Journal of Molecular Graphics and Modelling 27 (2008) 59–65

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Screening of persistent organic pollutants by QSPR classification models: A comparative study

Ester Papa *, Paola Gramatica

Department of Structural and Functional Biology, QSAR Research Unit in Environmental Chemistry and Ecotoxicology, University of Insubria, via Dunant 3, 21100 Varese, Italy

ARTICLE INFO

Article history:
Received 7 January 2008
Received in revised form 20 February 2008
Accepted 21 February 2008
Available online 4 March 2008

Keywords:
Quantitative structure–property relationships (QSPRs)
POP
Classification
Half-life
Persistence

ABSTRACT

A Quantitative Structure–Property Relationships (QSPRs) study for the prediction of the environmental persistence of a set of 250 heterogeneous organic compounds is here presented. Three *a priori* defined classes of environmental persistence were generated, by Hierarchical Cluster Analysis, from the combination of half-life data in air, water, soil and sediment available for all the studied compounds. QSPR classification models were successfully developed using different techniques (k-NN, CART and CP-ANN) and three interpretable theoretical molecular descriptors. Robust external validation was provided by statistical splitting and also on completely new data. The good performances of all these models were compared and their structural domains were analyzed. The analysis of the errors highlights a slight tendency of persistence overestimation, misclassifying chemicals from a lower to a higher class of persistence, in line with the precautionary principle. Finally, the reliability of the proposed QSPR models was verified further with new data from the literature. The structure-based classification models, applicable for the prediction of potential persistence of heterogeneous organic compounds, could be useful as preliminary support tools for the identification and prioritization of new potential POPs among already existing chemicals as well as "screening prior to synthesis" procedures to avoid the production, and consequent release into the environment, of new POPs.

1. Introduction

The persistence of organic compounds, is governed by the rate at which such compounds are removed by physical, chemical and biological processes, and is dependent on the environmental compartment. This property is a crucial issue when describing the environmental fate of pollutants. Persistent organic pollutants (POPs) are chemical substances which long-term persistence could lead to accumulation in the environment and biota and thus to effects due to chronic exposure. With today's evidence of the long-range transport (LRT) of such substances to regions where they have never been used or produced, and the consequent environmental threat they pose worldwide, the international community has called, on several occasions, for urgent global action to reduce and eliminate the release of such chemical substances [1,2]. The need to identify potentially persistent compounds is also strictly connected to the prioritization and screening of persistent, bioaccumulative, and toxic (PBT) chemicals, which use and production is being subjected to worldwide regulation [2–4].

The half-life of organic pollutants in various compartments is among the most commonly used criteria for studying persistence [2,5], and is regarded as a key quantity in the assessment of ecological and human health risk. Unfortunately practical efforts towards determining effective half-life have been hindered by serious difficulties as they are dependent on environmental conditions and laboratory tests and thus reliable data are available for a limited amount of organic compounds. Screening criteria that allow the priority ranking of substances needing a more detailed assessment are required and different approaches and models have been developed for the screening and the prediction of the environmental behavior of potential POPs [5–15].

Recent papers discuss and compare different multimedia models [11–13] which have been recommended by Klasmeier et al. [13] as being a more efficient alternative to the single media half-life approach proposed by the UNEP Stockholm Convention [2]. Multimedia models are indeed invaluable tools for chemicals with known experimental data (half-life data, partitioning properties, environmental conditions), but there is still a need for complementary methods for simpler and earlier identification of

* Corresponding author. Tel.: +39 0332 421552; fax: +39 0332 421554.
E-mail address: ester.papa@uninsubria.it (E. Papa).

1093-3263/$ – see front matter © 2008 Elsevier Inc.. All rights reserved.
doi:10.1016/j.jmgm.2008.02.004

new POPs, even without experimental data, and for directing the synthesis of safer alternatives to POPs, as requested also by the Stockholm Convention [2]. With regard to these topics, different Quantitative Structure–Property Relationships (QSPR)-based approaches have recently been proposed to study the environmental fate of chemicals [14–25]. In a recent study [15] we developed a simple QSPR regression model for the prediction of global persistence of a set of heterogeneous organic compounds; the comparability of the results, obtained by our multivariate approach, to the multimedia models, and the reliability of the predictions obtained by our proposed QSPR model were demonstrated.

In many practical applications for prioritizing compounds or for evaluating possible POP behavior in new chemicals produced by the chemical industry, it could be sufficient to have a simple classification scheme that divides compounds into classes of persistence by means of QSPR models of classification.

The first goal of this paper was to develop validated and robust QSPR classification models for the prediction of the POP-like behavior of heterogeneous chemicals in the environment. This aim was achieved first by identifying three classes of environmental persistence by Hierarchical Cluster Analysis, starting from half-life data in different environmental compartments. Then we developed robust and externally validated QSPR classification models to predict the persistence of heterogeneous chemicals, bearing in mind the demonstrated relevance of molecular structure in determining the intrinsic tendency of a molecule to be persistent in the environment [15].

This paper has the final aim to propose a different QSPR-based approach, already presented with preliminary results at international scientific meetings [24,25], as a tool for the identification and prioritization of existing or not yet synthesized potential POPs. This approach, which allows the prediction of the tendency to persistence of chemicals on structural basis only, can be particularly useful when other empiric approaches cannot be applied (i.e. UNEP single media approach, multimedia models) due to missing experimental data.

