

In Silico Assessment of Chemical Biodegradability

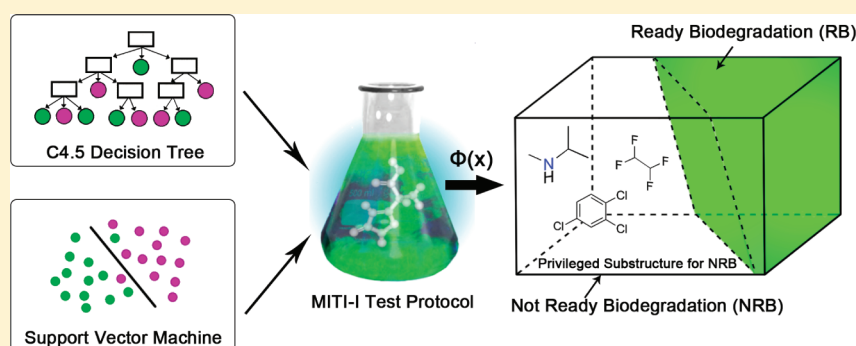
Feixiong Cheng,[†] Yutaka Ikenaga,[§] Yadi Zhou,[†] Yue Yu,[†] Weihua Li,[†] Jie Shen,[†] Zheng Du,[†] Lei Chen,[†] Congying Xu,[†] Guixia Liu,[†] Philip W. Lee,^{*,†,‡} and Yun Tang^{*,†}

[†]Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China

[‡]Graduate School of Agriculture, Kyoto University, Kitashirakawa Oiwake-cho, Sakyo-ku, Kyoto 606-8502, Japan

[§]Safety Assessment Division, Chemical Management Center, National Institute of Technology and Evaluation (NITE), 2-49-10 Nishihara, Shibuya-ku, Tokyo 151-0066, Japan

S Supporting Information



ABSTRACT: Biodegradation is the principal environmental dissipation process. Due to a lack of comprehensive experimental data, high study cost and time-consuming, *in silico* approaches for assessing the biodegradable profiles of chemicals are encouraged and is an active current research topic. Here we developed *in silico* methods to estimate chemical biodegradability in the environment. At first 1440 diverse compounds tested under the Japanese Ministry of International Trade and Industry (MITI) protocol were used. Four different methods, namely support vector machine, *k*-nearest neighbor, naïve Bayes, and C4.5 decision tree, were used to build the combinatorial classification probability models of ready versus not ready biodegradability using physicochemical descriptors and fingerprints separately. The overall predictive accuracies of the best models were more than 80% for the external test set of 164 diverse compounds. Some privileged substructures were further identified for ready or not ready biodegradable chemicals by combining information gain and substructure fragment analysis. Moreover, 27 new predicted chemicals were selected for experimental assay through the Japanese MITI test protocols, which validated that all 27 compounds were predicted correctly. The predictive accuracies of our models outperform the commonly used software of the EPI Suite. Our study provided critical tools for early assessment of biodegradability of new organic chemicals in environmental hazard assessment.

INTRODUCTION

In the past several decades, many pesticides and industrial chemicals, such as dichlorodiphenyltrichloroethane (DDT), chlordane, and dieldrin, were removed from the market due to their environmental persistence, bioaccumulation, and toxicity (PBT) properties. Persistence is defined as the length of time a substance remains in the environment. A common criterion of a substance's transformation is its biodegradation half-life in the environment. Biodegradation being the principal environment dissipation process is one of the most important parameters influencing the toxicity, transformation, and ultimate fate of an organic chemical in the aquatic and terrestrial ecosystems.^{1–3} Several standardized methods used to determine the extent of biodegradation in water, sediment, and soil environments have been developed by different regulatory organizations, such as the Organization for Economic Co-operation and Development

(OECD), International Organization for Standardization (ISO), Japanese Ministry of International Trade and Industry (MITI), National Institute of Technology and Evaluation (NITE), European Union (EU), and United States Environmental Protection Agency (US-EPA). The large number of existing chemicals in the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) database that have to be evaluated is well over 100 000, making the experimental testing of every single compound an almost impossible task.⁴

Because of the lack of comprehensive experimental data, high study cost, and animal welfare, the use of *in silico* approaches for assessing the PBT profiles of chemicals is encouraged.⁵ The US-EPA and European Chemical Agency (ECHA) are

Received: December 23, 2011

Published: February 14, 2012

responsible to prioritize the hazard/risks associated with the thousands of chemicals used in commerce. Much effort has been devoted to develop reliable quantitative structure–biodegradability relationship (QSBR) models to predict the biodegradation of new chemicals in natural systems. For a scientifically valid QSBR model, certain criteria must be met: (i) understandable mechanisms of biodegradation, (ii) classification of chemicals according to relative biodegradability, (iii) data sets of diverse chemical structures.¹

In the past decades, several statistical QSBR models had been developed.^{6–13} Howard et al. developed linear and nonlinear models using molecular weight (MW) and 36 fragment descriptors. The overall predictive accuracy was 90% of a linear model and 93% of a nonlinear model for the training set.⁸ Tunkel et al. built linear and nonlinear classification models of ready biodegradability (RB) versus not ready biodegradability (NRB) using 884 tested chemicals obtained from MITI. The overall predictive accuracy of the test set was 81% for both linear and nonlinear models.⁹ Hiromatsu et al. built an empirical flowchart to predict biodegradability. The model was validated using MITI data of 177 monobenzene derivatives and 168 acyclic compounds, resulting in the predictive accuracies at 94% and 88%, respectively.¹⁰ Philipp et al. combined QSBR with the systematic collection of biochemical knowledge to establish rules to predict the aerobic biodegradation of 194 N-heterocycles. The overall predictive accuracy of a 10-fold cross validation was 68% and 70% for the two models, respectively.¹¹ Detailed reviews about QSBR models can be found in the literature.^{1,14} Nevertheless, these models were built based on the limited chemical diverse data set or lack of an easily understandable mechanism of chemical biodegradation.

In this paper, we used a broad selection of machine learning and feature reduction methods to predict chemical biodegradability based on the largest heterogeneous data set of more than 1600 diverse compounds. 148 physicochemical descriptors and 7 types of fingerprints were calculated separately using open source tools and systematically selected by four different feature reduction methods, including linear correlation analysis, decision tree analysis, and genetic algorithm (GA). Models with reasonably high predictive accuracies were built using support vector machine (SVM), *k*-nearest neighbor (*k*-NN), naive Bayes (NB), and C4.5 decision tree (C4.5DT) algorithms. The impact of eight physicochemical descriptors on chemical biodegradable mechanism was carefully investigated. Some privileged substructures¹⁵ for RB or NRB chemicals were identified by combining substructure fragment and information gain analysis. Moreover in a blind test, 27 novel compounds were predicted first using the global model, and all the 27 compounds were correctly validated by experimental assays.

MATERIALS AND METHODS

Data Set Construction. The entire chemical data set was obtained from two sources. The first one was the Japanese MITI data set with biological oxygen demand (BOD) values, containing 389 new compounds released from 2000 to 2009 and 561 new compounds released before 2000 by Japanese NITE (<http://www.safe.nite.go.jp/english/>). These biodegradation data have not been used in previous publications of the QSBR model development. The other was the BIOWIN data set,⁹ including 884 compounds tested using MITI-I protocol, also known as OECD test guideline 301C.^{16,17} For the

modeling purpose, substances were scored +1 or –1 based on the BOD values in either 28-day tests (for the OECD version of MITI-I) or 14-day tests (the original MITI-I protocol, only some substances applicable): if BOD \geq 60% of maximum theoretical oxygen demand, the substance was scored +1 and labeled as RB; otherwise, the substance was scored –1 and labeled as NRB. The detailed description about the scoring criteria can be found in the literature.^{9,14}

Chemical 2D structures were obtained from the US-EPA Aggregated Computational Toxicology Resource (ACToR) database¹⁸ by CAS number mapping search using in-house scripts. All chemical structures were confirmed with the PubChem database.¹⁹ Entries containing inorganic compounds, noncovalent complexes, and mixtures were excluded from modeling. Salts were converted into the corresponding acids or bases; water molecules were removed from hydrates. In addition, the duplicated compounds were excluded. By applying these criteria, a large diverse biodegradability database containing 1440 unique compounds was obtained (Table 1).

Table 1. Detailed Statistical Number of Chemicals Used in the Training Set, Test Set, and External Validation Set

data sets	ready biodegradability	not ready biodegradability	total
training set	529	911	1440
test set	62	102	164
external validation set	4	23	27

In addition, 164 diverse compounds were collected from the US-EPA database and literature^{8,20} designated as the external test set. The CAS number, SMILES, RB, and NRB scores of 1604 unique compounds are listed in Supporting Information Table S1.

Calculation of Molecular Descriptors. *Molecular Descriptors.* Overall, about 500 different molecular descriptors were calculated using open source software of the PaDEL-Descriptor.²¹ Descriptors with more than 95% zero value or zero variance were removed. The remaining 148 descriptors (Supporting Information Table S2) were used for further chemical feature reduction.

Feature Reduction. Four data reduction approaches were used to reduce the features: (1) Correlation-based feature selection (CFS) was conducted by performing linear regressions for every descriptor. The top ten best correlation descriptors were selected. (2) Classification and regression tree algorithm (CART) was used.²² The Gini coefficient was used as a homogeneity measure for CART. (3) Chi-squared automatic interaction detector (CHAID) was used.²³ (4) The typical combinatorial method of the SVM with genetic algorithm (GASVM) was used to capture the most informative descriptors and a step-by-step selecting process was used. For each step, the population size is set to 25 and the current step will be terminated after reproducing 100 generations. In each generation, the Matthews Correlation Coefficient²⁴ from 5-fold cross validation for each chromosome was used as the fitness value of each chromosome. The mutation rate is set to 0.005, and the crossover rate is set randomly. During the selection process, the roulette wheel selection method developed in our previous work²⁵ was used, and the most fitted chromosome known as elite was retained for the whole reproduction process. At the end of each step, the final elites were chose for the next

selecting step. The whole feature selection will not be ended until the expected number of variables was obtained.

