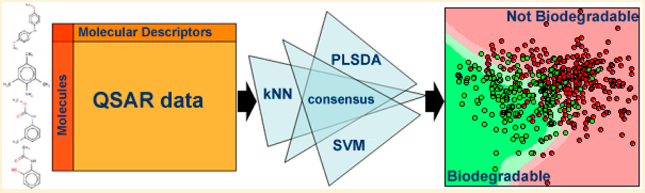
Quantitative Structure−Activity Relationship Models for Ready Biodegradability of Chemicals

Kamel Mansouri, Tine Ringsted, Davide Ballabio,* Roberto Todeschini, and Viviana Consonni

Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano Bicocca, Milano, Italy

Supporting Information

ABSTRACT: The European REACH regulation requires information on ready biodegradation, which is a screening test to assess the biodegradability of chemicals. At the same time REACH encourages the use of alternatives to animal testing which includes predictions from quantitative structure−activity relationship (QSAR) models. The aim of this study was to build QSAR models to predict ready biodegradation of chemicals by using different modeling methods and types of molecular descriptors. Particular attention was given to data screening and validation procedures in order to build predictive models. Experimental values of 1055 chemicals were collected from the webpage of the National Institute of Technology and Evaluation of Japan (NITE): 837 and 218 molecules were used for calibration and testing purposes, respectively. In addition, models were further evaluated using an external validation set consisting of 670 molecules. Classification models were produced in order to discriminate biodegradable and nonbiodegradable chemicals by means of different mathematical methods: *k* nearest neighbors, partial least squares discriminant analysis, and support vector machines, as well as their consensus models. The proposed models and the derived consensus analysis demonstrated good classification performances with respect to already published QSAR models on biodegradation. Relationships between the molecular descriptors selected in each QSAR model and biodegradability were evaluated.



1. INTRODUCTION

Substances which do not decay over a period of time pose a potential threat of accumulation and spread in the environment and organisms. Accumulation of persistent chemicals can, in the long run, show to be harmful because of the continuous exposure and the increasing chemical concentration in the surroundings.[1] The danger is, therefore, that the damages do not have to be immediate but can immerse after a longer period of time.

In Europe, legislators have consequently included persistency in the evaluation of chemicals in the framework of the European REACH regulation. REACH requires that chemicals produced or imported in quantities of more than 1 ton per year need information on ready biodegradation, which is a screening test for the assessment of biodegradability.[2] Thousands of chemicals exist in consumer products, and these can eventually end up in the environment. As an example, the EINECS list comprises more than 100 000 chemicals registered as being on the European Community market between 1971 and 1981.[3] Only a limited number of chemicals from the EINECS list have been tested for their biodegradability. Even for chemicals produced or imported at more than 1000 tons per year, the percentage of those with biodegradation data is merely 61%.[4] To increase the amount of data, REACH encourages the use of a weight-of-evidence approach, which means that all available information should be considered, including predictions from QSARs (quantitative structure−activity relationships) and read-across.

Several QSAR models have in the past been built to predict biodegradation with the use of different types of data, such as chemical half-life, expert judgment, and biodegradation screening tests. Table 1 collects some of the already published QSAR models which were built to classify molecules as ready or not ready biodegradable.[5−21] Both fingerprints (binary matrix stating the presence and absence of fragments/properties) and molecular descriptors have been used for modeling biodegradation.

In the study of Cheng, a consensus model was built where the average of several models were used to classify molecules.[22] The consensus model could correctly predict all ready biodegradable (RB) and not ready biodegradable (NRB) molecules in an external validation set of 27 molecules but the test set contained only four RB molecules. It was not evaluated if the external validation set was a good representation of the chemical space of the model, but four molecules may not be enough to cover the chemical domain.

Several structural features have been found to increase the time for biodegradation (for example halogens, chain branching, nitro groups, polycyclic residues, heterocyclic residues, and aliphatic ether bonds).[1] On the other hand,

Received: January 11, 2013
Published: March 7, 2013

Downloaded via UNIV OF SHEFFIELD on November 29, 2023 at 17:17:56 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

 © 2013 American Chemical Society dx.doi.org/10.1021/ci4000213 | J. Chem. Inf. Model. 2013, 53, 867−878

Table 1. Classification Models on Ready Biodegradation Published in the Literature[a]

ref	method	descriptors	training set (RB/NRB)	test set (RB/NRB)	test set Sn	test set Sp
Loonen et al. 1999[16]	PLSDA	F (127 struct frag)	670 (Na/Na)	224 (Na/Na)	0.80	0.85
Tunkel et al. 2000[17]	MLR	D (43 struct frag + mw)	589 (254/335)	295 (131/164)	0.80	0.82
Tunkel et al. 2000[17]	LR	D (43 struct frag + mw)	589 (254/335)	295 (131/164)	0.79	0.83
Cheng et al. 2012[22]	NB	D (10 physicochemical)	1440 (529/911)	164 (62/102)	0.71	0.91
Cheng et al. 2012[22]	*k*NN	D (12 physicochemical)	1440 (529/911)	164 (62/102)	0.73	0.91
Cheng et al. 2012[22]	SVM	F (79 E-state keys)	1440 (529/911)	164 (62/102)	0.61	0.93

[a]The original reference, modelling method, type and number of descriptors, number of molecules included in training and test sets (as well as number of molecules for ready biodegradable and not ready biodegradable classes), sensitivity, and specificity obtained in the test set are reported for each model. Partial least squares discriminant analysis (PLSDA), multiple linear regression (MLR), logistic regression (LR), naive Bayes (NB), *k* nearest neighbours (*k*NN), support vector machines (SVM), fingerprints (F), molecular descriptors (D), structural fragments (struct frag), molecular weight (mw), ready biodegradable (RB), not ready biodegradable (NRB), not available (Na), sensitivity (Sn, correctly predicted ready biodegradable), specificity (Sp, correctly predicted not ready biodegradable). Consider that Sn and Sp are expressed as ratios, while some of the original papers report them as percentages.

some structural features have been found to enhance biodegradability. These features include esters, amides, hydroxyl groups, aldehyde groups, carboxylic acid groups, unbranched linear alkyne chains and phenyl rings.[1] However, the presence of one of these structural features does not indicate an RB or NRB molecule but should only be taken as generalizations.[1] A physicochemical property which have been found to correlate with the rate of biodegradation is water solubility where soluble molecules tend to be more easily biodegradable compared to insoluble molecules.[23] Molecular weight has also been indicated as an important factor in relation to biodegradation because molecules with a molecular weight higher than 500 cannot be transported into bacterial cells.[24] For some large molecules like proteins and polysaccharides, extracellular enzymes can degrade the molecules into smaller entities which can pass through the cell membrane.[1]

The aim of this study was to build QSAR models to predict ready biodegradation of chemicals by using different modeling methods and types of molecular descriptors. Particular attention was given to data screening and validation procedures in order to carry out predictive models.[25] A set of 837 molecules was used for calibration purposes, while 218 molecules were used to test the calibrated QSAR models. In addition, models were further evaluated using an external validation set consisting of 670 molecules. Data were carefully screened to ensure accurate models based on correct experimental values and molecular structures. The considered classification modeling methods included linear, nonlinear, and local models, as well as consensus models. These methods were coupled with genetic algorithms in order to select the optimal subsets of molecular descriptors. The proposed QSAR models were interpreted in connection to the current knowledge on biodegradation. Finally, since new molecular descriptors were introduced in the proposed models, their link to biodegradability was discussed.

2. MATERIALS AND METHODS

2.1. Biodegradability Experimental Data. Experimental data of the Japanese Ministry of International Trade and Industry (MITI) test (I) were collected from the webpage of the National Institute of Technology and Evaluation (NITE) of Japan.[26] The test is one of the six approved screening tests for ready biodegradation from the Organization for Economic Cooperation and Development (OECD).[27] The MITI test measures the biochemical oxygen demand (BOD) in aerobic aqueous medium for 28 days (the original OECD protocol used a 14 day test period).[17] Chemicals with a BOD value higher than 60% are considered as RB whereas those with a BOD lower than 60% are regarded as NRB.[22,27,28]

The initial data set contained 1309 molecules. BOD values and classification judgments from NITE were given for all the collected molecules. The data set was screened to ensure that it was in accordance with the OECD test protocol (301 C) and that the correct chemical structures were used.