2. Methods

2.1. Data set

In this study a selection of overall half-life data for transformation in 4 environmental media (air, water, soil and sediment) [15,26] for 250 organic compounds, was used as data set to perform the here proposed QSPR approach. Even though these data are semi-quantitative and are based on expert judgement and actual experimental values, they have already been suggested by Webster et al. [6] as preferable for half life identification, particularly for screening purposes, and they are commonly used to run and develop the widely applied multimedia models. Due to the wide half-life range in the four selected media (from 5 to 55,000 h), we transformed the original data into logarithmic values. The studied dataset was representative of 18 chemical classes (data set reported in Table SI1) and included some of the most relevant environmental pollutants such as chlorobenzenes, polychlorobiphenyls (PCBs), polychlorodibenzodioxins (PCDD), polychlorodibenzofurans (PCDF), polycyclic aromatic hydrocarbons (PAHs), various pesticides, as well as heterogeneous industrial chemicals.

Empirical information on persistence for 10 reference chemicals used by Klasmeier et al. [13] (listed in Table SI1), as well as for 43 organic pollutants chosen from the UNECE [1,27] UNEP [2], EEB [28] priority lists (listed in Table SI4), was considered to further on check the external predictivity of our models.

2.2. Determination of persistency classes by Hierarchical Cluster Analysis

The Hierarchical Cluster Analysis method is applied to find clusters of chemicals in a multidimensional space, based on inter-object distances calculated, in this study, on the basis of the available half-life data. An agglomerative algorithm, starting from a situation in which each object is in its own cluster, defines clusters. The closest objects are then, step-by-step, joined together until at the end all objects are in one cluster. A binary tree called dendrogram can represent the hierarchy of clusters, which final number is obtained by cutting the tree at an arbitrary specified level. This multivariate analysis was here applied on the half-life data in four environmental media, by the SCAN program [29], for the definition of *a priori* classes of global environmental persistence. The Manhattan distance metric and the average linkage method were chosen as the best combination, among those available in the software [29], to perform the analysis on autoscaled half-life values. The term global persistence specifies that these classes are derived from a combination of half-life data in different compartments: air, water, soil, and sediment. This approach groups chemicals with a similar tendency for being persistent, into three *a priori* classes, Class 1 (HP, very high and high persistent compounds), Class 2 (MP, medium persistent compounds), and Class 3 (LP, low persistent compounds). These classes of global persistence where then used as the endpoint in the following QSPR modeling.

2.3. Descriptor calculation

A set of 662 theoretical molecular descriptors (zero-, mono-, and bi-dimensional), used as input for QSPR modeling, was computed by the software *DRAGON* [30]. The files for descriptor calculation, which contain information on atom and bond types, connectivity, and atomic spatial coordinates, were obtained with the software *HYPERCHEM* [31]. One of the advantages in using simple descriptors, which do not require the application of quantum mechanic methods for their calculation, is that they can be derived from the 2D structures of the chemicals or also from their SMILES code. The typologies of the calculated descriptors are summarized in Table SI2 of Supporting Information.

2.4. Selection of training set and prediction set for external validation

In order to obtain compounds for external validation [32,33], the available data set was split in a training set for the development of the models, and in a prediction set, which was never included during their development, and used to validate the external predictivity of the QSARs. The most significant principal components, calculated from each group of *DRAGON* molecular descriptors, were used to describe the relevant structural information of the chemicals in the splitting procedure. The splitting of the data set was realized by Kohonen Artificial Neural Network (K-ANN) [34,35] (also called Self-Organizing Maps—SOM) using the package KOALA [36]. The Kohonen network is based on a single layer of hidden neurons arranged in a box exhibiting a two-dimensional plane of response on its top. This plane is usually called "top map". The neurons (i.e. cells in the map) are vectors of weights, corresponding to the input variables (i.e. molecular descriptors). The map has a toroid geometry: each neuron has the same number of neighbors, including the neurons on the borders of the top map. The Kohonen ANN automatically adapts itself in such a way that similar input objects (chemicals) fall within the same neuron in the top map. The selection of the training set chemicals was performed by selecting the chemicals with the minimal distance from the centroid of each cell in the top map (architecture of the map: 8×8

neurons, 300 iterations and 48 variables). In line with this procedure, and in order to provide a robust external statistical validation of the models, the data set of 250 compounds was split as follows: 72 objects in the training set (TSET; 19 compounds of Class 1; 34 compounds of Class 2; 19 compounds of Class 3) and 178 objects in the prediction set (PSET; 43 compounds of Class 1; 84 compounds of Class 2; 51 compounds of Class 3).

2.5. QSPR modeling

Classification models are quantitative tools based on relationships between one or more independent variables (here the molecular descriptors) and a categorical response variable of integer numerical values, each representing the class of the corresponding sample (here the three classes of high, medium and low persistence).

For new compounds, where the class is unknown, the classification model predicts the assignment to one of the *a priori* defined class.

The classification methods used, to predict the classes of global persistence of the studied chemicals, are the k-Nearest Neighbour (k-NN), the Classification and Regression Tree (CART), and the Counter Propagation-Artificial Neural Networks (CP-ANN).