CART and CHAID were performed with PASW Statistics version 18 for Windows (<http://www.spss.com/statistics/>) with the default settings and linear correlation analysis with our in-house scripts. The GASVM method with MATLAB code can be obtained by author request.

Calculation of Molecular Fingerprints. Seven kinds of fingerprints implemented in PaDEL-Descriptor²¹ were evaluated in Supporting Information Table S2. These fingerprints include the CDK fingerprint (FP), CDK extended fingerprint (ExtFP), Estate fingerprint (EstateFP), MACCS keys (MACCS), PubChem fingerprint (PubChemFP), Substructure fingerprint (FP4), and Klekota-Roth fingerprint (KRFP). The detailed descriptions about these fingerprints can be found in the original literature.^{21,26}

Modeling Methods. Four different methods, including SVM, C4.5 DT, *k*-NN, and NB, were used. SVM algorithm was performed by the LIBSVM2.9 package.²⁷ C4.5 DT, *k*-NN, and NB were performed in Orange 2.0 (version 2.0b, available free of charge at Web site: <http://www.ailab.si/orange/>). In this study, all models developed here get a probability output instead of estimated target values (such as +1 or -1), which had been described in our previous published work.²⁴ In addition, a special applicability domain (AD) based on distance^{28–30} was introduced to avoid prediction for compounds which differ substantially from those in the training set.

Support Vector Machine (SVM). The SVM technique, originally developed by Vapnik³¹ for pattern recognition, aims at minimizing the structural risk under the frame of VC theory. Each molecule is represented by an eigenvector **t**, and the selected patterns t_1, t_2, \dots, t_n make up the components of **t**. For SVM training, the category label *y* was added. The *i*th molecule in the data set is defined as $M_i = (t_i, y_i)$, where $y_i = +1$ for the RB category and $y_i = -1$ for the NRB category. SVM gives a classifier:

$$f(\mathbf{t}) = \text{sgn} \left(\frac{1}{2} \sum_{i=1}^n \alpha_i K(\mathbf{t}_i, \mathbf{t}) + b \right) \quad (1)$$

Where, α_i is the coefficient to be learned and *K* is a kernel function. Parameter α_i is trained through maximizing the Lagrangian expression given below:

$$\begin{aligned} & \underset{\alpha_i}{\text{maximize}} \quad \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(t_i, t_j) \\ & \text{subject to:} \quad \sum_{i=1}^n y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq C \end{aligned} \quad (2)$$

The Gaussian radial basis function (RBF) kernel was used. The penalty parameter *C* and different kernel parameter γ were tuned based on the training set using a grid search strategy by 5-fold cross validation in order to obtain a SVM model with optimal performance.

C4.5 Decision Tree (C4.5 DT). C4.5 defines the possible decision tree by means of a hill-climbing search based on the statistical property measure called information gain. The elements of the tree generated by Iterative Dichotomiser 3 (ID3) and C4.5 are either leafs or decision nodes. A detailed descriptions of C4.5DT can be found in the original literature.³²

***k*-Nearest Neighbors (*k*-NN).** The *k*-nearest neighbor algorithm (*k*-NN) is a method to classify objects based on closest examples in the feature space.³³ In this study, the nearness is measured by hamming distance metrics and the parameter of *k* = 3 were used.

Naïve Bayes (NB). For NB classifier,³⁴ it generates the posterior probabilities which were given out directly based on the core function of eq 3.

$$P(C_i|X) = \frac{p_{C_i} p(X|C_i)}{\sum_j p_{C_j} p(X|C_j)} \quad (3)$$

Definition of Model Applicability Domain. In addition, a special applicability domain (AD) based on distance was introduced to avoid prediction for compounds which differ substantially from those in the training set. To test similarity, each compound was represented by a point in the *M*-dimensional vector (166-dimensional MACCS keys) space with the coordinates $X_{i1}, X_{i2}, X_{i3}, \dots, X_{iM}$, where X_i is the value of individual vector. The molecular similarity between any two molecules is characterized by the Euclidean distance. The Euclidean distance d_{ij} between compounds *i* and *j* in *M*-dimensional space was calculated as follows:

$$d_{ij} = \sqrt{\sum_{k=1}^M (X_{ik} - X_{jk})^2} \quad (4)$$

Compounds with the smallest distance have the highest similarity. The AD threshold, D_T was obtained by computing the distribution of distances (pairwise similarities) of compounds as follows:

$$D_T = \bar{\gamma} + Z\sigma \quad (5)$$

Where $\bar{\gamma}$ is the average Euclidean distance of the *k*-NN (*k* = 3) of each compound within the training set, σ is the standard deviation of these Euclidean distance, and *Z* is an arbitrary parameter to control the significance level. The default value of 0.5 was used, which formally places the boundary for which compounds will be predicted at one-half of the standard deviation. Finally, if the distance of the compound in the test set from at least one of its nearest neighbors in the training set exceeds this threshold, the predicting result is considered unreliable. The detailed descriptions can be found in the original literature.^{28–30}

Privileged Substructures Analysis. The privileged substructure fragments were explored using information gain³⁵ and substructure fragment analysis.^{15,36,37} The privileged structure term was first coined by Evans et al. in 1988, which was defined as “a single molecular framework able to provide ligands for diverse receptors”.³⁶ In the QSBR model, if a substructure was more frequently presented in RB chemical class, this substructure was called a privileged substructure involved in chemical biodegradation. The “frequency of a fragment” enrichment factor in an RB chemical class was defined as follows:

$$\begin{aligned} & \text{frequency of a fragment enrichment factor} \\ &= \frac{(N_{\text{fragment}}^{\text{RB}} \times N_{\text{total}})}{(N_{\text{fragment.total}} \times N_{\text{RB}})} \end{aligned} \quad (6)$$

where $N_{\text{fragment}}^{\text{RB}}$ is the number of compounds containing the fragment in the RB chemicals, N_{total} is the total number of compounds, $N_{\text{fragment_total}}$ is the total number of compounds containing this fragment, and N_{RB} is the number of compounds in the RB chemicals class.

Performance Evaluation of Models. Models were validated by a test set of 164 diverse compounds and the 5-fold cross validation techniques. All models were evaluated based on the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The sensitivity ($\text{SE} = \text{TP}/(\text{TP} + \text{FN})$), specificity ($\text{SP} = \text{TN}/(\text{TN} + \text{FP})$), and the overall predictive accuracy ($Q = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$) were calculated. In addition, the receiver operating characteristic (ROC) curve was also plotted. The ROC curve was used to graphically present the model behavior in a visual way. It shows the separation ability of a binary classifier by iteratively setting the possible classifier threshold.³⁸

Experimental Validation. In a blind test, 27 novel compounds were predicted by the above-mentioned combinatorial classification models and further assayed by the OECD test guideline 301C.^{16,17} OECD test guideline 301C is a standard screening test for chemicals for ready biodegradability in aerobic aqueous medium. The detailed description about OECD test guideline 301C can be found as follows:

The biodegradability of the test substance was determined under the following conditions. (1) Sludge sampling sites: sludge samples were collected from 10 or more sites in Japan and mixed. (2) Degradation testing apparatus: closed-system oxygen consumption measuring apparatus. (3) Basic culture medium: Mix 3 mL each of solution a, b, c, and d and add water up to 1 L. The four solutions were prepared as follows: (a) Potassium dihydrogen orthophosphate, KH_2PO_4 (8.50 g), dipotassium hydrogen orthophosphate, K_2HPO_4 (21.75 g), disodium hydrogen orthophosphate dodecahydrate, $\text{Na}_2\text{HPO}_4 \cdot 12\text{H}_2\text{O}$ (44.60 g), and ammonium chloride, NH_4Cl (1.70 g), are dissolved in water to 1 L. The pH value of the solution should be 7.2. (b) Magnesium sulfate heptahydrate, $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$ (22.50 g), is dissolved in water to make up 1 L. (c) Calcium chloride anhydrous, CaCl_2 (27.50 g), is dissolved in water to 1 L. (d) Iron(III) chloride hexahydrate, $\text{FeCl}_3 \cdot 6\text{H}_2\text{O}$ (0.25 g), is dissolved in water to 1 L. The pH of the solution should be 7.2. (4) Addition of the test substance and testing procedure: Prepare the following test vessels (300 mL each) and adjust them to the testing temperature. If the test substance could not be dissolved in water to reach the test concentration, the test substance was pulverized as finely as possible and no solvent or emulsifier was used. (4.1) A test vessel (300 mL size) contained the test substance (100 mg/L) in the basic culture medium. (4.2) Triplicate test vessels each contained the test substance (100 mg/L) and the active sludge (30 mg/L) in the basic culture medium. (4.3) A test vessel containing aniline (100 mg/L) and the active sludge (30 mg/L) in the basic culture medium was used as the active control. (4.4) A Test vessel containing the active sludge (30 mg/L) in the basic culture medium was used as a blank control sample. (5) Biodegradation test condition and analysis: Samples was incubated in the dark at $25 \pm 1^\circ\text{C}$ under stirring for 28 days. If the percentage degradation of the test substance calculated from the oxygen consumption exceeds 60%, the incubation was terminated after 14 days, in addition to the oxygen consumption measurement. Samples were analyzed for the remaining test substance and its possible degradation

products. (6) Calculation of the BOD value, method for calculating the degradability (%) from oxygen consumption:

$$\text{degradability (\%)} = \frac{\text{BOD}_{\text{test}} - \text{BOD}_{\text{base}}}{\text{TOD}} \quad (7)$$

BOD_{test} : the average value of biochemical oxygen consumption demand of the test substance (mg) in triplicate test vessels (4.2). BOD_{base} : biological oxygen consumption demand of the activated sludge (mg) in the test vessel (4.4). TOD : Theoretical oxygen demand required for complete oxidation of the test substance (mg).