2.2. Data Screening. The screening procedure was carried out on the basis of the steps described in the following paragraphs and summarized in Table 2.

Table 2. Results from the Screening Procedure of the MITI Data

reason for removal of molecules from the data set	number of removed molecules
chemical name and CAS number not in accordance	81
BOD replicates had more than 20% difference and classified differently	24
classification would change if nitrification was taken into account	4
experimental and NITE classification did not agree	54
disconnected structures	91
total number of removed molecules	254

2.2.1. Analysis of the Molecular Structures. The simplified molecular-input line-entry system (SMILES) format was used as the molecular structure representation. SMILES strings were collected from ChemSpider,[29] using a KNIME workflow.[30] When two CAS numbers were assigned to a chemical in the MITI database, then only "Biodegradation: CAS Registry No." was taken into consideration. When several names were specified in the MITI database, "Chemical Name in the Official Bulletin" was considered unless "Biodegradation: Name of chemical tested" was present. Here, 81 chemicals were removed because their CAS Registry Number and chemical names were not consistent in ChemSpider and the MITI database.

2.2.2. Handling of BOD Replicates. Replicate BOD values were given for 223 compounds in the MITI database. Most molecules with BOD replicates had three values. If one of the three values was significantly deviating from the two other BOD values according to Dixon's Q test with a 90% confidence limit,[31] then the deviating value was removed. If a molecule had a difference between BOD replicates higher than 20% and the replicate values classified the molecule into different categories (RB and NRB), then the molecule was removed. This was the

case of 24 molecules. For the remaining molecules with replicate values, the average BOD was used.

2.2.3. Unifying the Test Duration. In the MITI data set, 427 molecules had a test period shorter than 28 days, while the rest had a test period of 28 days. The BOD values based on test periods shorter than 28 days were extrapolated to 28 days as proposed in the literature.[20] This extrapolation could over- or underestimate the BOD values. However, experimental data was only used if the BOD value and the judgment by NITE classified the molecule in the same class. It was therefore assumed that classification errors due to the extrapolation could be neglected.

2.2.4. Handling Molecules with Nitrification. If a molecule contains nitrogen, then there is a possibility for nitrification in the ready biodegradation test.[27] 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. Four molecules which differed in their classification depending on the assumption of complete or no nitrification were removed.

2.2.5. Handling Divergences between Experimental and NITE Classification. As previously described, chemicals with BOD values higher than 60% are classified as biodegradable. When the experimental classification and the NITE judgment did not agree, molecules were removed. This was the case for 54 molecules.

2.2.6. Handling Disconnected Structures. The MITI data set included disconnected structures, such as salts, mixtures, isomer mixtures, and polymers. All 91 disconnected structures were removed because they could affect the subsequent calculation of molecular descriptors.

Table 2 summarizes the screening steps and the corresponding number of removed molecules. In the screening procedure, 254 molecules were removed. The remaining 1055 chemicals, with 356 RB and 699 NRB molecules, were used for modeling. The data set is provided in the Supporting Information file (SI) of this article.

2.3. Molecular Descriptors. The previously collected SMILES codes were used to calculate the molecular descriptors in DRAGON software version 6.[32] A two-dimensional structural representation was selected instead of 3D structures to avoid complex and irreproducible geometrical optimizations. The use of 3D descriptors could add valuable chemical information about the molecules. However, this type of descriptor requires a geometrical optimization and this can be an issue when applying the calibrated models to new molecules since the difference between the 3D conformers can affect the 3D descriptors values.

The calculated molecular descriptors were included in the following blocks of descriptors from DRAGON: constitutional indices, ring descriptors, topological indices, 2D matrix−based descriptors, functional group counts, atom-centered fragments, atom-type E-state indices, and 2D atom pairs (Table 3). A filtering of the descriptors was performed in DRAGON before exporting the descriptor values. Constant, near constant, and correlated descriptors were removed. In the latter case, for each pair of descriptors with a correlation coefficient higher than 98%, the one showing the largest pair correlation with all the other descriptors was excluded. A total number of 781 descriptors were exported from DRAGON for the subsequent modeling analysis.

Table 3. Number of Molecular Descriptors Initially Calculated by Using DRAGON

DRAGON block	number of descriptors
constitutional indices	32
ring descriptors	25
topological indices	34
2D matrix-based descriptors	84
functional group counts	88
atom centered fragments	69
atom-type E-state indices	37
2D atom pairs	412
total number of molecular descriptors	781

2.4. Modeling Methods. Three classification modeling methods were applied in order to find the appropriate relationship between molecular structures, encoded in molecular descriptors, and the biodegradability of chemicals: *k* nearest neighbors (*k*NN), partial least squares discriminant analysis (PLSDA), and support vector machines (SVM). The application of methods based on different mathematical strategies aimed to better explore the chemical space and balance potential biases related to each single modeling algorithm.

The *k*NN classification rule is conceptually quite simple:[33] a molecule is classified according to the classes of the *k* closest molecules, which means, it is classified according to the majority of its *k* nearest neighbors in the descriptors space. In this work, the Euclidean metric was used to measure distances between molecules. The *k* value giving the lowest classification error in cross-validation was selected as the optimal one.

PLSDA is a classification technique that profits the properties of partial least squares regression (PLS2-based method) with the discrimination power of a classification technique.[34,35] It finds fundamental relations between the matrix of descriptors and the class vector by calculating latent variables (LVs), which are orthogonal linear combinations of the original variables. PLSDA models were optimized in cross-validation to find a compromise between the classification performance and the number of selected LVs.

SVM define a decision boundary that optimally separates two classes by maximizing the distance between them.[36,37] The decision boundary can be described as an hyperplane that is expressed in terms of a linear combination of functions parametrized by support vectors, which consist in a subset of training molecules. SVM algorithms search for the support vectors that give the best separating hyperplane using a kernel function. During optimization, SVM search the decision boundary with maximal margin among all possible hyperplanes, where the margin can be intended as the distance between the hyperplane and the closest point for both classes. This procedure was carried out by means of a kernel based on a radial basis function.

Genetic algorithms (GAs) were applied to find the optimal subset of molecular descriptors.[38] GAs start from an initial random population of chromosomes, which are binary vectors representing the presence or absence of molecular descriptors. An evolutionary process is simulated to optimize a defined fitness function and new chromosomes are obtained by coupling the chromosomes of the initial population with genetic operations (crossover and mutation). The used fitness function was the classification error calculated in cross-validation.

dx.doi.org/10.1021/ci4000213 | J. Chem. Inf. Model. 2013, 53, 867−878

Consensus analysis was also applied in order to combine information and predictions obtained by the three different modeling techniques. In fact, the consensus approach can improve the quality of models by increasing their prediction reliability.[39] Consensus modeling has also been shown to diminish the effects of noisy data. Individual models contain varying amounts of noise, which can be reduced by averaging the predictions of several models.[40] The generation of a consensus analysis can be based on different strategies such as averaging, scoring, and probabilities.[39−42] In this work, two different consensus algorithms were adopted: (a) Each molecule was assigned to the most frequent class out of the three predictions obtained with the considered classification methods (*k*NN, PLSDA, and SVM). (b) A molecule was assigned only if the three models classified it in the same class; otherwise, it was not assigned.

2.5. Model Validation. Molecules were randomly divided into training and test sets, containing 80% and 20% of the total number of considered molecules, respectively. The selection was performed maintaining the class proportions, that is, the number of test molecules of each class was proportional to the number of training molecules of that class. The training set was used to select molecular descriptors and to build the classification models. Molecules of the test set were used just to evaluate the predictive ability of the trained models.

During model optimization and descriptor selection, a cross-validation procedure with five cancellation groups was used. Classification models were evaluated on the basis of specificity and sensitivity, which are the ability to correctly predict RB and NRB molecules, respectively. In particular, specificity (Sp) and sensitivity (Sn) were calculated with the following equations:

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \qquad \mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

where, TN and TP are the number of true negatives and true positives, and FN and FP are the number of false negatives and false positives, respectively. Being a two-class model, consider that the sensitivity of one class corresponds to the specificity of the other class. In addition, the nonerror rate (NER) was calculated as the average of specificity and sensitivity, while the classification error rate (ER) was calculated as the complement of NER. These indices were used in order to better estimate classification performances in presence of a data set with unequal number of molecules in each class. In this study error rate, specificity, and sensitivity are expressed as ratios and not as percentages.