A variable selection strategy, the Genetic Algorithm (GA) [37,38], was implemented for the k-NN method in the home-made software MobyDigs/Evolution developed by the Todeschini group (Milano-Bicocca University (Italy)). The application of the GA allows the generation of a population of models of different dimensionality, in decreasing order of predictive performance and model size, here ranging from one to three variables.

The values of the GA parameters, which can be modified to run this software for the classification method k-NN, used to develop our models, were: maximum k value = 10; population size for the evolution = 100; trade-off between cross-over and mutation (0–1) = 0.5, and selection bias % = 100 [38].

The best combination of modeling variables was then extracted by maximizing the overall percentage of correct assignments (Non-Error Rate) within the population of k-NN models.

These molecular descriptors were then also applied to generate CART and CP-ANN models, so the performance of the three models were compared on the basis of the same structural domain.

k-NN [39] is a classification method searching for the k nearest neighbours of each object in the data set; the basic idea is to identify the class of a compound based on the class of the k most similar compounds (five in this study), where the similarity is defined by calculating the Euclidean distances between the descriptor vectors. The k-NN method was here applied to autoscaled data and the *a priori* probability to belong to a class was set as proportional to the number of chemicals (objects) in the three *a priori* classes of persistence; the predictive power of the model was checked for k values between 1 and 10.

CART is a nonparametric unbiased classification strategy [40] normally based on automatic stepwise variable selection. As a final result, CART displays a binary, immediately applicable, classification tree: each nonterminal node corresponds to a discriminant variable (i.e. a molecular descriptor with its threshold value), and each terminal node corresponds to a single class (i.e. persistency). The CART model was developed using the SCAN program [29] with the splitting criterion of Gini [40], and the prior proportional class option.

CP-ANN is an adaptation of the Kohonen network to solve supervised problems (i.e. known *a priori* classes). The Kohonen layer gets input variables (the molecular descriptors) for the studied chemicals (objects). During the learning process of the network, the output values (i.e. the classes of persistence) are given to the output layer, which has the same topological arrangement as the Kohonen layer. During the learning the position of objects is projected to the output layer and the weights of the variables are corrected in such a way that they fit the output values of corresponding objects. The trained network can be used for predictions: a new object will be situated (in the Kohonen layer) on the neuron with the most similar weights. This position is then projected to the output layer, providing a predicted output value. A detailed description of CP-ANN architecture and learning strategy is given in the literature [41,42].

The architecture of the CP-ANN used here [36] was: 6×6 neurons, 400 iterations, 3 input variables (molecular descriptors) and 1 output layer (*a priori* class assignment for each studied compound).

The overall percentage of correct assignments, and the percentage of correct assignments for each different class were always used to check the internal (calculated for the TSET) and external (calculated for the PSET) performance of all the classification models.

2.6. Structural applicability domain

The structural applicability domain of the classification models was obtained by visualizing the studied compounds in a three-dimensional space representing the structural information encoded in the molecular descriptors. A further analysis of the applicability domain is also reported as Supporting Information (Table SI3).

3. Results

3.1. Categorization of POPs by Hierarchical Cluster Analysis

The logarithms of the half-life data for the 250 compounds in four environmental compartments [26], were combined by Hierarchical Cluster Analysis using the average linkage and the Manhattan distance metric [29], in order to group similarly persistent chemicals.

Three clusters were identified in the dendrogram, grouping compounds according to their cumulative half-lives (different classes in Fig. 1): Class 1 (high persistence), Class 2 (medium persistence), Class 3 (low persistence).

These three *a priori* defined classes were used as the endpoint to be modeled for the development of structure-based QSPR classification models.

3.2. QSPRs for persistence classification

Three different classification methods, k-NN, CART and CP-ANN, were applied to model the *a priori* defined classes of global persistence. The results of the modeling are reported in Table 1.

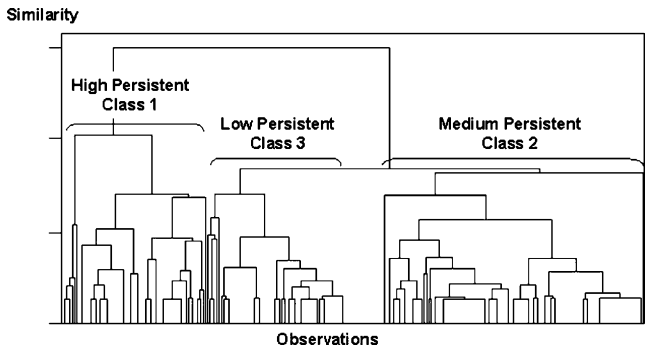

Fig. 1. Hierarchical Cluster Analysis on half-life data for 250 organic compounds in the various compartments (air, water, sediment and soil).

Table 1
Results for training and prediction set compounds (TSET and PSET, respectively) using the three classification methods (k-NN, CART and CP-ANN)

	a priori class	Assigned class			% correct	Overall % correct
		1	2	3		
1-A k-NN						
TSET	1	18	1	0	95	81
	2	3	26	5	76	
	3	0	5	14	74	
PSET	1	38	5	0	88	83
	2	7	65	12	77	
	3	0	7	44	86	
1-B CART						
TSET	1	18	1	0	95	83
	2	3	29	2	85	
	3	0	6	13	68	
PSET	1	37	4	2	86	75
	2	9	62	13	74	
	3	1	15	35	69	
1-C CP-ANN						
TSET	1	19	0	0	100	86
	2	1	28	5	82	
	3	0	4	15	79	
PSET	1	39	4	0	91	85
	2	6	68	10	81	
	3	0	6	45	88	

The classification results are reported as number of assignments for each class (compared to the *a priori* classes), percentage of correct assignments to each class, and total percentage of correct assignments in TSET and PSET.

