Available online at www.sciencedirect.com

ScienceDirect

Journal of Molecular Graphics and Modelling 25 (2007) 755–766

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Statistical external validation and consensus modeling: A QSPR case study for K_{oc} prediction

Paola Gramatica *, Elisa Giani, Ester Papa

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

Received 14 February 2006; received in revised form 26 June 2006; accepted 26 June 2006
Available online 4 August 2006

Abstract

The soil sorption partition coefficient (log K_{oc}) of a heterogeneous set of 643 organic non-ionic compounds, with a range of more than 6 log units, is predicted by a statistically validated QSAR modeling approach. The applied multiple linear regression (ordinary least squares, OLS) is based on a variety of theoretical molecular descriptors selected by the genetic algorithms-variable subset selection (GA-VSS) procedure. The models were validated for predictivity by different internal and external validation approaches. For external validation we applied self organizing maps (SOM) to split the original data set: the best four-dimensional model, developed on a reduced training set of 93 chemicals, has a predictivity of 78% when applied on 550 validation chemicals (prediction set). The selected molecular descriptors, which could be interpreted through their mechanistic meaning, were compared with the more common physico-chemical descriptors log K_{ow} and log S_w. The chemical applicability domain of each model was verified by the leverage approach in order to propose only reliable data. The best predicted data were obtained by consensus modeling from 10 different models in the genetic algorithm model population.

Keywords: Theoretical molecular descriptors; Genetic algorithms; Splitting; Soil sorption coefficient; K_{oc}; QSAR; OECD principles

1. Introduction

Sorption processes play a major role in determining the environmental fate, distribution and persistence of chemicals. An evaluation of the soil mobility of chemicals is a primary task in estimating the environmental distribution of chemicals. An important parameter in studying this process for organic chemicals is the soil sorption coefficient, expressed as the ratio between chemical concentration in soil and in water, normalized to organic carbon (K_{oc}). The experimental measurement of K_{oc} is difficult, expensive and time-consuming, thus a great deal of effort has been put into attempting the estimation of K_{oc} through statistical modeling. Indeed, the large number of existing, and new, chemical compounds calls for general and fast quantitative models to rapidly screen and assess the risk of chemicals. Several quantitative structure-activity/property relationships (QSAR/QSPR) models predicting the soil sorption partition coefficient for specific classes of chemicals (mainly pesticides), and global models for heterogeneous compounds, have been published in recent years, and there has been much discussion of the different approaches concerning molecular description and adopted methodologies in some reviews [1–3]. Many published QSAR models are based on correlations with experimental data, mainly with octanol/water partition coefficients (K_{ow}) and water solubility (S_w), others on molecular structure descriptors. All the methods, reviewed in the cited reviews, have been fairly successful in obtaining good models, internally validated for predictivity, but little work has been done to examine model predictivity and the chemical domain of application over a wide range of compounds, especially for new chemicals. In effect, a QSAR model must be validated for its predictivity before it can be used to predict the response of additional chemicals. Validating QSAR with external data (i.e. data not used in the model development), although the most demanding, is the best method of validation [4]. However the availability of an independent external validation set of several compounds is rare in QSAR. Thus, the input data set must be adequately split by experimental design or other splitting procedures [5–7] into representative training and validation/prediction sets.

* Corresponding author. Tel.: +39 0332 421573; fax: +39 0332 421554.
E-mail address: paola.gramatica@uninsubria.it (P. Gramatica).
URL: http://www.qsar.it

1093-3263/$ – see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2006.06.005

The object of this study is the proposal of different MLR–QSAR models of K_{oc} for a wide and highly heterogeneous data set of 643 non-ionic organic chemicals collected from three main sources [1,8,9]. The peculiarity of these models, that are based on different theoretical molecular descriptors selected by genetic algorithm as a variable subset selection procedure, is their development on a training set very much smaller than the prediction set (in a 1:6 ratio) and their applicability to a very heterogeneous set of chemicals. Consensus modeling, which considers more than just one QSAR model is also proposed and commented on. Recently published QSAR models, developed on big data sets of organic chemicals and not included in the previous reviews, are discussed and compared with our findings.