The classification models were further evaluated using an external validation set. This set was built merging two sources: 464 molecules of the data set modeled by Cheng et al.[22] and 206 molecules of the Canadian DSL database (Domestic Substances List).[43] Initially, 1604 compounds were collected from the set modeled by Cheng. These molecules were screened in order to remove compounds already present in our training or test sets. Moreover, the screening procedure used for the molecules of the MITI data set (described in section 2.2) was applied on this set of molecules. Some CAS numbers were missing, and thus, they were retrieved from ChemSpider matching molecular SMILES and chemical names. After the screening, 464 compounds were selected and included in the external validation set.

The considered DSL list consisted of more than 3500 compounds which meet the categorization of the Canadian Environmental Protection Act.[44] According to this categorization, compounds which are classified as persistent and bioaccumulative can be considered as not biodegradable. Therefore, the 420 persistent and bioaccumulative compounds of the DSL list were considered for inclusion in the external validation set. The SMILES structures were collected using the KNIME workflow described in section 2.2. After removing inorganic compounds, polymers, salts, and disconnected structures, as well as 21 compounds overlapping with the Cheng data set, 206 NRB molecules were added to the external validation set. Summarizing, the external validation set included a total of 670 compounds. The number of RB and NRB molecules of training, test, and external validation sets are summarized in Table 4.

Table 4. Number of Molecules Included in Training, Test, and External Validation Set

data	ready biodegradable	not ready biodegradable	total
training set	284	553	837
test set	72	146	218
external validation set	191	479	670

2.6. Software. A KNIME workflow[30] was used to collect and check SMILES notations from ChemSpider database.[29] Molecular descriptors were calculated by means of DRAGON.[32] SVM were calibrated using the library LIBSVM 3.1 implemented in C[45] and compiled in MATLAB 7.13.[46] GAs, models fitting, and predictions were performed in MATLAB 7.13[46] by means of routines built by the authors.

3. RESULTS AND DISCUSSION

3.1. Descriptor Selection and Model Calibration. The selection of molecular descriptors was organized into two subsequent steps in order to handle the large number of calculated descriptors (781) and to avoid potential overfitting of the QSAR models. GAs were separately calculated on each block of molecular descriptors (Table 3). Descriptors selected from each block were then merged and again GAs were used to find the most appropriate subset of molecular descriptors to calibrate the final QSAR models. The final models were selected taking into consideration the ER in cross-validation and a balanced ratio between the specificity and the sensitivity. It was also important that the final models used a reduced number of selected descriptors and for the PLSDA model also a low number of latent variables. This was done in order to make the models easy to interpret and at the same time to decrease the risk of overfitting. The same procedure was used by coupling GAs with the three considered modeling methods. The classification model based on *k*NN, PLSDA, and SVM included 12, 23, and 14 molecular descriptors, respectively. The obtained QSAR models were validated using the molecules included in both test and external validation sets, which did not participate in the descriptor selection and model calibration. Values of the selected molecular descriptors are provided in the Supporting Information (SI) of this article.

3.2. Classification Performances of the QSAR Models. The classification performances of the three QSAR models are collected in Table 5. The three classification models showed comparable performances. The ER in fitting and cross-validation was equal to 0.14 for all the computed models, while the error on the test set was equal to 0.14 with SVM and slightly higher (0.15) with both PLSDA and *k*NN. This balance

Table 5. Classification Results in Fitting, Cross-Validation, and on the Test Set of the Proposed QSAR Models and Their Consensus Analysis[a]

			fitting			5-fold cross-validation			test set		
model	desc	*k*/LVs/c	ER	Sn	Sp	ER	Sn	Sp	ER	Sn	Sp
*k*NN	12	6	0.14	0.84	0.89	0.14	0.84	0.88	0.15	0.81	0.90
PLSDA	23	5	0.14	0.88	0.83	0.14	0.88	0.83	0.15	0.83	0.87
SVM	14	5	0.14	0.81	0.92	0.14	0.80	0.91	0.14	0.82	0.91
consensus 1	41		0.11	0.86	0.91	0.11	0.87	0.91	0.13	0.82	0.92
consensus 2	41		0.07	0.91	0.95	0.07	0.91	0.95	0.09	0.88	0.94
			19% not assigned			20% not assigned			15% not assigned		

[a]For each model, the number of included descriptors, error rate (ER), specificity (Sp, correctly predicted not ready biodegradable), and sensitivity (Sn, correctly predicted ready biodegradable) are provided. The optimal parameters, *k* for *k*NN, the number of latent variables (LVs) for PLSDA, and the cost (c) for SVM, are reported in the table.

between model performances on the training and test sets can indicate the absence of overfitting, which is a possibility when dealing with variable selection on high dimensional data. Moreover, the quality of a QSAR classification model should also be evaluated on the basis of its ability to correctly predict each modeled class. Compared to other QSAR models on ready biodegradation,[16,17,22] the proposed models showed a good balance between specificity and sensitivity, which never had a difference higher than 0.11. In addition, specificity and sensitivity values calculated on the training and test sets were comparable, indicating robustness and reliability of the proposed models. The SVM and *k*NN models showed higher specificity than sensitivity for the RB class, that is, errors associated to NRB molecules predicted as biodegradable were lower. PLSDA, on the other hand, had the opposite behavior in the fitting and cross-validation results but the test set result showed the same tendency as the *k*NN and SVM models. The reason for the PLSDA model to have a higher sensitivity than specificity in the fitting and cross-validation results and the opposite scenario for the test set was not known but since the test set was chosen randomly, it is possible that this result is due to random variation.

Afterward, predictions obtained by the three classification QSAR models were merged and models based on the two different consensus approaches previously described were calculated. Classification performances of the consensus models are shown in Table 5. When assigning molecules to the most frequent class (consensus 1), classification results were improved, maintaining a reasonable balance between specificity and sensitivity on both the training and test sets. On the other hand, when not assigning molecules in the presence of divergence between the three classification models (consensus 2), ERs were further decreased. The classification ER on the test molecules was equal to 0.09, with a percentage of not assigned test molecules equal to 15%. This could mean that not assigned molecules were associated with lower reliability of prediction. In any case, the presence of 15% not classified molecules was balanced by the good classification performances of the consensus 2 model. As an example of the improved performance it can be mentioned that consensus 2 gave a specificity on the test set (ratio of correctly predicted NRB test molecules) equal to 0.94. The presence of not classified molecules is a matter of choice between high reliability with less molecules predicted or predictions for all the molecules with a lower reliability.

The external validation set supported the results for the three QSAR and consensus models by giving results relatively close to the cross-validation and test set validation. Results are collected in Table 6. The ERs of *k*NN, SVM, PLSDA, and consensus 1 were in the range between 0.17 and 0.18 and thus

Table 6. Classification Results on the External Validation Set of the Proposed QSAR Models and Their Consensus Analysis[a]

	ER	Sn	Sp
*k*NN	0.17	0.75	0.91
PLSDA	0.17	0.80	0.86
SVM	0.18	0.74	0.91
consensus 1	0.17	0.76	0.91
consensus 2	0.13	0.81	0.94
		13% not assigned	

[a]For each model, error rate (ER), specificity (Sp, correctly predicted not ready biodegradable), and sensitivity (Sn, correctly predicted ready biodegradable) are provided.

comparable with those obtained on the training and test sets (Table 5). The slightly higher ERs were expected due to the different sources of the external validation molecules, which could have slightly different classification thresholds with respect to the MITI data set used to train the models. Consensus 2 gave again the lowest ER on the external validation set (0.13) and 13% of not assigned molecules. Finally, all considered models showed the same conservative behavior on the external validation set, that is, specificity was always higher than sensitivity. Thus, NRB molecules were more accurately predicted and models did not tend to classify them as biodegradable.