The best k-NN model was chosen from the population of models by maximizing the accuracy (highest number of correct assignments in each class) and the model interpretability. The selected descriptors in this model by the Genetic Algorithm [37,38] are: molecular weight (MW), mean polarizability (Mp), and maximal electrotopological positive variation (MAXDP) [43]. The optimal number of *k* neighbours is 5.

The results of the k-NN model, calculated for the TSET and the PSET, are reported in Table 1-A. The model presents high percentages of correct overall assignments calculated for the TSET and the PSET (81% and 83%, respectively), as well as good accuracy in the single classes, ranging from 74% to 95% in the TSET, and from 77% to 88% in the PSET.

The second classification strategy used to predict the classes of environmental persistence was the Classification and Regression Tree (CART) method. The output of the CART model is a simple three nodes-classification tree (Fig. 2).

Also in this case the model presents good overall accuracy (Table 1-B) ranging from 75% (PSET) to 83% (TSET). More in detail, the accuracy is very good for high and medium persistent compounds (Class 1 TSET 95%, PSET 86%; Class 2 TSET 85%, PSET 74%), while it is lower, but still satisfying, for low persistent chemicals.

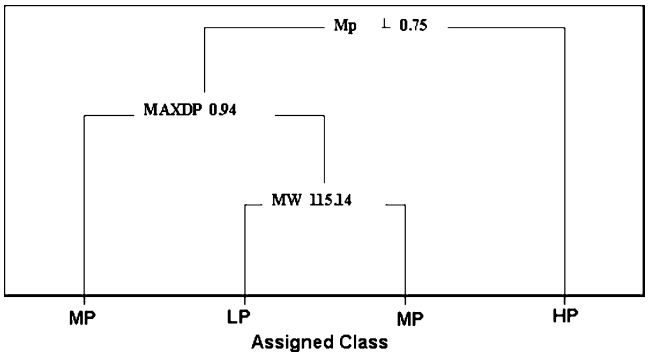

Fig. 2. Classification tree from the CART model.

Finally, the CP-ANN classification strategy gave the best classification results as reported in Table 1-C. The overall accuracy of this model is always high in both the TSET (86%) and the PSET (85%) reflecting the very good performances, never lower than 79%, obtained for the single classes of persistence.

4. Discussion

In this study three, *a priori* defined, classes of global persistence, identified by Hierarchical Cluster Analysis (Fig. 1), were modeled by applying different classification methods (k-NN CART and CP-ANN). These classes, defined on the basis of data of persistence available for the four media air, water, soil, and sediment, are in good agreement with the UNEP [2] classification criteria for persistence (air >2 days, surface water >60 days, soil and sediment >180 days). Thus, chemicals with high half-life values in all the media, or in three out of four media (values always exceeding the UNEP criteria), are grouped in the same cluster and assigned to the *a priori* Class 1 (High Persistence-HP); chemicals with a low global half-life fall in the central cluster (Low Persistence-LP; Class 3), not being persistent in any medium or being persistent in only one. Chemicals persistent from 1 to 8 months in one or more media are grouped in the cluster including medium persistent (MP) compounds, the *a priori* Class 2.

All the classification models were obtained using the same combination of descriptors selected by the Genetic Algorithm in the k-NN classification procedure on a TSET of 72 chemicals. The models were then strongly externally validated on a large prediction set (PSET), obtained by SOM splitting and including 70% (178 chemicals) of the original dataset of 250 compounds.

As it is evident from Table 1 all the models present always high performances which confirm the robustness and the high predictivity of these QSARs even when their external predictivity was checked on a large external prediction set. It is important to note that these classification results, with accuracy ranging from 100% to about 70%, are 25–50% higher than results existing in literature for similar studies [23].

Among the three presented strategies, the CP-ANN model gave the best and most balanced results, considering its internal (overall accuracy 86%) and external (overall accuracy 85%) performances as well as its robustness and accuracy in classifying individual classes. As clearly shown in Table 1-C, the model correctly assigns most of the studied compounds with high accuracy in all three classes (percentage of correct assignment ranges from 79% to 100% in TSET and from 81% to 91% in PSET). The high classification performance obtained for the PSET highlights the stability and the strong predictive power of the model.

The CART model gave the most unbalanced results when internal (83%) and external (75%) classification performance are compared, as well as among the three classes. However the simplicity of this approach, which anyway gives satisfactory results also for the PSET, represents an advantage when CART applicability is compared to the other classification techniques presented in this study. As shown in Fig. 2, the model consists of a simple binary tree, which classifies the compounds in the three classes on the basis of cut-off values defining the CART algorithm. The first node alone is able to classify all the highly persistent chemicals (Class 1). The discriminant variable Mp assigns compounds with a value >0.75 (high polarizability) to this class, while the remaining compounds need further structural information to be classified.