RESULTS

Chemical Space Analysis. Chemical diversity is very important when building a QSBR model. The chemical space distribution (defined by MW and Ghose–Crippen LogKow (ALogP)) of the training set and test set are graphically presented in Figure 1. The test set shared a similar chemical space of the training set. The experimental BOD with numerical values of 817 new molecules (Supporting Information Table S1) in the training set was presented for the first time in this publication.

Performance of Descriptor-Based Models. Physicochemical descriptors selected by the CFS, CART, CHAID, and GASVM methods were summarized in Supporting Information Table S3. The 10 highest scoring descriptors were selected by CFS, CART, and CHAID methods, and the 12 highest scoring descriptors were selected by GASVM methods.

The binary classification models of RB versus NRB chemicals were built using different physicochemical descriptors subsets selected by four different feature reduction methods. The performance of descriptor-based classification models is summarized in Table 2. Comparing the overall performance of different modeling methods, the combinational models of CART-NB, CHAID-SVM, and GASVM-kNN yields the best predictive performance. The high area under the receiver operating characteristic (AUC) are 0.856, 0.844, and 0.873 for CART-NB, CHAID-SVM, and GASVM-kNN models, respectively.

Performance of Fingerprint-Based Models. The combinatorial QSBR models using the seven different fingerprints and four different modeling methods were evaluated. The performance of fingerprints-based models is summarized in Table 3. Comparing the combination of different modeling methods and fingerprints, four kinds of fingerprints, including MACCS keys, EStateFP, PubChemFP, and KRFP yielded the best overall predictive performance. The model of SVM with EStateFP yielded the best performance (SP of 93.1% and AUC of 0.884).

The 5-fold cross-validation technique was used to evaluate the model robustness. Seven global models were built by combining training set and test set using the seven excellent combinatorial modeling methods. The performance of the test set in 5-fold cross-validation is summarized in Table 4. The high AUC value of 0.841–0.969 was obtained for seven good combinatorial QSBR models. These global models with high predictive accuracy could provide useful tools to predict biodegradability of new chemicals in the environmental hazard assessment.

Experimental Validation of Combinatorial Classification Models. The generalization ability of a model decides the usefulness and reliability of models. In a blind test, 27 novel

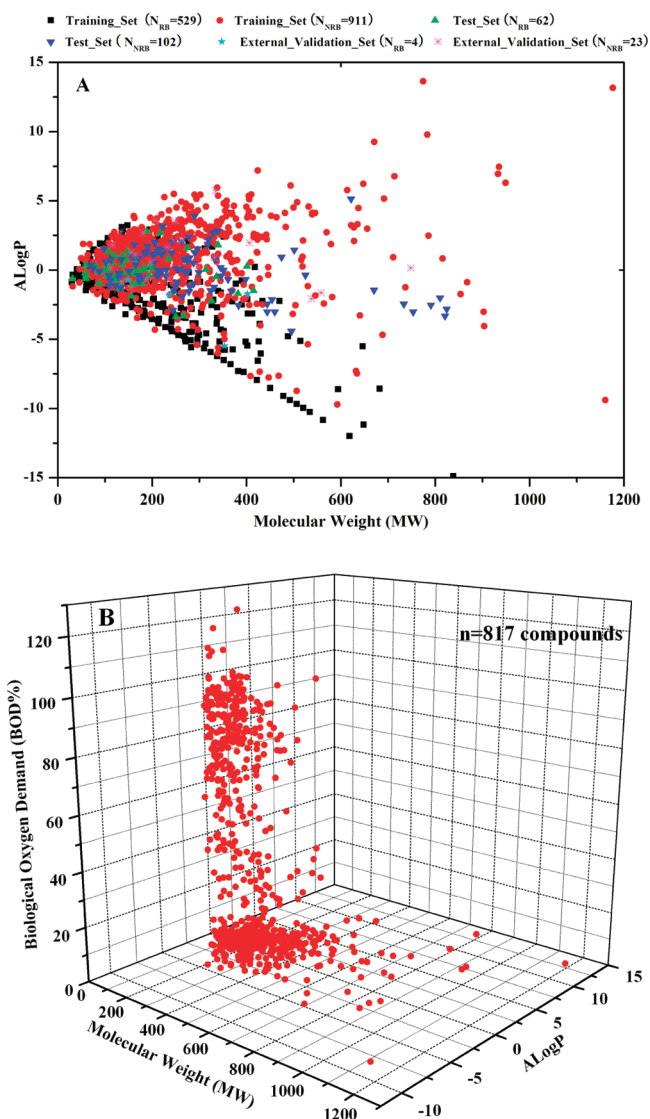


Figure 1. Diversity analysis of the training set ($n = 1440$ compounds), test set ($n = 164$ compounds), external validation set ($n = 27$ compounds), and the 817 novel compounds with biological oxygen demand (BOD%) numerical value. (A) Chemical space defined by molecular weight (MW) and Ghose–Crippen LogKow (ALogP). Note that one data instance (MW = 1625.7, Drug Iotrolan) is not presented in this figure. (B) Distribution of BOD% in a chemical space defined by MW and ALogP. Both figures use the same color scheme. N_{RB} : the number of ready biodegradable chemicals. N_{NRB} : the number of not ready biodegradable chemicals.

chemicals were predicted first using the seven best global classification models and were further assayed using the MITI-I test protocol. The detailed experimental and predicted results with probability output were given in Scheme 1 and Supporting Information Table S4. Six global models, except the GASVM-kNN model, yield the excellent prediction results (Table 5). The high AUC range is from 0.891 to 1.0 for these six global models. In addition, the consensus model using classic classifiers fusion techniques of the mean method described in our previously published work²⁴ was investigated. The overall predictive accuracy of the consensus model was 100.0%; that is, all of the 27 chemicals were predicted correctly. We also predicted 27 novel compounds in the external validation set using the models of *Biowin5* and *Biowin6* implemented in the

software EPI Suite v4.10 (<http://www.epa.gov/oppt/exposure/pubs/episuite.dll.htm>). Comparing the performance (Table 5) of our models with the models of *Biowin5* and *Biowin6*, our methods outperform *Biowin5* and *Biowin6* which were published by Tunkel et al.⁹

Role of the Applicability Domain (AD). It is well-known that the predictive reliability is a very important issue given the fact that any quantitative structure–activity/property relationships (QSAR/QSPR) model is characterized by its AD.³⁹ The seven best modeling methods after applying the AD based on the distance method indicated the improved predictive accuracies for the test set (Table 6). For example, the SE 44.0%, SP 93.2%, Q 80.8%, and AUC 0.793 were obtained for CART-NB model that chemicals were out of domain (OD). The higher performance of SE 89.2%, SP 85.7%, Q 87.7%, and AUC 0.890 was obtained respectively for in domain (ID) chemicals. The detailed description about ID and OD chemicals in the test set are given in Supporting Information Table S1. This observation confirmed that the use of AD obviously improved the predictive accuracies of models, although the improvement came at the expense of lower chemical diversity.^{40,41}

Privileged Substructure for Chemical Biodegradation.

To further explore the structural features of RB and NRB chemicals. Combining information gain and substructure fragment analysis^{35,37} were performed on the entire data set of 1631 unique compounds (by combining the training set, test set, and external validation set) using FP4 and ERFP substructure fingerprints. The privileged substructure fragments characterizing RB and NRB chemicals and the frequency of fragment enrichment factor were identified in Figure 2 and Supporting Information Schemes S1 and S2 and Table S5.

The patterns of quaternary_carbon, qlkylchloride, qlkylbromide, diarylether, amine, tertiary_aliph_amine, primary_arom_amine, secondary_mixed_amine, tertiary_mixed_amine, phenol, arylchloride, arylbromide, nitro, sulfonic_acid, and phosphoric_acid_derivative were presented more frequently in NRB chemicals than RB chemicals (Supporting Information Scheme S1 and Table S5). This indicated that if a chemical included these substructure fragments, this chemical is favorable for NRB. The patterns of alcohol, primary_alcohol, carboxylic_acid, and carboxylic_ester were presented more frequent in RB chemicals than NRB chemicals (Figure 2). Moreover, the patterns of secondary_arom_amine, tertiary_arom_amine, chloroalkene, arylfluoride, aryliodide, hetero_S, and trifluoromethyl were only presented in NRB chemicals. If one pattern was only presented in toxic class, this pattern was called structural alert. It showed that if a chemical included these patterns, this chemical may be NRB.

Klekota–Roth fingerprint is a very good chemical substructure fingerprint enriching compound biological activity.²⁶ Fifteen highly specific privileged substructures were identified by combining information gain and Klekota–Roth substructure analysis (Supporting Information Scheme S2 and Table S5). The patterns of nitrobenzene, ethylbenzene, 2,4,6-trichlorobenzene, 2,4-dichlorobenzene, chlorobenzene, and 2,4-diamino ethane were presented more frequently in NRB chemicals than RB chemicals. The structural alerts of 2,3,5-trichlorobenzene, 2,5-dichlorobenzene, and 1,1,2,2-tetrafluoroethane were only presented in NRB chemicals.