Considering all the obtained results in classification (summarized in Tables 5 and 6), as well as the model complexity (represented by the number of selected molecular descriptors), the proposed QSAR models and their consensus models had equal or better classification performances with respect to models already published in the literature (Table 1). In particular, models selected in this study showed balanced results on training, test, and external validation sets, suggesting reliable predictions and the absence of potential overfitting. This is in contrast to some of the models published in the literature which showed greater difference between the calibration and validation results, that is, ER equal to 0.00 and 0.18 on the training and test molecules, respectively.[22]

3.3. Descriptor Interpretation. One of the fundamentals of QSAR is that models should be reduced to a set of descriptors which is information rich but as small as possible, in order to ensure stability of the model and reliability of its predictions.[47,48] The symbols of the descriptors, the descriptor

dx.doi.org/10.1021/ci4000213 | J. Chem. Inf. Model. 2013, 53, 867−878

Table 7. List of Molecular Descriptors Selected in the QSAR Models

symbol	description	DRAGON block	model
B01[C-Br]	presence/absence of C−Br at topological distance 1	2D atom pairs	PLSDA
B03[C-Cl]	presence/absence of C−Cl at topological distance 3	2D atom pairs	PLSDA
B04[C-Br]	presence/absence of C−Br at topological distance 4	2D atom pairs	PLSDA
C%	percentage of C atoms	constitutional indices	*k*NN−PLSDA
C-026	R−CX−R	atom centered fragments	SVM
F01[N-N]	frequency of N−N at topological distance 1	2D atom pairs	*k*NN
F02[C-N]	frequency of C−N at topological distance 2	2D atom pairs	SVM
F03[C-N]	frequency of C−N at topological distance 3	2D atom pairs	*k*NN
F03[C-O]	frequency of C−O at topological distance 3	2D atom pairs	PLSDA
F04[C-N]	frequency of C−N at topological distance 4	2D atom pairs	*k*NN−PLSDA
HyWi_B(m)	hyper-Wiener-like index (log function) from Burden matrix weighted by mass	2D matrix-based	PLSDA
J_Dz(e)	Balaban-like index from Barysz matrix weighted by Sanderson electronegativity	2D matrix-based	*k*NN
LOC	lopping centric index	topological indices	PLSDA
Me	mean atomic Sanderson electronegativity (scaled on Carbon atom)	constitutional indices	PLSDA
Mi	mean first ionization potential (scaled on carbon atom)	constitutional indices	PLSDA
N-073	Ar2NH/Ar3N/Ar2N−Al/R···N···R	atom centered fragments	PLSDA
nArCOOR	number of esters (aromatic)	functional group counts	SVM
nArNO2	number of nitro groups (aromatic)	functional group counts	PLSDA
nCb-	number of substituted benzene C(sp2)	functional group counts	*k*NN−SVM
nCIR	number of circuits	ring descriptors	PLSDA
nCp	number of terminal primary C(sp3)	functional group counts	*k*NN
nCrt	number of ring tertiary C(sp3)	functional group counts	SVM
nCRX3	number of CRX3	functional group counts	PLSDA
nHDon	number of donor atoms for H-bonds (N and O)	functional group counts	SVM
nHM	number of heavy atoms	constitutional indices	*k*NN
nN	number of nitrogen atoms	constitutional indices	SVM
nN-N	number of N hydrazines	functional group counts	PLSDA−SVM
nO	number of oxygen atoms	constitutional indices	*k*NN−PLSDA
NssssC	number of atoms of type ssssC	atom-type E-state indices	*k*NN−SVM
nX	number of halogen atoms	constitutional indices	SVM
Psi_i_1d	intrinsic state pseudoconnectivity index−type 1d	topological indices	PLSDA
Psi_i_A	intrinsic state pseudoconnectivity index—type S average	topological indices	SVM
SdO	sum of dO E-states	atom-type E-state indices	PLSDA
SdssC	sum of dssC E-states	atom-type E-state indices	*k*NN
SM6_B(m)	spectral moment of order 6 from Burden matrix weighted by mass	2D matrix-based	SVM
SM6_L	spectral moment of order 6 from Laplace matrix	2D matrix-based	PLSDA
SpMax_A	leading eigenvalue from adjacency matrix (Lovasz−Pelikan index)	2D matrix-based	PLSDA
SpMax_B(m)	leading eigenvalue from Burden matrix weighted by mass	2D matrix-based	SVM
SpMax_L	leading eigenvalue from Laplace matrix	2D matrix-based	*k*NN−PLSDA−SVM
SpPosA_B(p)	normalized spectral positive sum from Burden matrix weighted by polarizability	2D matrix-based	PLSDA
TI2_L	second Mohar index from Laplace matrix	2D matrix-based	PLSDA

blocks, and a brief description of the molecular descriptors from DRAGON which were selected in this study are collected in Table 7. The numbers of molecular descriptors included in each of the proposed models are comparable to those published in the literature (Table 1). In order to evaluate how the selected descriptors related to ready biodegradability, principal component analysis (PCA) was performed separately on the descriptors selected in the *k*NN and SVM models. PCA models were calculated on the training set, while test set molecules were projected onto the PCA model. Scores and loadings plots were used to discuss the behavior of the selected descriptors in relation to the knowledge on biodegradability found in the literature. The descriptors included in the PLSDA model were directly analyzed by means of the latent variables calculated by PLSDA.

3.3.1. Molecular Descriptors of the kNN Model. The *k*NN model included 12 descriptors (Table 7). The results of the PCA analysis on the set of 12 descriptors are shown in Figure 1. The score plot of the first and fourth principal components (PC1 and PC4) explained together 34% of the variance and is shown in Figure 1A, while the corresponding loading plot is presented in Figure 1B.

The combination of PC1 and PC4 gave a reasonable separation between RB and NRB molecules. The majority of NRB molecules had positive scores on PC1 (Figure 1A). As shown in the loading plot (Figure 1B), the descriptors which were responsible for the highest positive values on PC1 were nCb-, F01[N-N], F04[C-N], and F03[C-N]. These descriptors encode information on substituted benzene and nitrogen. This fits with the fact that NRB molecules contain more cyclic and nitro groups than RB molecules.[1] The descriptor "number of heavy atoms" (nHM) is also located in the positive side of PC1, and this can be related to the fact that RB molecules do not contain heavy atoms. On the other hand, most of the RB molecules had a low value on PC1. One of the descriptors which were correlated with low values on PC1 was the

dx.doi.org/10.1021/ci4000213 | J. Chem. Inf. Model. 2013, 53, 867−878

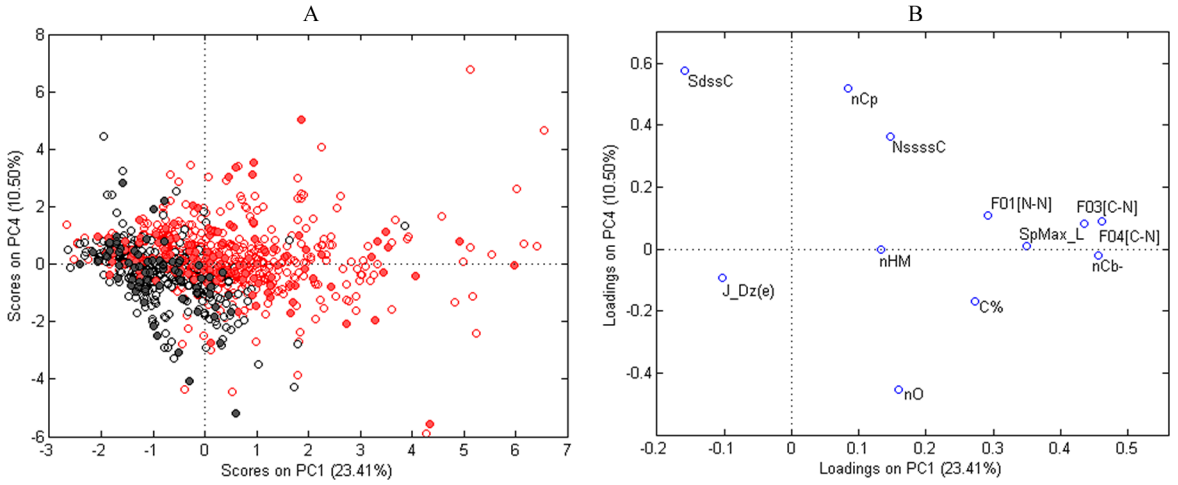


Figure 1. PCA of the descriptors used in the *k*NN model. Scores plot (A) and loadings plot (B) of the first and fourth principal components (explained variance equal to 34%). Ready biodegradable molecules are colored in black, and not ready biodegradables are in red; training molecules are marked with empty circles, and test molecules are marked with full circles. Labels of the molecular descriptors refer to symbols listed in Table 7.