The discriminant variable in the second node is MAXDP, which classifies the more apolar chemicals of medium persistence, mainly

hydrocarbons, as belonging to Class 2 (MAXDP threshold value $\leq$0.94). Medium persistent compounds with MAXDP value greater than the threshold value (i.e. more heterogeneous and polar due to the presence of heteroatoms, mainly halogens), need more dimensional information to be separated from the low persistent chemicals of Class 3. At the last node the threshold value for the variable MW, splits the remaining compounds between Class 2 (MW > 115.4) and Class 3. The need for these two sequential nodes and of two different molecular descriptors highlights some difficulty to separate Class 2 and Class 3 compounds. This could be due to high structural heterogeneity in the dataset combined to a partial structural overlapping of medium and low persistent compounds (Class 2 and Class 3, respectively).

In fact, even if this model has similar overall internal (TSET) percentage of classification as the k-NN model (83%), it has the lowest accuracy in Class 3 assignments in both the training (68%) and the prediction set (69%), as shown in Table 1-B. This shows the more reduced external predictive ability of CART when compared to the other strategies (the overall percentage of correct classification in the PSET is 75%), mainly due to the tendency to overestimate the persistence of low persistent chemicals (Class 3 misclassifications). However, even though this aspect should be considered when CART is applied for the classification of new chemicals, it is important to take into account that overestimation of persistence from low (Class 3) to medium (Class 2) can be considered a minor error than underestimation, according to the precautionary principle.

The k-NN model gave good and quite balanced overall results (81% correct assignments in the TSET and 83% in the PSET), and although it had lower accuracy than the CP-ANN model for Class 1 and Class 2, however the model presents strong predictive ability with high PSET accuracies (range 77–88%), comparable to the TSET (range 74–95%).

A final comment should be made considering the misclassifications of highly/medium persistent compounds as medium/low persistent by the three classification models. In fact, it is interesting to note (Table 2) that none of the persistent compounds are misclassified from Class 1 of high persistence to Class 3 of low persistence by CP-ANN and KNN, which make only moderate misclassifications among adjacent classes, while just 5% of the compounds *a priori* belonging to Class 1 of high persistence (in the PSET) are misclassified by CART to Class 3.

The three discriminant structural variables selected in this study (molecular weight, mean polarizability, and maximum electrotopological variation) are all bi-dimensional descriptors independent of chemical conformation, thus easily calculable from the bi-dimensional structural graph of a compound. The calculation of these variables can be performed by the appropriate software [30] even starting from the SMILES string of a chemical. These variables have been already mentioned as being related to processes of degradation in the environment [15,43,44]. Bigger and more complex chemicals are generally more persistent than smaller ones (with a lower MW) while electronic features (Mp and MAXDP) relate to polarizability and to a chemical's ability to form electrostatic and dipole–dipole interactions in the surrounding media. These features therefore can directly influence bioavailability and the partitioning of chemicals into different environmental compartments, and can indirectly determine their availability for different degradation pathways. It is also interesting to note that the descriptor MAXDP had already demonstrated its ability to model an important environmental partition property such as the soil sorption partition parameter Koc [43,44]. The relation between sorption and persistence is evident, as the more sorbed chemicals are the most recalcitrant to biotic and abiotic degradations, and are thus the more persistent. In addition, MAXDP, which gives information on the molecule's electrophilicity and is also related to the presence of halogen atoms, was a modeling descriptor which was also included in the global half-life index regression model [15].

Table 2
Percentage of misclassifications to adjacent classes for training and prediction set compounds (TSET and PSET, respectively) using the three classification methods (k-NN, CART and CP-ANN)

	a priori class	% misclassified to the adjacent Class 1	% misclassified to the adjacent Class 2	% misclassified to the adjacent Class 3	% correct
k-NN					
TSET	1	–	5	–	95
	2	9	–	15	76
	3	–	26	–	74
PSET	1		12		88
	2	8		14	77
	3		14		86
CART					
TSET	1	–	5	–	95
	2	9	–	6	85
	3	–	32	–	68
PSET	1	–	9	(5)[a]	86
	2	11	–	15	74
	3	(2)[a]	29	–	69
CP-ANN					
TSET	1	–	–	–	100
	2	3	–	15	82
	3	–	21	–	79
PSET	1	–	9	–	91
	2	7	–	12	81
	3	–	12	–	88

[a] Misclassification to not adjacent class (CART model only) are reported in brackets.

4.1. Analysis of the applicability domain

Since all the classification models were generated using the same combination of descriptors a direct comparison can be made between them in the same structural domain, also to give a better explanation and interpretation of the observed errors. The studied compounds were plotted in the space of the three molecular descriptors MW, Mp and MAXDP (Fig. 3) and labelled according to their *a priori* class of persistence.

The three structural descriptors are able to distinguish among these classes, which are satisfactorily separated in the graph. The less persistent compounds (squares) are mainly grouped on the right side, the highly persistent compounds (triangles) are located on the left side, while the medium persistent chemicals (dots) are grouped in the centre.

This graphic representation can be helpful in explaining misclassifications made by the different models. The chemicals misclassified by all the methods (numbered in Fig. 3), as well as their descriptors values, are listed in Table SI3 and distinguished in under- and over-estimation errors ("u" and "o"). Predicted consensus classes, based on unanimous misclassifications, are also reported.