One of the most important aspects of the models proposed in this paper is that the fundamental points set down by OECD principles [10,11] for regulatory acceptability of QSARs are fulfilled during model development. These requirements are: (i) validation for predictivity (both by internal and external validation by different statistical approaches [12]), (ii) the checking of the chemical applicability domain by the leverage approach, (iii) the mechanistic interpretation of molecular descriptors. In this way only reliably predicted data are proposed.

2. Data and methods

2.1. Experimental data

The experimental data of the soil sorption partition coefficient, normalized on organic carbon (K_{oc}), of 643 heterogeneous organic compounds were collected from the literature [1,8,9] and compiled into a single database (see Table 1 in supporting information). Some of the chemicals in the literature databases have more than one K_{oc} value, the result of being derived from different sources; in these cases the median was adopted. The modelled data were expressed in logarithmic units (log K_{oc}), with a range from −0.31 to 6.02 in the training set and from 0 to 6.33 in the prediction set, for chemicals with a log K_{ow} range of −2.11 to 8.39 (from −0.87 to 7.45 in the training set and from −2.11 to 8.39 in the prediction set). The data set is highly heterogeneous, and includes practically all the principal functional groups present in pesticides and various organic pollutants.

2.2. Theoretical molecular descriptors

The molecular descriptors for the given compounds were mainly calculated using *DRAGON* software [13] on the (x,y,z)-atomic co-ordinates of the minimal energy conformations determined by the MM+ method in *HYPERCHEM* Package [14]. A total of 1079 molecular descriptors of differing types were calculated to describe compound structural diversity and used as input variables for variable selection by genetic algorithm. The descriptor typology is: (a) 0D-48 constitutional (atom and group counts), (b) 1D-154 functional groups, (c) 1D-120 atom centered fragments, (d) 2D-119 topological, (e) 2D-64 descriptors of Burden (BCUTs: Burden-CAS-University of Texas eigenvalues), (f) 2D-21 Galvez Indices from the adjacency matrix, (g) 2D-96 various auto-correlations from the molecular graph, (h) 2D-33 connectivity index, (i) 2D-47 information index, (j) 2D-47 walk and path counts, (k) 2D-44 eigenvalue-based indices, (l) 3D-41 Randic molecular profiles, (m) 3D-74 geometrical descriptors, (n) 3D-99 Weighted Holistic Invariant Molecular descriptors (WHIMs) and (o) 3D-197 Geometry, Topology and Atom-Weights AssemblY (GETAWAY) descriptors. The list and meaning of the molecular descriptors is provided by the *DRAGON* package, and the calculation procedure is explained in detail, with related literature references, in the *Handbook of Molecular Descriptors* [15]. Furthermore, descriptors from EPISuite [16] are added: log K_{ow} and log S_w.

Constant values and descriptors found to be correlated pairwise were excluded in a pre-reduction step (one of any two descriptors with a correlation greater than 0.97 was removed to reduce redundant information), thus 479 molecular descriptors underwent subsequent variable selection.

2.3. Model building, variable selection and validation of internal predictivity

Multiple linear regression (MLR) analysis and variable selection were performed by the last version of the software MOBY DIGS [17], using the ordinary least squares regression (OLS) method for the modeling and the genetic algorithm-variable subset selection (GA-VSS) method [18] for the variable selection. The genetic algorithm was applied to the input set of 479 molecular descriptors for each chemical of the studied training sets and the related response, in order to extract the best set of molecular descriptors, which are, in combination, the most relevant variables in modeling the response of the training set chemicals. In the last version of the MOBY-DIGS software [17] the descriptors can be distributed, from the very beginning, into one or more populations, thus the GA model searching is performed contemporarily on the different populations of models, that can run separately: in our study we simultaneously developed three separate model populations based on different typologies of DRAGON molecular descriptors (1D, 2D and 3D). The models were initially developed by the all-subset-procedure: until two variable models were achieved for each different population; this technique explored all the low dimension combinations. Then, starting from the obtained populations, each of 100 models, the GA explores new combinations, selecting the variables by a mechanism of reproduction/mutation analogous to that of biological population evolution. In this first step the optimized parameter is the cross-validated correlation coefficient R^2_{cv} or Q^2_{LOO} (*leave-one-out*).