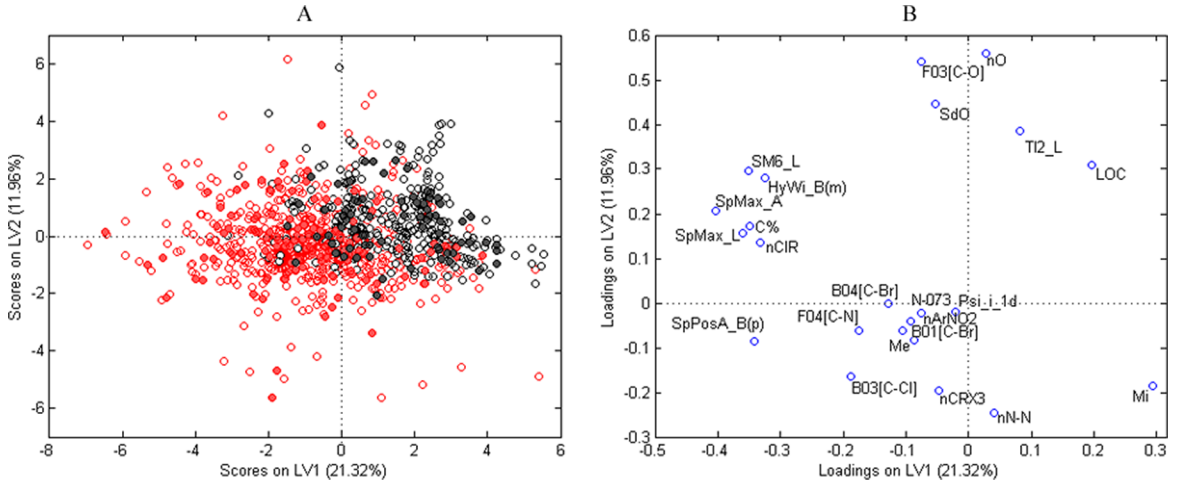


Figure 2. Scores plot (A) and loadings plot (B) of the first and second latent variables of the PLSDA model (explained variance equal to 33%). Ready biodegradable molecules are colored in black, and not ready biodegradables are in red; training molecules are marked with empty circles, test molecules are marked with full circles. Labels of the molecular descriptors refer to symbols listed in Table 7.

descriptor giving information about carbon with two single bonds and one double bond (SdssC). SdssC might be correlated with RB molecules because according to the literature, this class of molecules tend to be less branched compared to the NRB ones.[1] Also PC4 contained information on molecular branching, since two of the most important descriptors on this component were quaternary carbon and carbon bound to three terminal atoms (NssssC, nCp). Having the same upper right side orientation as the NRB, these two descriptors are therefore negatively correlated with biodegradability, thus confirming that branching decreases biodegradation.

3.3.2. Molecular Descriptors of the PLSDA Model. The PLSDA model included 23 descriptors (Table 7). The score plot of the first and second latent variables (LV1 and LV2), explaining together 33% of variance, is shown in Figure 2A.

RB molecules were grouped in the upper right side of the score plot, having positive scores on both LV1 and LV2. Matrix-based descriptors were placed on the extreme left side of the loadings plot (Figure 2B), thus correlating with NRB. These descriptors contain information on the molecular branching, and they might therefore be connected with NRB molecules since the degree of branching has an influence on a molecule's ability to biodegrade.[1] Descriptors with information on cycles, halogens, and nitrogen (nCIR, B03[C-Cl], F04[C-N], B04[C-Br], B01[C-Br], N-073, and nCRX3) had negative loadings on LV1. Cycles, halogens, and nitrogen are more often seen in NRB compared to RB compounds, and their connection with NRB molecules is therefore in alignment with knowledge from the literature.[1]

Descriptors related to the presence of oxygen (nO, F03[C-O], and SdO) had positive loadings on LV2 and, thus, were responsible for the separation of the RB and NRB classes on this latent variable. On the other side, descriptors related to the presence of nitrogen and halogens (B03[C-Cl], nCRX3, nNN) had negative loadings on LV2. This result fits with the knowledge on biodegradation, since the presence of functional groups with oxygen atoms increase biodegradability, while NRB molecules tend to have more nitrogen and halogens.[1]

3.3.3. Molecular Descriptors of the SVM Model. The SVM model included 14 molecular descriptors (Table 7). The results of the PCA analysis on this set of 14 descriptors are shown in

dx.doi.org/10.1021/ci4000213 | J. Chem. Inf. Model. 2013, 53, 867−878

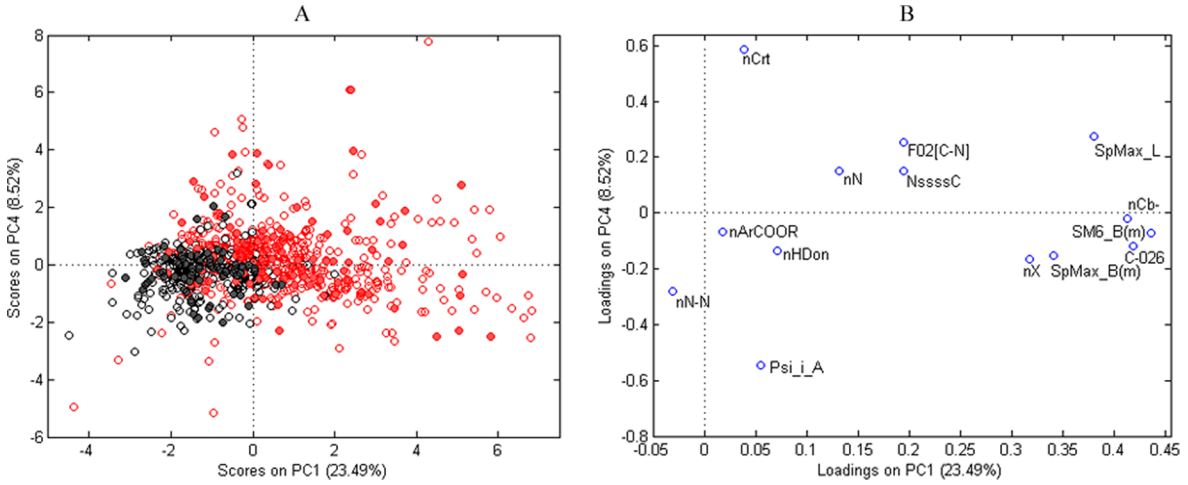


Figure 3. PCA of the descriptors used in the SVM model. Scores plot (A) and loadings plot (B) of the first and fourth principal components (explained variance equal to 32%). Ready biodegradable molecules are colored in black, and not ready biodegradables are in red; training molecules are marked with empty circles, and test molecules are marked with full circles. Labels of the molecular descriptors refer to symbols listed in Table 7.

Figure 3. The score plot of the first and fourth principal components (PC1 and PC4), explaining together 32% of variance, is shown in Figure 3A.

The first principal component was able to separate the two classes in the presence of some overlap (Figure 3A). In particular, RB molecules were characterized by negative scores, while the majority of NRB molecules were placed on the right side of PC1, having positive scores. The majority of the descriptors had positive loadings on PC1, as shown in Figure 3B. The most important descriptors for PC1 had information on molecular branching, aromatic groups and halogens (SpMax_L, SM6_B(m), C-026, nCb-, nX), as well as the presence of nitrogen (NssssC, F02[C-N], nN). All these descriptors are related to the NRB class of molecules, which is characterized by positive scores on PC1. This is in accordance with the literature since NRB molecules in general have more nitrogen groups and aromatic groups with halogens compared to RB molecules.[1] PC4 is less successful in separating the two classes. However, the nCrt descriptor, which encodes information about rings, had the greatest positive loading value on PC4. PC4 seems to show a tendency for lower values among the RB molecules compared to the second class, which is expected, since rings are more often seen in NRB.

Summarizing, descriptors selected in each QSAR model encoded similar information about the presence of halogens, chain branching, nitro groups, rings and some functional groups, which are related to biodegradability. Moreover, it was seen that the information on the relationships between chemical structures and biodegradability was consistent in the three QSAR models even though the descriptors were selected independently in each of the three proposed models.