Most of the overestimations listed in Table SI3, and numbered in Fig. 3, are related to compounds that belong to a given persistence class according to their original persistence data, but fall within the space of another *a priori* class, on the basis of the structural features encoded by the three model descriptors. These chemicals can be identified as outliers for the response. Differently, borderline chemicals falling, for structural reasons, within a region of partial overlapping of two adjacent classes were mainly underestimated. Even if it is not possible to highlight these

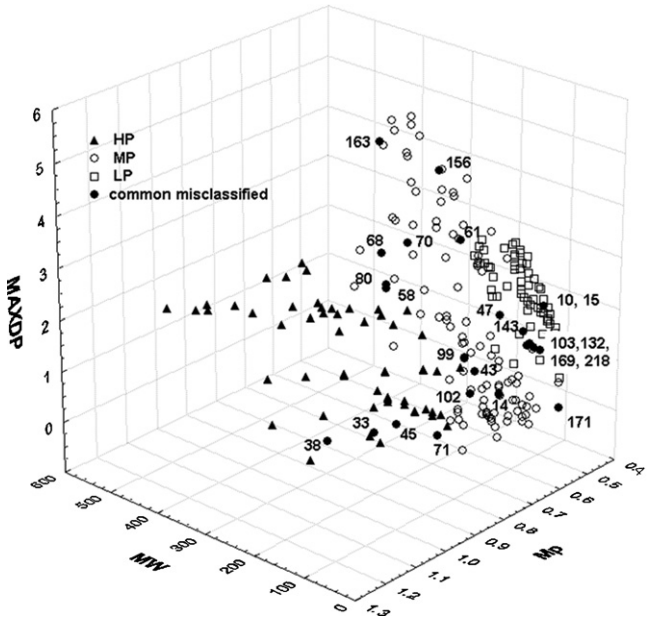

Fig. 3. Plot of the chemicals in the space of the three molecular descriptors of the proposed models. The compounds are labelled according to the classes of persistence and those commented on in the text are numbered according to the ID (Tables SI1–SI3 in Supporting information).

compounds as structural outliers, since they fall into the structural space defined by the three descriptors included in the model, they might however present some structural features that cannot be taken fully into account by the model's variables, with a consequent misclassification in the neighbour class.

The compounds misclassified by all the classification models are mainly overestimated in the adjacent class of persistence by the structural descriptors, thus, according to the precautionary principle, these errors can be considered as moderate misclassifications. The only apparently dangerous errors could be the misclassification of 1,2-dichloropropane (no. 43), and 1,2 dichloroethane (no. 102). These compounds should be in Class 1 of high persistence according to their input half-life data, but are predicted as less persistent by all the methods (even as low persistent by CART). This discrepancy was already highlighted in our previous study [15], thus new experimental determination of half-life values for these compounds is recommended.

4.2. Additional external validation

The external validation of our models, which predictivity was already demonstrated on the large PSET never included during the model development, was further on verified predicting the classes of potential persistence of about 50 additional chemicals taken from different literature sources [1,2,13,27,28].

A first set was chosen including 10 chemicals which overall persistence was studied by Klasmaier et al. [13] for the development of a multimedia model for persistence and long-range transport screening. On the basis of empirical evidence and of the multimedia models [13], six out of these ten chemicals (PCB-28, PCB-101, PCB-180, a-HCH, HCB, CCl_4 and aldrin) are POPs (Class 1); atrazine and biphenyl are described as moderately persistent (Class 2) while *p*-cresol is not persistent (Class 3).

These ten chemicals were correctly predicted using the three QSPR models here developed and were correctly assigned to the high (PCB-28, PCB-101, PCB-180, a-HCH, HCB, CCl_4, and aldrin), medium (atrazine and biphenyl) and not persistent (*p*-cresol) class (see also Table SI1).

Our classification models were then applied to a second set of 43 priority chemicals, that where additionally selected from different priority lists [1,2,27,28] (Table SI4). All the compounds commented as POPs in these priority lists have been, as expected, classified by all the models in Class 1 of high persistence (i.e. endrin, mirex, endosulfan, heptachlor, pentachlorophenol, some PAHs and chloronaphtalenes, BDEs, etc.). The chemicals, already recognized in the priority lists as of less concern for persistence, have been classified here also as belonging to Class 2 of medium persistence.

Finally our results were compared to those calculated (for persistence) by the US-EPA PBT Profiler [3] with 75% of agreement between the two approaches (results shown in Table SI4). The agreement of our results with those reported above and obtained by different approaches such as multimedia models, or expert judgment (in the case of priority lists), gives a further confirmation of the reliability and predictivity of our QSPR models, already strongly validated on the prediction set statistically generated by SOM splitting.

5. Conclusions

In conclusion, in this study three QSPR models were proposed, able to predict the intrinsic tendency of chemicals to be persistent, using few structural descriptors (calculable from the SMILES code by the software DRAGON [30] also freely available online at http://www.vcclab.org/lab/edragon). These models represent an extension of a previous study, which was focused on the development of a QSPR regression model for persistence [15], by applying the classification approach, more often used by authorities for the regulation of chemicals. The here presented classification models can distinguish the potential POP-like behavior of compounds among three different classes of global persistence and directly screen those potentially very persistent (assigned to Class 1, and exceeding three or four out of the four UNEP criteria for persistence) from the others.