The GA selection was stopped when increasing the model size did not increase the Q^2 value to any significant degree (here, up to four variables). During evolution, new model populations can be created from the genetic heritage (the variables) of other existing populations, and these model populations can migrate to other existing populations that have their own genetic heritage. This leads to the exploration of a larger model space and increases the possibilities of developing a better quality population. Once the populations have evolved sufficiently, a final subset of models can be selected from the existing populations; this results in the final population of 100 satisfactory regression models, ordered according to their decreasing internal predictive performance, verified by Q^2.

On these final models, several different validation tools are further applied, such as the bootstrap technique, external validation, and Y scrambling.

The bootstrap approach was applied to verify robustness and internal predictivity. This procedure generates K n-dimensional groups by a repeated random selection of n-chemicals from the original data set. The model obtained on the first selected chemicals is used to predict the values for the excluded compounds and then Q^2 is calculated for each model. The bootstrapping was repeated 5000 times for each validated model.

The proposed models were also checked by permutation testing: new parallel models were developed based on fit to randomly reordered Y-data (Y scrambling), and the process was repeated several times (300 iterations). The resulting models obtained on the data set with randomized response should have significantly lower Q^2 values than the proposed ones because the relationship between the structure and response is broken. This is proof of the proposed model's validity as it can be reasonably excluded that the originally proposed model was obtained by chance correlation [4,19].

2.4. Chemical domain

The chemical domain of the studied chemicals in the models was verified by the leverage approach to verify prediction reliability [4,19]. The plot of standardised residuals versus leverages (hat diagonals), i.e. the Williams graph, obtained by the *SCAN* package [20], verified the presence of response outliers (i.e. compounds with cross-validated standardized residuals greater than three standard deviation units, $\pm 3\sigma$) and chemicals very structurally influential in determining model parameters. Also the data predicted by the models were verified for reliability by their leverage, so that only predicted data for chemicals belonging to the chemical domain of the training set would be proposed. In fact, leverage can be used as a quantitative measure of the model applicability domain suitable for evaluating the degree of extrapolation: it represents a sort of compound "distance" from the model experimental space. Prediction must be considered unreliable for compounds with a high leverage value ($h > h^*$, the critical value being $h^* = 3p'/n$, where p' is the number of model variables plus one, and n is the number of the objects used to calculate the model). Conversely, when the leverage value of a compound is lower than the critical value, the probability of accordance between predicted and actual values is as high as that for the training set chemicals [21].

2.5. External validation: splitting training/prediction sets

In order to obtain compounds for external validation, the available set of chemicals is split into a training set and an external prediction set. Two different splitting methods are applied. The splitting of the data set, realized by Kohonen map-artificial neural network or self organizing maps (SOM) [22,23] using the package KOALA [24], takes advantage of the clustering capabilities of SOM, allowing the selection of a meaningful training set and a representative prediction set. The 41 most significant principal components, calculated from each group of *DRAGON* molecular descriptors, were used to describe the relevant structural information of the chemicals. This structural information and the response were used as variables to build a Kohonen map (10 × 10 neurons, 300 epochs). At the end of 300 epochs of net training, similar chemicals are close to each other in the multi-dimensional descriptor space (falling within the same cell of the top map), i.e. they carry the same structural information. To select the training set of chemicals, it is assumed that the compound closest to each cell centroid is the most representative of all the chemicals within the same cell. Thus, the selection of the training set chemicals was performed by the minimal distance from the centroid of each cell in the top map. The remaining objects, close to the training set chemicals, were used for the prediction set (14.5% in training set–85.5% in prediction set).

Additional splitting was carried out by random selection through activity sampling, thus by ordering the chemicals according to their descending experimental values, selecting the most and the least active, and taking every fourth chemical from the set in the prediction set to be used after model development for the external validation (25% of the total data set).