3.4. Improving Interpretability of the Models. The proposed models for ready biodegradability are based on several substructure descriptors but also on some matrix-based molecular descriptors (HyWi_B(m), J_Dz(e), SM6_B(m), SM6_L, SpMax_A, SpMax_B(m), SpMax_L, SpPosA_B(p), TI2_L). These descriptors are calculated from Laplacian (**L**), Barysz (**Dz**), and Burden (**B**) matrices, which are derived from the adjacency matrix (**A**). The adjacency matrix, or vertex adjacency matrix, is an important source for the calculation of molecular descriptors.[49] This is one of the fundamental graph theoretical matrices and represents the whole set of connections between adjacent pairs of atoms.[50] The adjacency matrix gives information about branching, which is demonstrated to be relevant for biodegradation modeling. This was confirmed in the PCA plots, where these matrix-based descriptors were always related to NRB compounds. Nevertheless, matrix-based molecular descriptors were never included before in already published biodegradation QSAR models and thus their relationship with biodegradation could not be directly verified and needs further investigation.

2D matrix-based descriptors are topological indices calculated by applying a set of basic algebraic operators to different graph-theoretical matrices representing the H-depleted molecular graph of molecules.[49]

Laplace matrix **L** is obtained by the difference between a diagonal vertex degree matrix and the adjacency matrix **A**:

$$[\mathbf{L}]_{ij} = \begin{cases} -1 & \text{if } (i, j) \in \mathrm{E(G)} \\ \delta_i & \text{if } i = j \\ 0 & \text{if } (i, j) \notin \mathrm{E(G)} \end{cases}$$

where δ_i is the ith vertex degree, that is, the number of vertices adjacent to vertex i and E(G) is the set of graph edges.

Burden matrices $\mathbf{B}(w)$ are augmented adjacency matrices defined to account for heteroatoms and bond multiplicity as the following:

$$[\mathbf{B}(w)]_{ij} = \begin{cases} \sqrt{\pi_{ij}^*} & \text{if } (i, j) \in \mathrm{E(G)} \\ \dfrac{w_i}{w_C} & \text{if } i = j \\ 0.001 & \text{if } (i, j) \notin \mathrm{E(G)} \end{cases}$$

The diagonal elements are atomic carbon-scaled properties (e.g., mass (m), polarizability (p)); the off-diagonal elements corresponding to pairs of bonded atoms are the square roots of conventional bond orders π^* (i.e., 1, 2, 3, and 1.5 for single, double, triple, and aromatic bonds, respectively); all other matrix elements are set at 0.001.

Barysz matrices $\mathbf{Dz}(w)$ are weighted distance matrices that were defined on the basis of a generalization of Barysz

dx.doi.org/10.1021/ci4000213 | J. Chem. Inf. Model. 2013, 53, 867−878

Table 8. List of OLS Models Built to Describe the Matrix-Based Descriptors[a]

descriptor	R^2	Q^2	model equations
HyWi_B(m)	0.94	0.94	−0.297 + 0.001MW + 0.745MWC01 + 3.446Eta_alpha_A
J_Dz(e)	0.82	0.79	−1.712 − 0.376nCIC + 2.212Xindex − 0.215NRS + 1.143piPC02
SM6_B(m)	0.87	0.86	3.432 + 0.005ZM1Mad + 1.927B01[C-Br] + 5.88Eta_alpha_A + 0.566piPC02
SM6_L	0.99	0.99	−0.416 + 1.04SRW08 + 3.683X0A
SpMax_A	0.96	0.95	0.495 + 0.267SRW08 − 0.087RDCHI
SpMax_B(m)	0.78	0.78	3.267 + 6.509B01[C-X] + 3.193B01[C-Br] + 0.135piPC04
SpMax_L	0.93	0.93	−8.339 + 1.678SRW08 + 9.693X1A − 1.49MWC01
SpPosA_B(p)	0.86	0.86	1.613 + 0.858Eta_alpha_A − 0.831Mi + 0.004C%
TI2_L	0.94	0.94	1.763 + 1.747MSD + 2.408RDCHI

[a]The squared correlation coefficient in fitting (R^2), in cross-validation with five cancellation groups (Q^2), and the model equations are provided.

weighting scheme in terms of conventional bond orders π^* and any atomic property:[51]

$$[\mathbf{Dz}(w)]_{ij} = \begin{cases} d_{ij}(w, \pi^*) & \text{if } i \neq j \\ 1 - \dfrac{w_C}{w_i} & \text{if } i = j \end{cases}$$

$$d_{ij}(w, \pi^*) = \sum_{b=1}^{d_{ij}} \left(\frac{1}{\pi_b^*} \frac{{w_C}^2}{w_{b(1)} w_{b(2)}} \right)$$

where w_C is any atomic property (e.g., Sanderson electronegativity (e)) of the carbon atom and w_i the corresponding value of the ith atom; $d_{ij}(w,\pi^*)$ is a weighted topological distance calculated by summing the edge weights over all bonds involved in the shortest path between vertices v_i and v_j; the subscripts $b(1)$ and $b(2)$ represent the two vertices incident to the considered bth edge.

The topological indices that were derived from these graph matrices and found to be related to ready biodegradability of organic compounds are briefly defined below.

The hyper-Wiener-like index, HyWi_B(m), is calculated according to the following formula:

$$\mathrm{HyWi\_B(m)} = \ln\left\{ 1 + \frac{1}{2} \sum_{i=1}^{\mathrm{nSK}} \sum_{j=i}^{\mathrm{nSK}} \left([\mathbf{B}(\mathrm{m})]_{ij}^2 + [\mathbf{B}(\mathrm{m})]_{ij} \right) \right\}$$

where nSK is the number of non-H atoms and defines the matrix dimension. This index is sensitive to molecule size and for a given size it takes minimum values for linear hydrocarbons while increases both with the number of heavy atoms and branching involving multiple bonds.

The Balaban-like index, J_Dz(e), is similar to Randić connectivity index but calculated with a normalization factor that makes it independent of the molecule size and cyclicity degree:

$$\mathrm{J\_Dz(e)} = \frac{\mathrm{nBO}}{\mathrm{nCIC} + 1} \sum_{i=1}^{\mathrm{nSK}-1} \sum_{j=i+1}^{\mathrm{nSK}} a_{ij} [VS_i(\mathbf{Dz};\, \mathrm{e}) VS_j(\mathbf{Dz};\, \mathrm{e})]^{-1/2}$$

where the vertex degrees are replaced by the matrix row sums VS and elements a_{ij} of the adjacency matrix are introduced to account only for contributions from bonded atom pairs; nBO is the number of graph edges, and nCIC the number of independent rings in the molecule.

SpMax_A, SpMax_B(m), SpMax_L, SpPosA_B(p), SM6_B(m), SM6_L, and TI2_L are spectral indices calculated as function of the matrix eigenvalues.[52] In particular, SpMax is the leading eigenvalue, that is, the largest eigenvalue of the matrix spectrum, and SpPosA is the normalized spectral positive sum index, that is, the sum of the positive eigenvalues divided by the number of non-H atoms in order to reduce molecule size influence.[53] The leading eigenvalue of the adjacency matrix **A** (SpMax_A) is the well-known Lovasz−Pelikan index,[54] which was demonstrated to be related to molecular branching. Both SpMax_A and SpMax_L demonstrated to be able to characterize a large group of NRB molecules containing halogens (especially F and Cl) as the substituents in nonterminal positions along the molecular structure.

The spectral moment of sixth order (SM6) is the sum of the sixth power of all of the matrix eigenvalues. Since SM6_B(m) and SM6_L are derived from modified adjacency matrices, these indices are to some extent related to the number of self-returning walks of length six in the molecule, which can also be expressed as linear combinations of counts of certain fragments contained in the molecular graph.[55] These indices tend to increase with molecular branching and cyclicity. Moreover, the index SM6_B(m) is able to characterize a group of about 50 NRB compounds including heavy atoms (e.g., Sn and Br) and with large ramification or number of rings. The second Mohar index (TI2_L) is calculated as the inverse of the smallest nonzero eigenvalue of the Laplace matrix, which is weighted by the number of non-hydrogen atoms.[56] This index does not account for the presence of different heteroatoms in molecules but is very sensitive to structural features such as branching and cyclicity. It increases with the number of non-H atoms, and in a series of equal-sized molecules it discriminates between linear chains (high values) and branched/cyclic structures that typically are not ready biodegradable.