Table 2. Performance of Classification Models for the Training Set and Test Set Using Four Different Feature Selection Methods and Modeling Methods^a

modeling methods	training set				test set			
	SE (%)	SP (%)	Q (%)	AUC	SE (%)	SP (%)	Q (%)	AUC
CART-SVM	85.1	92.3	89.6	0.950	59.7	89.2	78.0	0.797
CART- <i>k</i> NN	100.0	100.0	100.0	1.00	59.7	85.3	75.6	0.748
CART-C4.5DT	85.8	93.3	90.6	0.947	59.7	89.2	78.1	0.786
CART-NB	79.6	78.8	79.1	0.869	71.0	91.2	83.5	0.856
CHAID-SVM	76.0	85.9	82.3	0.891	61.3	91.2	79.9	0.844
CHAID- <i>k</i> NN	100.0	100.0	100.0	1.00	71.0	83.3	78.7	0.834
CHAID-C4.5DT	81.5	88.7	86.0	0.900	56.5	92.2	78.7	0.782
CHAID-NB	71.5	78.6	76.0	0.815	69.4	85.3	79.3	0.817
CFS-SVM	36.7	87.3	68.7	0.793	24.2	91.2	65.9	0.495
CFS- <i>k</i> NN	99.8	99.8	99.8	1.00	37.1	71.6	58.5	0.563
CFS-C4.5DT	52.2	82.4	71.3	0.737	25.8	83.3	61.6	0.584
CFS-NB	59.0	69.9	65.9	0.705	32.3	78.4	61.0	0.620
GASVM-SVM	76.7	89.4	84.7	0.910	66.1	91.2	81.7	0.829
GASVM- <i>k</i> NN	100.0	100.0	100.0	1.00	72.6	91.2	84.2	0.873
GASVM-C4.5DT	88.5	94.2	92.1	0.955	61.3	88.2	78.1	0.752
GASVM-NB	81.3	78.9	79.8	0.868	69.4	87.3	80.5	0.844

^aCFS: correlation-based features selection. CART: the classification and regression tree algorithm. CHAID: chi-squared automatic interaction detector. GASVM: genetic algorithm and support vector machine (SVM). *k*-NN: *k*-nearest neighbors. C4.5DT: C4.5 decision tree. NB: naïve Bayes. SE: sensitivity. SP: specificity. Q: the overall predictive accuracy. AUC: the area under the receiver operating characteristic curve.

DISCUSSION

Comparison of Different Modeling and Feature Reduction Methods. In this paper, four different modeling methods and four different feature reduction methods were used to develop highly predictive models for chemical biodegradability prediction. Comparing the four different modeling methods, namely SVM, C4.5DT, *k*-NN, and NB, the overall performance of SVM was better than the other three ones, especially for fingerprint data description. For example, SVM model with estate fingerprints (EstateFP-SVM) gave the highest AUC value of 0.884 for the test set (Table 3). These results are in agreement with our previously published work that SVM algorithms are good modeling methods in chemical metabolic property prediction and chemical toxicity prediction.^{24,42,43}

Comparing four different feature reduction methods, except CFS, the three feature reduction methods of CART, CHAID, and GASVM yielded the excellent performance. For the *k*-NN modeling method, the highest AUC value of 0.873 was obtained for the test set using the GASVM feature selection method. Unfortunately, the generalization ability of GASVM-*k*NN model was poor, compared with results of the external validation set using the other six global models (Table 2). It is important that the chemical descriptors used in the model are biodegradably relevant, which was not selected by chance. Descriptors of ALogP, nRing, and BCUTp-1h were selected by our used feature reduction methods (Supporting Information Table S3), which are the best or near-best relevant or informative molecular descriptor set (Figure 3).

Diversity of Data Set. The diversity of the data set is another key issue for QSBR models, especially for global models. In the past decades, several QSBR models were developed based on a very limited chemical space, such as a limited set of homologous chemicals.^{1,10,14,20,44} These models provided reliable predictions within narrow AD, but they could not be applied widely due to the narrow AD.⁴⁵ We constructed the combinatorial classification models based on the largest heterogeneous data set of 1604 unique compounds so far,

including 817 unpublished molecules with the numerical BOD value in Supporting Information Table S1 and Figure 1B. The range of the overall predictive accuracy was 74.1–100.0% for the external validation set using the seven best combinational global models and consensus model (Table 5). The data indicated that the models developed here using this large heterogeneous database had good generalization ability.

Relevance of Selected Chemical Descriptor to Biodegradability Mechanism. Apart from applying a high precision modeling algorithm and high quality data, selection of molecular descriptors is also important for optimizing the models. To increase the interpretation of models, the relationships between the biodegradation of 1631 chemicals and 8 key physicochemical descriptors, including ALogP, XLogP, MW, TopoPSA (topological polar surface area), nRing (number of rings), BCUTp-1h (nlow highest polarizability weighted BCUTS), nHBAcc (number of hydrogen bond acceptors), and nHBDOn (number of hydrogen bond donors), are presented in Figures 3 and 4. Student's *t* test was used to evaluate the significance of the difference between paired samples and the means.

ALogP is distributed between −14.89 and 13.62, with a mean of 0.26. XLogP is distributed between −5.61 and 27.66, with a mean of 2.56 (Figure 3). The mean value of ALogP was −0.76 and 0.84 for 595 RB chemicals and 1036 NRB chemicals, respectively, with a *p*-value of 2.66×10^{-41} . This shows that NRB chemicals are likely to have higher ALogP value. As a complementary test, the linear correlation analysis of ALogP and BOD numerical value of 817 novel chemicals (Supporting Information Table S1) is presented in Figure 4. ALogP has a linear correlation ($R = -0.3$) with BOD. The mean value of XLogP was 2.59 and 2.54 for RB and NRB chemicals, respectively, with a *p*-value of 0.77. This indicates that XLogP does not have any significant difference. There is no linear correlation between XLogP and BOD (Figure 4).

MW is an estimation of molecular size and complexity, which was generally used to model chemical biodegradability.⁹ MW is distributed between 30.01 and 1625.73, with a mean of 215.94

Table 3. Performance of Classification Models for the Training Set and Test Set Using Different Fingerprints and Modeling Methods^a

modeling methods	training set				test set			
	SE (%)	SP (%)	Q (%)	AUC	SE (%)	SP (%)	Q (%)	AUC
MACCS-SVM	91.1	98.0	95.5	0.988	59.7	96.1	82.3	0.828
MACCS-kNN	96.6	99.3	98.3	0.997	58.1	82.4	73.2	0.789
MACCS-C4.5DT	88.3	93.6	91.7	0.952	67.7	92.2	82.9	0.815
MACCS-NB	77.9	74.6	75.8	0.844	58.1	85.3	75.0	0.802
FP4-SVM	85.4	94.7	91.3	0.965	62.9	88.2	78.7	0.824
FP4-kNN	89.6	96.2	93.8	0.982	62.9	69.6	67.1	0.734
FP4-C4.5DT	75.8	92.2	86.2	0.899	43.6	79.4	65.9	0.688
FP4-NB	77.1	79.3	78.5	0.869	53.2	74.5	66.5	0.698
PubChemFP-SVM	85.3	93.3	90.3	0.960	66.1	87.3	79.3	0.843
PubChemFP-kNN	98.7	99.6	99.2	1.00	64.5	77.5	72.6	0.749
PubChemFP-C4.5DT	85.8	94.7	91.5	0.949	56.5	93.1	79.3	0.755
PubChemFP-NB	74.7	74.9	74.8	0.818	50.0	82.4	70.1	0.756
EStateFP-SVM	79.8	91.5	87.2	0.936	61.3	93.1	81.1	0.884
EStateFP-kNN	87.2	92.7	90.6	0.955	59.7	77.5	70.7	0.739
EStateFP-C4.5DT	73.2	90.0	83.8	0.878	59.7	85.3	75.6	0.765
EStateFP-NB	70.9	79.7	76.5	0.840	45.2	86.3	70.7	0.822
ExtFP-SVM	80.0	90.7	86.7	0.949	58.1	90.2	78.0	0.816
ExtFP-kNN	97.2	98.2	97.9	0.996	67.7	73.5	71.3	0.750
ExtFP-C4.5DT	85.1	94.2	90.8	0.951	58.1	82.4	73.2	0.718
ExtFP-NB	85.6	69.5	75.4	0.817	46.8	86.3	71.3	0.725
FP-SVM	88.8	92.1	90.9	0.968	56.5	92.2	78.7	0.814
FP-kNN	90.2	97.9	95.1	0.994	71.0	74.5	73.2	0.754
FP-C4.5DT	86.2	90.7	89.0	0.937	59.7	83.3	74.4	0.798
FP-NB	85.1	67.1	73.7	0.807	48.4	86.3	72.0	0.732
KRFPC-SVM	85.3	95.8	91.9	0.976	61.3	86.3	76.8	0.855
KRFPC-kNN	100.0	100.0	100.0	1.00	79.0	56.9	65.2	0.731
KRFPC-C4.5DT	54.8	90.7	77.5	0.782	59.7	76.5	70.1	0.691
KRFPC-NB	80.7	95.4	90.0	0.972	50.0	74.5	65.2	0.667
KRFP-SVM	84.3	95.9	91.7	0.972	64.5	89.2	79.9	0.857
KRFP-kNN	99.6	99.5	99.5	0.999	77.4	66.7	70.7	0.763
KRFP-C4.5DT	83.7	94.6	90.6	0.938	54.8	81.4	71.3	0.780
KRFP-NB	75.6	91.9	85.9	0.946	53.2	79.4	69.5	0.725

^aSVM: support vector machine. *k*-NN: *k*-nearest neighbors. C4.5DT: C4.5 decision tree. NB: naïve Bayes. MACCS: MACCS keys. FP4: substructure fingerprints. FP: CDK fingerprint. ExtFP: CDK extended fingerprint. PubChemFP: PubChem fingerprints. EStateFP: estate fingerprint. KRFP: Klekota–Roth fingerprint. KRFPC: Klekota–Roth fingerprint count. SE: sensitivity. SP: specificity. Q: the overall predictive accuracy. AUC: the area under the receiver operating characteristic curve.