2.6. Predictivity

The predictive power of the regression model developed on the selected training set is estimated on the predicted values of prediction set chemicals, by the external Q^2 that is defined [25]:

$$Q^2_{\text{ext}} = 1 - \frac{\sum_{i=1}^{\text{pred}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{\text{pred}} (y_i - \bar{y}_{\text{tr}})^2}$$

where y_i and $\hat{y}_i$ are, respectively, the measured and predicted (over the prediction set) values of the dependent variable, and $\bar{y}_{\text{tr}}$ the averaged value of the dependent variable for the training set; the summations cover all the compounds in the prediction set.

Other measures applied to define the accuracy in predictivity of the proposed QSARs are the coefficient of determination R^2 calculated for the prediction chemicals by applying the model developed on the training set, and the root mean squared error (RMSE) for the two sets. The RMSE summarizes the overall error of the model: it is calculated as the root square of the sum of squared errors in prediction divided by their total number.

Standard error of estimate (s), together with the coefficient of determination (R^2) are also reported for each model.

2.7. Principal component analysis

Principal component analysis (PCA) and multidimensional scaling (MDS) for data exploration were performed on auto-scaled data by the SCAN [20] and/or STATISTICA [26] packages.

3. Results and discussion

Table 1 lists the best, recently published, QSAR models developed on big data sets (161–643 chemicals) and not included in the cited reviews [1–3]; they are in chronological sequence with the main statistical parameters. For reference purposes the fundamental paper of Sabljic et al. [1] is also included.

The aim of this work is the development, using theoretical molecular descriptors, and the proposal of externally validated general QSAR models for the prediction of K_{oc} for a wide and heterogeneous set of non-ionic organic compounds. A similar approach was applied in our preliminary study on a set of pesticides [28]. In the present paper we have increased the dimension and structural diversity of the data set (the biggest studied so far) and developed different MLR models, always verifying their applicability and generalizability. This is done by adequately splitting the available set of experimental data into a very reduced representative training set (even less than 15% of the original data set, Kahn et al. [32] used the 20%) for model development and a big prediction set (more than 85% of the original data) for model performance inspection. The application of a single and general QSAR model, based on theoretical molecular descriptors, to a large set of heterogeneous compounds could be very useful for screening big data sets and for planning new, environmentally friendly, chemicals. The great advantage of theoretical descriptors is that they can be calculated homogeneously by a defined software for all chemicals, even those not yet synthesized, the only need being a hypothesized chemical structure.

An analysis of already published models (Table 1) allows an understanding of the main structural features related to soil sorption ability. Indeed, in the various models, the same structural aspects, mainly those related to molecular size, are represented by alternative descriptors with a similar meaning for K_{oc} modeling: for instance log K_{ow}, fragments parameters, topological descriptors and MW. Some useful electronic parameters are captured in the different alternative descriptors such as E-state indices [9,30], quantum chemical descriptors (PNSA-1, η[AM1], [AM1], P^{max}) [32], correction factors [8,27], the number of phenyls, oxygen, nitrogen and sulfur atoms [31] and the number of hydrogen acceptor atoms (nHA, that is the same as using nF plus nO plus nN), the number of nitro groups (nNO), or the maximum positive intrinsic state difference related to the electrophilicity of the molecule (MAXDP) [28].

In order to find a relationship between log K_{oc} and the structural features of the chemicals, a wide and diversified set of theoretical molecular descriptors that take into account different structural features (mono-dimensional, bi-dimensional and three-dimensional) was used. We used many different types of molecular descriptors as the modeling input variables, in order to have the possibility of catching, in the models, all the relevant structural features really related to the studied response. As we could not have *a priori* knowledge of which descriptors, or which particular descriptor combinations, could be related to the response, and therefore which could be usefully used in models for prediction aims, we applied genetic algorithms [18] as the variable selection procedure (GA-VSS) to select only the best combinations of those descriptors most relevant to obtaining models with the highest predictive power for K_{oc}.