In order to improve the interpretability of the proposed QSAR models, matrix-based descriptors were further analyzed to elucidate the information they encode. For this purpose, ordinary least squares (OLS) regression was used to investigate the existing relationships between these targeted matrix-based descriptors and other DRAGON molecular descriptors, which

dx.doi.org/10.1021/ci4000213 | J. Chem. Inf. Model. 2013, 53, 867−878

were used as the independent variables in the regression models. A variable selection procedure based on GAs was carried out to search for the optimal subset of DRAGON descriptors related to each matrix-based descriptor. Regression models were optimized on the basis of the squared correlation coefficient Q^2 calculated in cross-validation with five cancellation groups.[57,58]

In addition to the already calculated descriptors (Table 3), the following DRAGON blocks were considered for this analysis: connectivity indices, topological information indices, walk and self-returning walk counts, and Extended Topochemical Atom (ETA) indices. These indices were selected as they encode to different extent information about molecular branching.[49] Statistics of the OLS models calculated for each matrix-based descriptor are collected in Table 8. Regression models included a maximum number of four descriptors; high and balanced performance in fitting (R^2) and cross-validation (Q^2) demonstrated good consistency as well as ability in describing the chemical information encoded by matrix-based descriptors. In particular, the regression models for HyWi_B(m), SM6_L, SpMax_A, SpMax_L, and TI2_L had R^2 and Q^2 higher than 0.9; the models for SM6_B(m) and SpPosA_B(p) gave R^2 and Q^2 higher than 0.8, while just two models (J_Dz(e), SpMax_B(m)) had Q^2 between 0.78 and 0.79.

On the basis of these results, it could be concluded that the considered matrix-based descriptors mainly encode chemical information related to branching, cyclicity, and molecular size, which were demonstrated to be important factors related to biodegradability. In addition, the obtained OLS models also proved that matrix-based descriptors are highly information rich, since they were modeled by several other descriptors, each encoding different chemical information (Table 9). This feature makes matrix-based descriptors particularly interesting to QSAR modeling, since using descriptors able to encode different molecular features can lead to more parsimonious models including a limited number of variables.

Table 9. Molecular Descriptors Selected in the OLS Models Describing the Matrix-Based Descriptors

symbol	description	DRAGON Block
B01[C-Br]	presence/absence of C−Br at topological distance 1	2D atom pairs
B01[C-X]	presence/absence of C−X at topological distance 1	2D atom pairs
C%	percentage of C atoms	constitutional indices
Eta_alpha_A	eta average core count	ETA indices
Mi	mean first ionization potential (scaled on Carbon atom)	constitutional indices
MSD	mean square distance index (Balaban)	topological indices
MW	molecular weight	constitutional indices
MWC01	molecular walk count of order 1	walk and path counts
nCIC	number of rings (cyclomatic number)	ring descriptors
NRS	number of ring systems	ring descriptors
piPC02	molecular multiple path count of order 2	walk and path counts
piPC04	molecular multiple path count of order 4	walk and path counts
RDCHI	reciprocal distance sum Randic-like index	connectivity indices
SRW08	self-returning walk count of order 8	walk and path counts
X0A	average connectivity index of order 0	connectivity indices
X1A	average connectivity index of order 1	connectivity indices
Xindex	Balaban X index	information indices
ZM1Mad	first Zagreb index by Madan vertex degrees	topological indices

4. CONCLUSIONS

The aim of this study was to develop reliable classification QSAR models for ready biodegradability. Experimental values were collected from the MITI database and screened to obtain a consistent data set that meets the requirements of the OECD guidelines. The structure representations of the compounds, as well as the collected experimental data, were accurately verified. The resulting data set was split into training and test sets before modeling. Genetic algorithms coupled with three different classification algorithms (*k*NN, PLSDA, and SVM) were applied in order to select the optimal subset of molecular descriptors. The three models and the derived consensus analysis demonstrated good statistics in fitting and cross-validation as well as high accuracy in prediction for the test set with respect to already published models on biodegradation. The lowest ER in classification was reached by means of *k*NN, which gave an ER equal to 0.12 for the test set, and consensus analysis, which gave an error rate equal to 0.06 with 23% of not assigned molecules. The developed models were further validated using an external validation set collected from different sources, and good classification performances were obtained. The proposed models showed a balance between specificity and sensitivity values, as well as similar performances in training, test, and external validation sets, which can indicate the absence of overfitting. The potential relationships between the selected molecular descriptors and biodegradability were evaluated by comparing with information from the literature.

Matrix-based molecular descriptors, which were used for the first time to model biodegradability, were further analyzed. The information they encoded was evaluated by means of regression OLS models based on other types of molecular descriptors. Relationships between matrix-based descriptors and biodegradability were highlighted, since they contained information about molecular branching and size. In general, this family of molecular descriptors appeared to be interesting for QSAR modeling, since they were information rich and thus by using them the total number of descriptors to be used to model a defined endpoint could be reduced.

■ ASSOCIATED CONTENT

Supporting Information

Training, test, and external validation sets, CAS numbers, SMILES structures, and experimental class (ready/not ready biodegradable: RB/NRB) of the molecules as well as the values of molecular descriptors included in the QSAR models. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: davide.ballabio@unimib.it. Mailing address: Department of Environmental Sciences, University of Milano-Bicocca

P.zza della Scienza, 1-20126 Milano, Italy. Telephone: +39-02-6448.2801. Fax: +39-02-6448.2839.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under Grant Agreement no. 238701 of the Marie Curie ITN Environmental Chemoinformatics (ECO) project: http://www.eco-itn.eu/.

■ REFERENCES

(1) Boethling, R. S. Designing Biodegradable Chemicals. In *Designing Safer Chemicals*; DeVito, S. C., Garrett, R. L., Eds.; American Chemical Society: Washington, DC, 1996; Vol. *640*, pp 156−171.

(2) European Commission. Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 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. *Off. J. Eur. Union* **2006**, *L396*, 1−849.

(3) Rorije, E.; Loonen, H.; Müller, M.; Klopman, G.; Peijnenburg, W. J. G. M. Evaluation and application of models for the prediction of ready biodegradability in the MITI-I test. *Chemosphere* **1999**, *38*, 1409−1417.

(4) Allanou, R.; Hansen, B. G.; Van der Bilt, Y. *Public availability of data on EU high production volume chemicals*; European Communities: Italy, 1999; Report EUR 18996 EN.

(5) Pavan, M.; Worth, A. P. Review of estimation models for biodegradation. *QSAR Comb. Sci.* **2008**, *27*, 32−40.

(6) Raymond, J. W.; Rogers, T. N.; Shonnard, D. R.; Kline, A. A. A review of structure-based biodegradation estimation methods. *J. Hazard. Mater.* **2001**, *84*, 189−215.

(7) Jaworska, J. S.; Boethling, R. S.; Howard, P. H. Recent developments in broadly applicable structure-biodegradability relationships. *Environ. Toxicol. Chem.* **2003**, *22*, 1710−1723.

(8) Baker, J. R.; Gamberger, D.; Mihelcic, J. R.; Sabljić, A. Evaluation of artificial intelligence based models for chemical biodegradability prediction. *Molecules* **2004**, *9*, 989−1003.

(9) Geating, J. *Literature study of the biodegradability of chemicals in water*; U.S. EPA: Cincinnati, OH, 1981; Report EPA-600/2-81-175.

(10) Niemi, G. J.; Veith, G. D.; Regal, R. R.; Vaishnav, D. D. Structural features associated with degradable and persistent chemicals. *Environ. Toxicol. Chem.* **1987**, *6*, 515−527.

(11) Boethling, R. S.; Sabljić, A. Screening-level model for aerobic biodegradability based on a survey of expert knowledge. *Environ. Sci. Technol.* **1989**, *23*, 672−679.

(12) Howard, P. H.; Boethling, R. S.; Stiteler, W.; Meylan, W.; Beauman, J. Development of a predictive model for biodegradability based on BIODEG, the evaluated biodegradation data base. *Sci. Total Environ.* **1991**, *109−110*, 635−641.