The three developed models, based on the same molecular descriptors selected by Genetic Algorithm (MW, Mp, MAXDP) in k-NN, always gave very good classification results, verified by their accuracy on the single classes of environmental persistence assignment, both for training and prediction sets. The relevance of these structural features in describing the compounds' persistence in the environment is highlighted, even if the methods are based on different modeling approaches. The percentage of accuracies, ranging from about 70 to 100% calculated for all the models in both the training and the prediction set, confirms the high quality of our QSPRs, which are 25–50% more predictive than other classification models existing in literature and developed for a similar data set [23].

The CP-ANN model gave the best results, giving the highest overall classification percentages in the training and the prediction set, as well as the most balanced results in classifying individual classes, with a percentage of correct classifications always higher than 79%.

The CART model, though with slightly lower performance than the other two, was the simplest one being based on an easily applicable classification tree.

The structural domain of the models was investigated by a graphical view of the three classes of persistence distributed into the structural descriptors space.

The reliability of our models was finally successfully validated by comparing our results with those achieved by Klasmeier et al. [13] using multimedia models on 10 reference chemicals, as well as correctly identifying the most persistent chemicals in the UN and EU priority lists. A good agreement (75%) was also found between

our results and predictions made by the US-EPA PBT profiler. These results demonstrate that our QSPR approach can be used preliminarily and complementarily to multimedia models to predict persistence on a structural basis, without the use of other experimental data.

Moreover, since no method other than QSPR is applicable to detect the potential persistence of completely new compounds, all the models here presented could also be applied in "screening prior to synthesis". This procedure could avoid the production, and consequent release into the environment, of new possible POPs, orienting the synthesis towards safer alternatives [2]. The proposed modeling approach is also in line with the "proactive approach" [45] of the new European legislation for the Registration, Evaluation and Authorization of Chemicals (REACH) [4].

Acknowledgement

We thank Dr. Leon van der Wal for his critical comments on the draft of the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2008.02.004.