Table 4. Performance of Seven Global Classification Models Built on 1604 Compounds Using the Seven Best Modeling Methods in 5-fold Cross-Validation^a

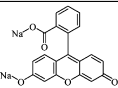
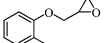

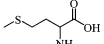
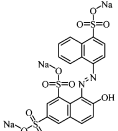
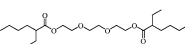
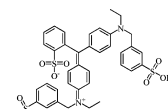
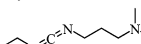
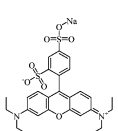
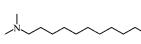

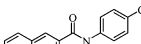
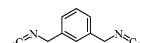

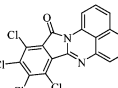
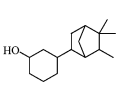
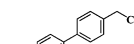
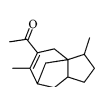

modeling methods	SE (%)	SP (%)	Q (%)	AUC
CART-NB	80.2	77.7	78.6	0.860
CHAID-SVM	86.0	87.3	83.8	0.900
GASVM-kNN	90.0	81.2	78.6	0.841
MACCS-SVM	84.1	93.7	89.5	0.947
PubChemFP-SVM	94.7	93.1	89.3	0.946
EStateFP-SVM	94.6	89.7	84.6	0.912
KRFP-SVM	91.2	96.0	91.2	0.969

^aCART: the classification and regression tree algorithm. CHAID: chi-squared automatic interaction detector. NB: naïve Bayes. SVM: support vector machine. *k*-NN: *k*-nearest neighbors. MACCS: MACCS keys. PubChemFP: PubChem fingerprints. EStateFP: estate fingerprints. KRFP: Klekota–Roth fingerprint. SE: sensitivity. SP: specificity. Q: the overall predictive accuracy. AUC: the area under the receiver operating characteristic curve.



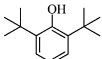


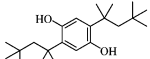
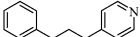

(Figure 3). The mean values of MW were 176.49 and 238.61 for RB and NRB chemicals, respectively, with a *p*-value of 7.8×10^{-18} . And the low negative linear correlation of -0.21 was observed between MW and BOD numerical values (Figure 4). RB chemicals are favorable for lower MW values. Environmental experts have demonstrated it for 100 years. The number of rings of chemicals indicated high significant difference of the mean nRing of RB and NRB chemicals as the lower *p*-value of 3.42×10^{-49} . The mean values of nRing were 0.53 and 1.36 for RB and NRB chemicals, respectively, and it is a moderately negative linear correlation ($R = -0.38$) between nRing and chemical BOD (Figure 4). This showed that a chemical with the higher numbers of ring is unfavorable for chemical biodegradability.

BCUTp-1h belongs to BCUT descriptors, which constitute a novel class of molecular descriptors encoding both substructure topological (or topographical) and atomic information relevant to the strength of ligand–receptor interactions. The mean value of BCUTp-1h was 7.70 and 9.54 for RB and NRB chemicals, respectively, with the lowest *p*-value of 9.97×10^{-82} . There is a good negative linear correlation ($R = -0.47$) between

Scheme 1. Detailed Predicted Results of the Consensus Global Probability Model and Experimental Results Using the OECD MITI Test Protocol for the External Validation Set of 27 Novel Chemicals^a

CAS RN	Structure	Indirect Analysis	Direct Analysis				Experimental Results	*Predicted Results ^b
		BOD [%]	TOC*1 [%]	UV*2 [%]	GC*2 [%]	HPLC*2 [%]		
518-47-8		0	0	-	-	0	NRB	0.311
2210-79-9		0	2	-	-	90	NRB	0.267
95-13-6		0	-	-	-	1	NRB	0.278
59-51-8		81	82	-	-	89	RB	0.556
2611-82-7		2	2	-	-	0	NRB	0.013
94-28-0		92	-	-	-	100	RB	0.750
2650-18-2		2	0	-	-	0	NRB	0.011
1892-57-5		0	4	-	-	0	NRB	0.111
3520-42-1		6	0	-	-	0	NRB	0.017
21542-96-1		36 35	- -	- -	66 >99	- -	RB	0.502
37609-25-9		66	-	-	-	92	RB	0.624
92-78-4		1	-	-	-	0	NRB	0.101
3634-83-1		0	-	-	-	>99	NRB	0.049
281-23-2		15	-	-	0	-	NRB	0.179
20749-68-2		0	-	-	-	1	NRB	0.020
3407-42-9		0	-	-	3	-	NRB	0.074
1667-10-3		0	-	-	-	3	NRB	0.147
32388-55-9		0	-	-	3	-	NRB	0.162
583-57-3		0	-	-	2	-	NRB	0.431

Scheme 1. continued

CAS RN	Structure	Indirect Analysis		Direct Analysis			Experimental Results	*Predicted Results ^b
		BOD [%]	TOC*1 [%]	UV*2 [%]	GC*2 [%]	HPLC*2 [%]		
98-06-6		0	-	-	27	-	NRB	0.242
571-58-4		0	-	-	-	2	NRB	0.245
128-39-2		0	-	-	-	11	NRB	0.031
88-16-4		0	-	-	0	-	NRB	0.045
2432-14-6		0	-	-	-	0	NRB	0.207
903-19-5		0	-	-	1	-	NRB	0.028
2057-49-0		4	-	-	23	-	NRB	0.188
355-80-6		0	-	-	0	-	NRB	0.131

^aRB: ready biodegradability. NRB: not ready biodegradability. BOD: biological oxygen demand. ^bIf a chemical was predicted with a probability greater than 0.5, this chemical is RB; otherwise, this chemical is NRB. The detailed description about the consensus model using the classic classifiers fusion techniques of the mean method and probability output can be found in our previously published work.²⁴

Table 5. Performance of the Global Classification Models Built on 1604 Compounds Using the Seven Best Modeling Methods for the External Validation Set (27 Novel Compounds)

modeling methods	TP	TN	FP	FN	SE (%)	SP (%)	Q (%)	AUC
CART-NB	3	20	3	1	75.0	87.0	85.2	0.957
CHAID-SVM	4	22	1	0	100.0	95.7	96.3	0.978
GASVM-kNN	1	19	4	3	25.0	82.6	74.1	0.603
MACCS-SVM	3	23	0	1	75.0	100.0	96.3	1.00
PubChemFP-SVM	4	23	0	0	100.0	100.0	100.0	1.00
EStateFP-SVM	3	21	2	1	75.0	91.3	88.9	0.891
KRFP-SVM	3	22	1	1	75.0	95.7	92.6	0.978
consensus model ^b	4	23	0	0	100.0	100.0	100.0	1.00
Biowin5 ^c	3	20	3	1	75.0	87.0	85.2	
Biowin6 ^c	2	21	2	2	50.0	91.3	85.2	

^aCART: the classification and regression tree algorithm. CHAID: chi-squared automatic interaction detector. GASVM: genetic algorithm and support vector machine (SVM). NB: naïve Bayes. *k*-NN: *k*-nearest neighbors. MACCS: MACCS keys. PubChemFP: PubChem fingerprints. EStateFP: estate fingerprints. KRFP: Klekota–Roth fingerprint. TP: true positives. TN: true negatives. FP: false positives. FN: false negatives. SE: sensitivity. SP: specificity. Q: the overall predictive accuracy. AUC: the area under the receiver operating characteristic curve. ^bConsensus model was fused by mean methods. ^cBiowin5 and Biowin6 are the linear and nonlinear MITI biodegradation models, respectively, published by Tunkel et al.,⁹ which had to be implemented in the software EPI Suite v4.10 (<http://www.epa.gov/oppt/exposure/pubs/episuite.htm>).

BCUTp-1h and BOD of 817 chemicals (Figure 4). The data indicated that higher BCUTp-1h value is unfavorable for chemical biodegradability. Hydrogen binding ability is commonly represented by nHBacc and nHBDOn. The *p*-value between the mean nHBacc and nHBDOn for the RB and NRB chemicals were 0.37 and 0.04, indicative of no or low significant difference. There is no linear correlation between nHBacc and nHBDOn and chemical BOD (Figure 4). The data shows that hydrogen binding ability is not a key factor for chemical biodegradation. Biodegradation is a complex chemical and biological process consisting of many steps, which depend

not only on the amount and structure of the chemical, but also on environmental conditions into which the chemical is released.⁴⁵ It is very difficult to explain the biodegradation mechanism using individual or several simple chemical descriptors.

Visualization Analysis of Substructure Alerts for Chemical Biodegradation. In order to visually explore the structural features of RB and NRB chemicals, some privileged substructures for RB and NRB chemicals were identified in Supporting Information Schemes S1 and S2 by combining information gain and substructure fragment analysis on the

Table 6. Performance of in Domain (ID) and out of Domain (OD) Chemicals in the Test Set for the Seven Best Combinatorial Classification Models after Applying Application Domain Assessment

methods	ID				OD			
	SE (%)	SP (%)	Q (%)	AUC	SE (%)	SP (%)	Q (%)	AUC
CART-NB	89.2	85.7	87.7	0.890	44.0	93.2	80.8	0.793
CHAID-SVM	86.2	86.5	85.7	0.904	79.3	24.0	93.2	0.768
GASVM-kNN	90.4	85.7	86.2	0.900	76.8	93.2	82.8	0.801
MACCS-SVM	90.0	75.7	92.9	0.899	80.1	36.0	97.3	0.704
PubChemFP-SVM	89.9	81.1	82.1	0.876	70.4	44.0	89.2	0.777
EStateFP-SVM	87.6	70.3	89.3	0.871	77.7	48.0	94.6	0.858
KRFP-SVM	87.1	78.4	89.3	0.897	85.8	44.0	89.2	0.786

^aSVM: support vector Machine. NB: naïve bayes. *k*-NN: *k*-nearest neighbors. C4.5DT: C4.5 decision tree. MACCS: MACCS keys. EStateFP: estate fingerprint. KRFP: Klekota–Roth fingerprint. CART: the classification and regression tree algorithm. CHAID: chi-squared automatic interaction detector. GASVM: genetic algorithm and support vector machine. SE: sensitivity. SP: specificity. Q: the overall predictive accuracy. AUC: the area under the receiver operating characteristic curve.