(13) Howard, P. H.; Boethling, R. S.; Stiteler, W. M.; Meylan, W. M.; Hueber, A. E.; Beauman, J. A.; Larosche, M. E. Predictive model for aerobic biodegradability developed from a file of evaluated biodegradation data. *Environ. Toxicol. Chem.* **1992**, *11*, 593−603.

(14) Boethling, R. S.; Howard, P. H.; Meylan, W.; Stiteler, W.; Beauman, J.; Tirado, N. Group contribution method for predicting probability and rate of aerobic biodegradation. *Environ. Sci. Technol.* **1994**, *28*, 459−465.

(15) Gamberger, D.; Horvatić, D.; Sekušak, S.; Sabljić, A. Applications of experts' judgement to derive structure-biodegradation relationships. *Environ. Sci. Pollut. Res. Int.* **1996**, *3*, 224−228.

(16) Loonen, H.; Llndgren, F.; Hansen, B.; Karcher, W.; Nlemela, J.; Hiromatsu, K.; Takatsuki, M.; Peunenburg, W.; Rorije, E.; Struijs, J. Prediction of biodegradability from chemical structure: Modeling of ready biodegradation test data. *Environ. Toxicol. Chem.* **1999**, *18*, 1763−1768.

(17) 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. *Environ. Toxicol. Chem.* **2000**, *19*, 2478−2485.

(18) Huuskonen, J. Prediction of biodegradation from the atom-type electrotopological state indices. *Environ. Toxicol. Chem.* **2001**, *20*, 2152−2157.

(19) Jaworska, J.; Dimitrov, S.; Nikolova, N.; Mekenyan, O. Probabilistic assessment of biodegradability based on metabolic pathways: catabol system. *SAR QSAR Environ. Res.* **2002**, *13*, 307−323.

(20) Sedykh, A.; Klopman, G. Data analysis and alternative modelling of MITI-I aerobic biodegradation. *SAR QSAR Environ. Res.* **2007**, *18*, 693−709.

(21) Alikhanidi, S.; Takahashi, Y. Pesticide Persistence in the Environment-Collected Data and Structure-Based Analysis. *J. Comp. Chem. (Japan)* **2004**, *3*, 59−70.

(22) 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 assessment of chemical biodegradability. *J. Chem. Inf. Model.* **2012**, *52*, 655−669.

(23) Klopman, G.; Balthasar, D. M.; Rosenkranz, H. S. Application of the computer-automated structure evaluation (CASE) program to the study of structure-biodegradation relationships of miscellaneous chemicals. *Environ. Toxicol. Chem.* **1993**, *12*, 231−240.

(24) Nendza, M. Prediction of Persistence. In *Predicting Chemical Toxicity and Fate*; Cronin, M., Livingstone, D., Eds.; CRC Press: Boca Raton, FL, 2004.

(25) Gramatica, P.; Cassani, S.; Roy, P. P.; Kovarich, S.; Yap, C. W.; Papa, E. QSAR Modeling is not "Push a Button and Find a Correlation": A Case Study of Toxicity of (Benzo-)triazoles on Algae. *Mol. Inform.* **2012**, *31*, 817−835.

(26) CHRIP National Institute of Technology and Evaluation (NITE) of Japan, Chemical Risk Information Platform http://www.safe.nite.go.jp/english/kizon/KIZON_start_hazkizon.html (accessed Jan 16, 2012).

(27) Organisation for Economic Co-operation and Development. *Test No. 301: Ready Biodegradability*; OECD Publishing: Paris, 1992.

(28) *EPA 712-C-98-076 Fate, Transport, and Transformation Test Guidelines*; EPA: Washington, D.C., 1998; OPPTS 835.3110.

(29) *ChemSpider*; Royal Society of Chemistry: Cambridge, http://www.chemspider.com/ (accessed Oct 29, 2012).

(30) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*; Springer: New York, 2007.

(31) 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. *Anal. Chem.* **1991**, *63*, 139−146.

(32) *DRAGON* (Software for Molecular Descriptor Calculations) version 6.0.28; Talete srl: Milano, Italy, 2012.

(33) Kowalski, B. R.; Bender, C. F. The K-nearest neighbor classification rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation. *Anal. Chem.* **1972**, *44*, 1405−1411.

(34) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(35) Ståhle, L.; Wold, S. Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. *J. Chemom.* **1987**, *1*, 185−196.

(36) Cortes, C.; Vapnik, V. Support-Vector Networks. *In Machine Learning* **1995**, 273−297.

(37) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory; COLT '92*, Pittsburgh, PA, July 27−29; ACM: New York, 1992; pp 144−152.

(38) Leardi, R.; Lupiáñez González, A. Genetic algorithms applied to feature selection in PLS regression: How and when to use them. *Chemom. Intell. Lab. Syst.* **1998**, *41*, 195−207.

(39) Baurin, N.; Mozziconacci, J.-C.; Arnoult, E.; Chavatte, P.; Marot, C.; Morin-Allory, L. 2D QSAR Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 276−285.
(40) Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A.; Xie, Q.; Tong, W. Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis* **2004**, *19*, 365−377.
(41) Ganguly, M.; Brown, N.; Schuffenhauer, A.; Ertl, P.; Gillet, V. J.; Greenidge, P. A. Introducing the Consensus Modeling Concept in Genetic Algorithms: Application to Interpretable Discriminant Analysis. *J. Chem. Inf. Model.* **2006**, *46*, 2110−2124.
(42) Hewitt, M.; Cronin, M. T. D.; Madden, J. C.; Rowe, P. H.; Johnson, C.; Obi, A.; Enoch, S. J. Consensus QSAR Models: Do the Benefits Outweigh the Complexity? *J. Chem. Inf. Model.* **2007**, *47*, 1460−1468.
(43) Environment Canada DSL (Domestic Substances List) http://www.ec.gc.ca/lcpe-cepa/default.asp?lang=En&n=5F213FA8-1&wsdoc=D031CB30-B31B-D54C-0E46-37E32D526A1F (accessed Nov 4, 2012).
(44) CEPA 1999 Canadian Environmental Protection Act http://laws-lois.justice.gc.ca/eng/regulations/SOR-2000-107/page-1.html#footnotea_e-ID0EFBCA (accessed Nov 14, 2012).
(45) Chang, C.-C.; Lin, C.-J. *LIBSVM: a library for support vector machines*; National Taiwan University, Department of Computer Science: Taipei, Taiwan, 2001.
(46) *MATLAB*, version 7.13.0.564; MathWorks: Natick, MA, 2011; www.mathworks.com.
(47) Topliss, J. G.; Edwards, R. P. Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.
(48) Manallack, D. T.; Livingstone, D. J. Artificial neural networks: Application and chance effects for QSAR data analysis. *Med. Chem. Res.* **1992**, *2*, 181−190.
(49) Todeschini, R.; Consonni, V. *Molecular descriptors for chemoinformatics*; Wiley-VCH: New York, 2009.
(50) Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1992; pp 225−273.
(51) Ivanciuc, O. QSAR Comparative Study of Wiener Descriptors for Weighted Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1412−1422.
(52) Consonni, V.; Todeschini, R. New spectral indices for molecule description. *Match* **2008**, *60*, 3−14.
(53) Balaban, A. T.; Ciubotariu, D.; Medeleanu, M. Topological indices and real number vertex invariants based on graph eigenvalues or eigenvectors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 517−523.
(54) Lovasz, L.; Pelikan, J. On the eigenvalue of trees. *Period Math Hung* **1973**, *3*, 175−182.
(55) Barysz, M.; Trinajstić, N. A novel approach to the characterization of chemical structures. Int. *J. Quantum Chem. Quant. Chem. Symp.* **1984**, *18*, 661−673.
(56) Trinajstić, N.; Babic, D.; Nikolić, S.; Plavšić, D.; Amić, D.; Mihalić, Z. The Laplacian matrix in chemistry. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 368−376.
(57) Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the Definition of the Q2 Parameter for QSAR Validation. *J. Chem. Inf. Model.* **2009**, *49*, 1669−1678.
(58) Consonni, V.; Ballabio, D.; Todeschini, R. Evaluation of model predictive ability by external validation techniques. *J. Chemom.* **2010**, *24*, 194−201.