References

[1] The 1998 Aarhus Protocol on Persistent Organic Pollutants (POPs) http://www.unece.org/env/lrtap/pops_h1.htm (accessed May 30, 2007).
[2] UNEP, Stockholm Convention on Persistent Organic Pollutants, United Nations Environment Program, Geneva, Switzerland, 2001, http://www.pops.int.
[3] U.S. Environmental Protection Agency PBT-Profiler, http://www.pbtprofiler.net.
[4] REACH Regulation (EC) No. 1907/2006, http://eur-lex.europa.eu/LexUriServ/site/en/oj/2006/l_396/l_39620061230en00010849.pdf.
[5] D. Muir, P.H. Howard, Are there other persistent organic pollutants? A challenge for environmental chemists, Environ. Sci. Technol. 40 (2006) 7157–7166.
[6] E. Webster, D. Mackay, F. Wania, Evaluating environmental persistence, Environ. Toxicol. Chem. 17 (1998) 2148–2158.
[7] B.D. Rodan, D.W. Pennington, N. Eckley, R.S. Boethling, Screening of persistent organic pollutants: techniques to provide a scientific basis for POPs criteria in international negotiations, Environ. Sci. Technol. 33 (1999) 3482–3488.
[8] T. Gouin, D. Mackay, E. Webster, F. Wania, Screening chemicals for persistence in the environment, Environ. Sci. Technol. 34 (2000) 881–884.
[9] A. Beyer, D. Mackay, M. Matthies, F. Wania, E. Webster, Assessing long-range transport potential of persistent organic pollutants, Environ. Sci. Technol. 34 (2000) 699–703.
[10] D.W. Pennington, An evaluation of chemical persistence screening approaches, Chemosphere 44 (2001) 1589–1601.
[11] K. Fenner, M. Scheringer, M. MacLeod, M. Matthies, T. McKone, M. Stroebe, A. Beyer, M. Bonnell, A.C. Le Gall, J. Klasmeier, D. Mackay, D. Van de Meent, D. Pennington, B. Scharenberg, N. Suzuki, F. Wania, Comparing estimates of persistence and long-range transport potential among multimedia models, Environ. Sci. Technol. 39 (2005) 1932–1942.
[12] J.A. Arnot, D. Mackay, E. Webster, J.M. Southwood, Screening level risk assessment model for chemical fate and effects in the environment, Environ. Sci. Technol. 40 (2006) 2316–2323.
[13] J. Klasmeier, M. Matthies, M. MacLeod, K. Fenner, M. Scheringer, M. Stroebe, A.C. Le Gall, T. McKone, D. Van De Meent, F. Wania, Application of multimedia models for screening assessment of long-range transport potential and overall persistence, Environ. Sci. Technol. 40 (2006) 53–60.
[14] T. Oberg, Virtual screening for environmental pollutants: structure activity relationships applied to a database of industrial chemicals, Environ. Toxicol. Chem. 25 (2006) 1178–1183.
[15] P. Gramatica, E. Papa, Screening and ranking of POPs for global half-life: QSAR approaches for prioritization based on molecular structure, Environ. Sci. Technol. 41 (2007) 2833–2839.
[16] A. Sabljic, QSAR models for estimating properties of persistent organic pollutants required in evaluation of their environmental fate and risk, Chemosphere 43 (2001) 363–375.
[17] A. Sabljic, W. Peijnenburg, Modeling lifetime and degradability of organic compounds in air, soil, and water systems—(IUPAC Technical Report), Pure Appl. Chem. 73 (2001) 1331–1348.
[18] L. Carlsen, J.D. Walker, QSARs for prioritizing PBT substances to promote pollution prevention, QSAR Comb. Sci. 22 (2003) 49–57.
[19] P. Gramatica, P. Pilutti, E. Papa, Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling, J. Chem. Inf. Comput. Sci. 44 (2004) 1794–1802.
[20] P. Gramatica, P. Pilutti, E. Papa, A tool for the assessment of VOC degradability by tropospheric oxidants starting from chemical structure, Atmos. Environ. 38 (2004) 6167–6175.
[21] O.G. Mekenyan, S.D. Dimitrov, T.S. Pavlov, G.D. Veith, POPs: A QSAR system for developing categories for persistent, bioaccumulative and toxic chemicals and their metabolites, SAR QSAR Environ. Res. 16 (2005) 103–133.
[22] D. Aronson, R. Boethling, P. Howard, W. Stiteler, Estimating biodegradation half-lives for use in chemical screening, Chemosphere 63 (2006) 1953–1960.
[23] R. Kuhne, R.U. Ebert, G. Schüürmann, Estimation of compartmental half-lives of organic compounds—structural similarity versus EPI-Suite, QSAR Comb. Sci. 26 (2007) 542–549.
[24] E. Papa, P. Gramatica, Structurally-based tools for the screening and prediction of the environmental persistence of organic chemicals, in: Proceedings of the 10th EuCheMS-DCE International Conference, Rimini, Platform, 4–9 September, 2005.
[25] E. Papa, P. Gramatica, Screening of POPs by QSAR classification models, in: Proceedings of the 17th Annual Meeting SETAC-Europe, Porto, Portugal, Platform, 20–24 May, 2007.
[26] D. Mackay, W.Y. Shiu, K.C. Ma, Physical–Chemical Properties and Environmental Fate Handbook, CRCnet-BASE CD-ROM, Chapman and Hall/CRC, Boca Raton, FL, USA, 2000.
[27] D. Lerche, E. van de Plassche, A. Schwegler, F. Balk, Selecting chemical substances for the UN-ECE POP protocol, Chemosphere 47 (2002) 617–630.
[28] EEB Comments on European Parliament and Council Decision establishing a list of priority substances in the field of water policy (COM (2000))47 final. And: Commission proposal on a procedure for selection of priority hazardous substances according to Article 16.3 WFD (ENV/140900/01rev), http://www.eeb.org/publication/2000/1016_EEB_Comments_PS-IdentPHS_WFD.pdf (accessed February 20, 2008).
[29] SCAN, Software for Chemometric Analysis, rel. 1.1 for Windows, Minitab, USA, 1995.
[30] DRAGON for Windows (Software for Molecular Descriptor Calculations). Version 5.4, Talete s.r.l., Milan, Italy, 2006, http://www.talete.mi.it.
[31] HyperChem rel. 7.03 for Windows, Autodesk, Inc., Sausalito, CA, USA, 2002.
[32] A. Tropsha, P. Gramatica, V.J.K. Gombar, The importance of being Earnest: validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 22 (2003) 69–77.
[33] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26 (2007) 694–701.
[34] J. Gasteiger, J. Zupan, Neural networks in chemistry, Angew. Chem. Int. Ed. Engl. 32 (1993) 503–527.
[35] J. Zupan, M. Novic, I. Ruisánchez, Kohonen and counter propagation artificial neural networks in analytical chemistry, Chemometr. Int. Lab. Syst. 38 (1997) 1–23.
[36] R. Todeschini, Milano Chemometrics and QSAR Research group, KOALA-Software for Kohonen Artificial Neural Networks, Version 1.0 for Windows 2001, Milan, Italy.
[37] R. Leardi, R. Boggia, M. Terrile, Genetic algorithms as a strategy for feature selection, J. Chemometr. 6 (1992) 267–281.
[38] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, Mobydigs: software for regression and classification models by genetic algorithms, in: R. Leardi (Ed.), Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks, Elsevier, Amsterdam, The Netherlands, 2003.
[39] M.A. Sharaf, D.L. Illman, B.R. Kowalski, Chemometrics, Wiley, New York, 1986.
[40] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, CRC Press, Boca Raton, Florida, 1998.
[41] R. Hecht-Nielsen, Application of counter-propagation networks, Neural Networks 1 (1988) 131–140.
[42] J.E. Dayhoff, Neural Network Architectures, An Introduction, Van Nostrand Reinhold, New York, 1990.
[43] P. Gramatica, M. Corradi, V. Consonni, Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors, Chemosphere 41 (2000) 763–777.
[44] P. Gramatica, E. Giani, E. Papa, Statistical external validation and consensus modeling: a QSPR case study for Koc prediction, J. Mol. Graph Model. 25 (2007) 755–766.
[45] J.A. Field, C.A. Johnson, J.B. Rose, What is "emerging"? Environ. Sci. Technol. 40 (2006) 7105.