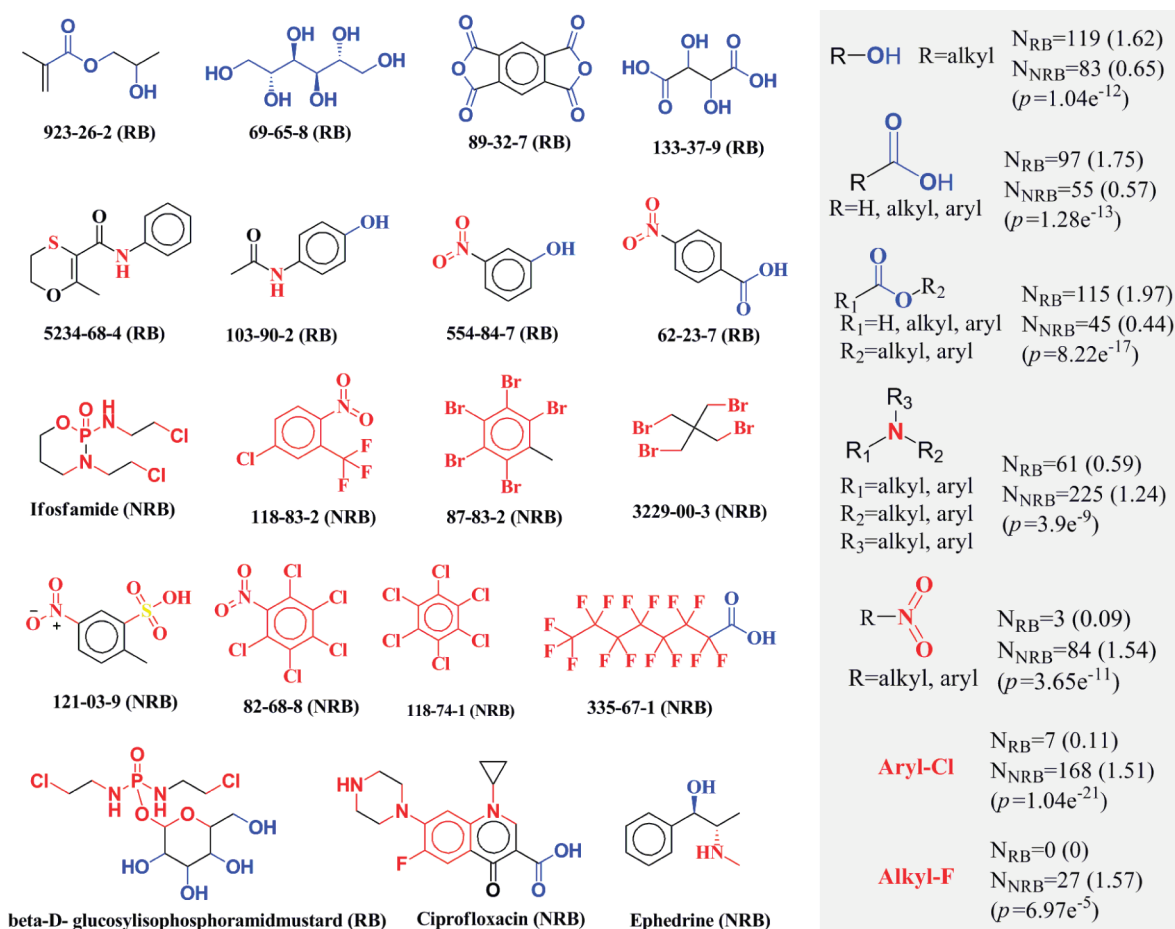


Figure 2. Some privileged and unprivileged substructures in example compounds are favorable for ready biodegradability (RB, depicted with blue) and not ready biodegradability (NRB, depicted with red). N_{RB} is the number of RB chemicals in RB class with specified fragment t . N_{NRB} is the number of NRB chemicals in NRB class with specified fragment t . The data in brackets represents the “frequency of a fragment” enrichment factor. p -value: Student’s t test was used to evaluate the significance of the difference between paired samples and the means.

entire 1631 compounds. To increase the useful implications for environmental experts, our discussion will be limited to the most representative structure alerts.

Alcohol. The alcohol group is a very common group presented in many organic chemicals. In the entire data set of 1631 compounds, 202 chemicals were presented alcohol group, of which 119 compounds are RB and 83 compounds are NRB with a p -value of 1.04×10^{-12} . We focused on NRB chemicals which influence the toxicity, persistence, and ultimate fate of an

organic chemical in the aquatic and terrestrial ecosystems. For chemicals, CAS 923-26-2, 69-65-8, 89-32-7, and 133-37-9, having an alcohol group are RB (Supporting Information Table S1). This is in agreement with the common perspective that chemicals including the group of primary alcohol are easy to degrade by an oxidation or conjugation reaction. Chemicals CAS 111-48-8, ciprofloxacin and ephedrine having an alcohol group and unfavorable group for RB at the same time, are NRB (Figure 2). When alcohol is considered a favorable group

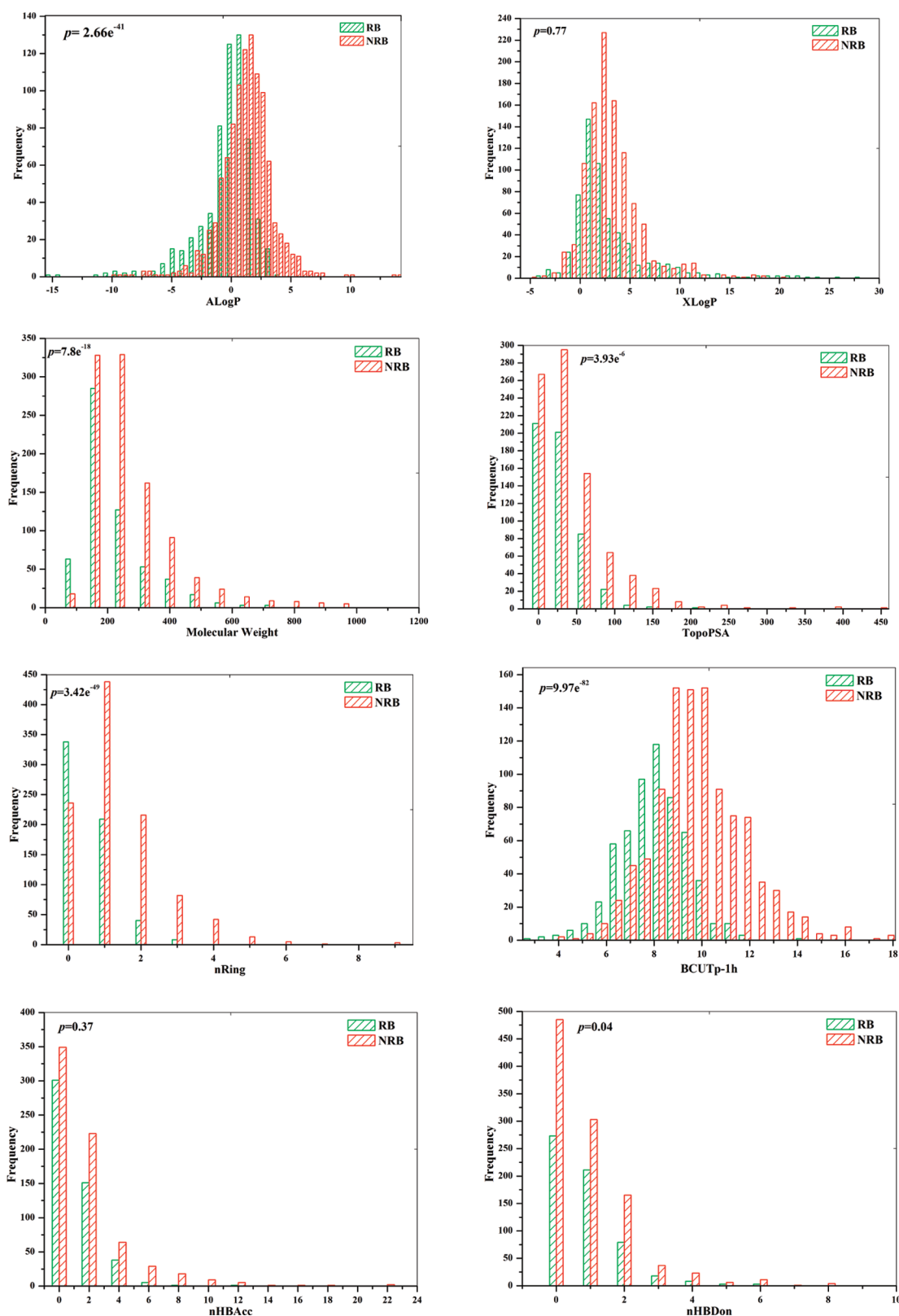


Figure 3. Distributions of eight chemical properties, including Ghose–Crippen LogKow (ALogP), XlogP, molecular weight (MW), topological polar surface area (TopoPSA), number of rings (nRing), nlow highest polarizability weighted BCUTS (BCUTp-1h), number of hydrogen bond acceptors (nHBAcc), and number of hydrogen bond donors (nHBDDon) for ready biodegradability (RB) and not ready biodegradability (NRB) classes. p -value: Student's t test was used to evaluate the significance of the difference between paired samples and the means.

without considering the other part of a compound, it often leads to a wrong prediction.

Amine. The amine group is present in 286 chemicals, of which 225 are NRB and only 61 chemicals are RB with a p -value of 3.9×10^{-9} . The detailed description is given in

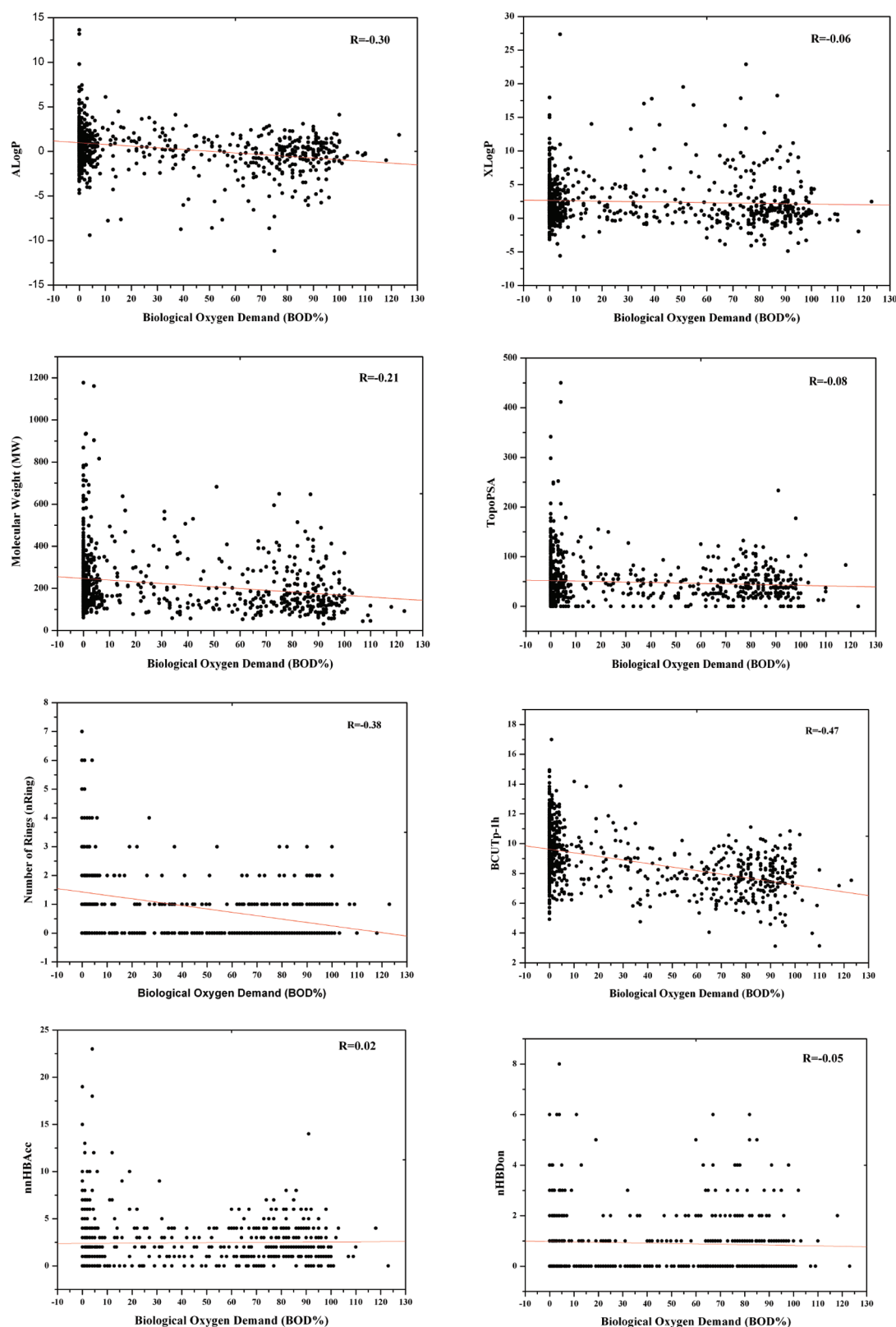


Figure 4. Correlations between eight representative chemical descriptors, including Ghose–Crippen LogKow (ALogP), XlogP, molecular weight (MW), topological polar surface area (TopoPSA), number of rings (nRing), nlow highest polarizability weighted BCUTs (BCUTp-1h), number of hydrogen bond acceptors (nHBacc), and number of hydrogen bond donors (nHBDon) versus biological oxygen demand (BOD%) of 817 compounds.

Supporting Information Scheme S1 and Table S5. The amine group contained several overlapping fragments, aliphatic amine, aromatic amine (primary_om_amine, secondary_om_

amine, and tertiary_om_amine), and mixed amine. Forty-seven chemicals contained Tertiary_aliph_amine group, of which 42 compounds are NRB and only 5 compounds (such as

CAS 7779-27-3) are RB with a p -value of 1.84×10^{-4} . Only two compounds having the aromatic amine and mixed amine group are RB (Supporting Information Scheme S1). The groups of aromatic amine and mixed amine are good structural alerts to alert NRB.

Nitro. The nitro group is present in 87 chemicals, of which 84 are NRB and only 3 chemicals are RB with a p -value of 3.65×10^{-11} . The CAS number of three RB compounds having nitro group are CAS 62-23-7 (*p*-nitrobenzoic acid), 554-84-7 (*m*-hydroxybenzoic acid), and 552-16-9 (*o*-nitrobenzoic acid) (Supporting Information Table S1). We found that the three RB chemicals have the nitro group and the groups of hydroxy and carboxylic acid at the same time. The groups of hydroxy and carboxylic acid are favorable for chemical RB. This indicated that if the favorable groups and structural alerts are present in a chemical at the same time, it is very difficult to identify RB or NRB chemicals.

Halogen. The halogen is a very common group presented in organic chemistry, and it is often considered a typical structural alert to chemical biodegradability. The detailed descriptions of alkyl and aromatic halogen were presented in Supporting Information Scheme S1. If a chemical has the group of alkylfluoride or arylfluoride, this chemical is NRB. This indicated that the fluoride is a good structural alert, which is in agreement with the common chemical knowledge that the fluoride bond is a strong bond and it is very difficult to be biodegraded. The alkylchloride group is present in 90 chemicals, of which 72 are NRB and only 18 chemicals are RB with a p -value of 8.24×10^{-4} . The arylchloride group is present in 175 chemicals, of which 168 are NRB and only 7 chemicals are RB with a low p -value of 1.04×10^{-21} . When the seven RB chemicals are double checked, such as CAS 17639-93-9, 79-43-6, 2549-51-1, etc., these also have the groups of carboxylic acid or carboxylic ether (Supporting Information Table S1). Moreover, some more specific substructure alerts including chloride atoms were identified in Supporting Information Scheme S2.

Carboxylic Acid and Carboxylic Ether. The carboxylic acid group is presented in 152 chemicals, of which 55 are NRB and 97 are RB with a p -value of 1.28×10^{-13} . The carboxylic ether is present in 160 chemicals, of which 45 are NRB and 115 are RB with a p -value of 8.22×10^{-17} . This is in agreement with the common perspectives that carboxylic ester is easy to biodegrade by hydrolysis reaction or carboxylic acid is easy to degrade by oxidation or conjugation reaction. When the NRB chemicals are double checked, we found that the NRB chemicals having carboxylic acid or carboxylic ether groups also have the structural alerts of alkylfluoride at the same time, such as chemicals CAS 307-55-1, 376-06-7, 1976-9-3, 526-78-3, 20679-58-7, etc. (Supporting Information Table S1). It is confirmed that if the favorable groups for RB and structural alerts are present in a chemical at the same time, it is very difficult to identify RB or NRB. Our groups are active to determine the relationship between chemical structural feature and biodegradability, in order to establish additional specific structural alert information.

Although substructure fragment analysis can identify some structure alerts or privileged substructures for RB or NRB chemicals, they cannot characterize the spatial arrangement of these privileged fragments, or how to characterize if multiple privileged fragments or structural alerts are found simultaneously in the same chemical. For example, ciprofloxacin, ephedrine, and beta-D-glucosylisophosphoramidmustard have

the privileged and unprivileged substructures at the same time (Figure 2). The chemical of beta-D-glucosylisophosphoramidmustard is RB, but the drugs of ciprofloxacin and ephedrine are NRB. Nonetheless, these meaningful substructure fragments or structural alerts identified in this publication could potentially provide some visual alert function in the field of environmental hazard assessment.

CONCLUSION

In summary, our modeling study exceeds in four respects: (i) Four different modeling methods and four different feature reduction methods were systemically used to develop the best combinatorial classification probability models of RB versus NRB chemicals based on the largest MITI data set published so far. (ii) All models with high predictive accuracy were validated not only by a diverse test set, but also by the laboratory biodegradation experiments. (iii) The significance of AD in this study confirmed that the use of AD improved the prediction accuracy of models. (iv) The information gain and substructure fragment analysis were combined to extract the privileged fragments for ready versus not ready biodegradability chemicals. Several identified privileged substructures or structural alerts improved the interpretation of models for environmental assessment experts.

Models developed in this study and modeling methodologies combined with the MITI test protocol will provide critical information and useful tools for designing new biodegradable chemicals or predicting the biodegradability of new chemicals in the environmental hazard assessment.

ASSOCIATED CONTENT

Supporting Information

Additional details on materials and methods used and supplemental tables (Tables S1–S5) and schemes (Schemes S1 and S2). This material is available free of charge via the Internet at <http://pubs.acs.org>.

AUTHOR INFORMATION

Corresponding Author

*Tel.: +86-21-64251052. Fax: +86-21-64253651. E-mail address: philiplee2007@gmail.com (P.L.); ytang234@ecust.edu.cn (Y.T.).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Program for New Century Excellent Talents in University (Grant NCET-08-0774), the National Natural Science Foundation of China (Grant 21072059), the 111 Project (Grant B07023), the Shanghai Committee of Science and Technology (11DZ2260600), and the Fundamental Research Funds for the Central Universities (WY1113007).

REFERENCES

- (1) 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.
- (2) Howard, P. H.; Muir, D. C. Identifying new persistent and bioaccumulative organics among chemicals in commerce. *Environ. Sci. Technol.* **2010**, *44*, 2277–2285.

- (3) Howard, P. H.; Muir, D. C. Identifying new persistent and bioaccumulative organics among chemicals in commerce II: pharmaceuticals. *Environ. Sci. Technol.* **2011**, *45*, 6938–6946.
- (4) Rorije, E.; Loonen, H.; Muller, M.; Klopman, G.; Peijnenburg, W. J. Evaluation and application of models for the prediction of ready biodegradability in the MITI-I test. *Chemosphere*. **1999**, *38*, 1409–1417.
- (5) Rusyn, I.; Daston, G. P. Computational toxicology: realizing the promise of the toxicity testing in the 21st century. *Environ. Health Perspect.* **2010**, *118*, 1047–1050.
- (6) Cuissart, B.; Touffet, F.; Cremilleux, B.; Bureau, R.; Rault, S. The maximum common substructure as a molecular depiction in a supervised classification context: experiments in quantitative structure/biodegradability relationships. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1043–1052.
- (7) Andreini, C.; Bertini, I.; Cavallaro, G.; Decaria, L.; Rosato, A. A simple protocol for the comparative analysis of the structure and occurrence of biochemical pathways across superkingdoms. *J. Chem. Inf. Model.* **2010**, *51*, 730–738.
- (8) Howard, P. H.; Boethling, R. S.; Stiteler, W. M.; Meylan, W. M.; Hueber, A. E.; Beaman, 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.
- (9) 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.
- (10) Hiromatsu, K.; Yakabe, Y.; Katagiri, K.; Nishihara, T. Prediction for biodegradability of chemicals by an empirical flowchart. *Chemosphere*. **2000**, *41*, 1749–1754.
- (11) Philipp, B.; Hoff, M.; Germa, F.; Schink, B.; Beimbom, D.; Mersch-Sundermann, V. Biochemical interpretation of quantitative structure-activity relationships (QSAR) for biodegradation of N-heterocycles: a complementary approach to predict biodegradability. *Environ. Sci. Technol.* **2007**, *41*, 1390–1398.
- (12) Hou, B. K.; Wackett, L. P.; Ellis, L. B. Microbial pathway prediction: a functional group approach. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1051–1057.
- (13) DeLisle, R. K.; Dixon, S. L. Induction of decision trees via evolutionary programming. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 862–870.
- (14) Jaworska, J. S.; Boethling, R. S.; Howard, P. H. Recent developments in broadly applicable structure-biodegradability relationships. *Environ. Toxicol. Chem.* **2003**, *22*, 1710–1723.
- (15) Horton, D. A.; Bourne, G. T.; Smythe, M. L. The combinatorial synthesis of bicyclic privileged structures or privileged substructures. *Chem. Rev.* **2003**, *103*, 893–930.
- (16) OECD 301. Guideline for Testing of Chemicals. *Ready Biodegradability*. OECD: Paris, 1992.
- (17) Ericson, J. F. Evaluation of the OECD 314B activated sludge die-away test for assessing the biodegradation of pharmaceuticals. *Environ. Sci. Technol.* **2010**, *44*, 375–381.
- (18) Judson, R.; Richard, A.; Dix, D.; Houck, K.; Elloumi, F.; Martin, M.; Cathey, T.; Transue, T. R.; Spencer, R.; Wolf, M. ACToR—Aggregated Computational Toxicology Resource. *Toxicol. Appl. Pharmacol.* **2008**, *233*, 7–13.
- (19) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623–633.
- (20) Boethling, R. S.; Lynch, D. G.; Jaworska, J. S.; Tunkel, J. L.; Thom, G. C.; Webb, S. Using Biowin, Bayes, and batteries to predict ready biodegradability. *Environ. Toxicol. Chem.* **2004**, *23*, 911–920.
- (21) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466–1474.
- (22) Breiman, L. *Classification and regression trees*, 1st ed.; Chapman & Hall/CRC: Boca Raton, 1984.
- (23) Sonquist, J. A.; Morgan, J. N. *The detection of Interaction Effects; Survey research center*; University of Michigan: Ann Arbor, 1964; p 296.
- (24) Cheng, F.; Yu, Y.; Shen, J.; Yang, L.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. Classification of Cytochrome P450 Inhibitors and non-Inhibitors using Combined Classifiers. *J. Chem. Inf. Model.* **2011**, *51*, 996–1011.
- (25) Shen, J.; Du, Y.; Zhao, Y.; Liu, G.; Tang, Y. In silico prediction of blood-brain partitioning using a chemometric method called genetic algorithm based variable selection. *QSAR Comb. Sci.* **2008**, *72*, 635–645.
- (26) Klekota, J.; Roth, F. P. Chemical substructures that enrich for biological activity. *Bioinformatics.* **2008**, *24*, 2518–2525.
- (27) Chang, C. C.; Lin, C.-J. LIBSVM: a library for support vector machines. <http://www.csie.ntu.edu.tw/~cjlin/libsvm> (accessed Jan 18, 2010).
- (28) Tropsha, A.; Golbraikh, A. Predictive QSAR modeling workflow, model applicability domains, and virtual screening. *Curr. Pharm. Des.* **2007**, *13*, 3494–3504.
- (29) Zhu, H.; Rusyn, I.; Richard, A.; Tropsha, A. Use of cell viability assay data improves the prediction accuracy of conventional quantitative structure-activity relationship models of animal carcinogenicity. *Environ. Health Perspect.* **2008**, *116*, 506–513.
- (30) Rodgers, A. D.; Zhu, H.; Fourches, D.; Rusyn, I.; Tropsha, A. Modeling liver-related adverse effects of drugs using knearest neighbor quantitative structure-activity relationship method. *Chem. Res. Toxicol.* **2010**, *23*, 724–732.
- (31) Corinna, C.; Vladimir, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.
- (32) Quinlan, J. R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, 1993.
- (33) Cover, T. M.; Hart, P. E. Nearest neighbor pattern classification. *IEEE. T. Inform. Theory.* **1967**, *13*, 21–27.
- (34) Watson, P. Naive Bayes classification using 2D pharmacophore feature triplet vectors. *J. Chem. Inf. Model.* **2008**, *48*, 166–178.
- (35) Shen, J.; Cheng, F.; Xu, Y.; Li, W.; Tang, Y. Estimation of ADME properties with substructure pattern recognition. *J. Chem. Inf. Model.* **2010**, *50*, 1034–1041.
- (36) Evans, B. E.; Rittle, K. E.; Bock, M. G.; DiPardo, R. M.; Freidinger, R. M.; Whitter, W. L.; Lundell, G. F.; Veber, D. F.; Anderson, P. S.; Chang, R. S.; et al. Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem.* **1988**, *31*, 2235–2246.
- (37) Jensen, B. F.; Vind, C.; Padkjaer, S. B.; Brockhoff, P. B.; Refsgaard, H. H. In silico prediction of cytochrome P450 2D6 and 3A4 inhibition using Gaussian kernel weighted k-nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors. *J. Med. Chem.* **2007**, *50*, 501–511.
- (38) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics.* **2000**, *16*, 412–424.
- (39) Weaver, S.; Gleeson, M. P. The importance of the domain of applicability in QSAR modeling. *J. Mol. Graphics Modell.* **2008**, *26*, 1315–1326.
- (40) Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical assessment of QSAR models of environmental toxicity against *Tetrahymena pyriformis*: focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733–1746.
- (41) Sushko, I.; Novotarskyi, S.; Korner, R.; Pandey, A. K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Muller, K. R.; Xi, L.; Liu, H.; Yao, X.; Oberg, T.; Hormozdiari, F.; Dao, P.; Sahinalp, C.; Todeschini, R.; Polishchuk, P.; Artemenko, A.; Kuz'min, V.; Martin, T. M.; Young, D. M.; Fourches, D.; Muratov, E.; Tropsha, A.; Baskin, I.; Horvath, D.; Marcou, G.; Muller, C.; Varnek, A.; Prokopenko, V. V.; Tetko, I. V. Applicability domains for classification problems: Benchmarking of distance to models for Ames mutagenicity set. *J. Chem. Inf. Model.* **2010**, *50*, 2094–2111.
- (42) Cheng, F.; Shen, J.; Yu, Y.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. In silico prediction of *Tetrahymena pyriformis* toxicity for diverse

industrial chemicals with substructure pattern recognition and machine learning methods. *Chemosphere*. **2011**, 82, 1636–1643.

(43) Cheng, F.; Yu, Y.; Zhou, Y.; Shen, Z.; Xiao, W.; Liu, G.; Li, W.; Lee, P. W.; Tang, Y. Insights into molecular basis of cytochrome p450 inhibitory promiscuity of compounds. *J. Chem. Inf. Model.* **2011**, 51, 2482–2495.

(44) Hao, R.; Li, J.; Zhou, Y.; Cheng, S.; Zhang, Y. Structure-biodegradability relationship of nonylphenol isomers during biological wastewater treatment process. *Chemosphere* **2009**, 75, 987–994.

(45) Pavan, M.; Worth, A. P. Review of Estimation Models for Biodegradation. *QSAR Comb. Sci.* **2008**, 27, 32–40.