3.1. Internal and external validation

The real usefulness of QSAR models is not just their ability to reproduce known data, verified by their fitting power (R^2), but is mainly their possibility of predictive application for new chemicals. For this reason, even though the model calculations were performed by maximizing the explained variance in cross-

Table 1
List of recently published QSAR models of log K_{oc} and statistical parameters

No.	Train	Test	Model descriptors	R^2	Q^2_{LOO}	R^2_{ext}	*s* train	*s* test	Reference
1	390		log K_{ow}	0.63	0.62		0.557		Sabljic et al. [1]
2	592		74 fragments, 24 structural factors	0.97					Tao et al. [8]
3	430	162	86 fragments, 19 structural factors	0.97					Tao et al. [8]
4	400	143	$^1\chi^v$, $^2\chi$, $^6\chi^{cb}$, $\sum F_i n_i$	0.87					Tao and Liu [27]
5	543		$^1\chi^v$, $^2\chi$, $^6\chi^{cb}$, $\sum F_i n_i$	0.86					Tao and Liu [27]
6	141	20	MW, nNO, nHA, CIC, MAXDP, Ts	0.84	0.82	0.67	0.35	0.54	Gramatica et al. [28]
7	68	274	2 PLS various descriptors	0.69–0.75	0.49–0.65				Andersson et al. [29]
8	403	165	log *S* (calc.)	0.80	0.8	0.76	0.515	0.6	Huuskonen [9]
9	403	165	log *S* (calc.), HBA, NAR, MW, I_{acid}	0.85	0.845	0.79	0.448	0.6	Huuskonen [9]
10	403	165	log K_{ow} (calc.)	0.79	0.78	0.73	0.528	0.7	Huuskonen [9]
11	403	165	log K_{ow} (calc.), NAR, ROT, MW, I_{acid}	0.86	0.85	0.80	0.434	0.6	Huuskonen [9]
12	143	20	$^1\chi$, 11 E-state indices (S_i)	0.82	0.79	0.79	0.37	0.45	Huuskonen [30]
		38				0.74		0.65	
13	82	43	N_Φ, M_W, N_N, N_O, N_S	0.94	0.93	0.91	0.33	0.3	Delgrado et al. [31]
14	344		log K_{ow}, PNSA-1, η[AM1], [AM1], P^{max}	0.76	0.75		0.409		Kahn et al. [32]
15	68	276	log K_{ow}, η[AM1]	0.76	0.73	0.70	0.439	0.5	Kahn et al. [32]
16	93	550	VED1, nHAcc, MAXDP, CIC0	0.82	0.80	0.78	0.567	0.559	Present study
17	93	550	Consensus model	0.82		0.81	0.542	0.534	Present study

validation (by Q^2_{LOO}) with GA-VSS (see Section 2.1), the choice of the best models was made by a very strong verification of both the internal predictivity (by Q^2_{Boot}) and, especially, the external predictivity.

Y-randomisation was also applied to exclude the possibility of chance correlation, i.e. fortuitous correlation without any predictive ability. It gave the following results: the random models, performed using a scrambled order of the experimental rate constant, were found to have significantly lower R^2 and Q^2 than the original models, corroborating the statistical reliability of the actual models.

We examined all the most important internal validation criteria, namely cross-validation (LOO), bootstrapping and Y-scrambling, that appear to be necessary conditions, though still not sufficient, for the model to have high predictive power [12]. In fact we, like other authors [33–38], are strongly convinced, from personal experience, that models with high estimated ‘‘predictivity’’, highlighted only by internal validation methods, can be less predictive and even unpredictive when verified on new chemicals not used in developing the model [4].

Thus, for a stronger evaluation of model applicability for prediction on new chemicals, the effective predictive capability of a model was evaluated by the ‘‘external’’ validation procedure, i.e. by comparing the predictions made for molecules excluded from the model generation step with their actual experimental activity.

3.2. Splitting of the data set

Given a single data set, which is a typical situation in QSAR modeling, the external validation at the model development step can only be achieved by splitting the original data set into a training set, used to establish the QSAR model, and a prediction set, to evaluate model performance. Rational division of the experimental data set into training and prediction set is a crucial part in the development and validation of reliable QSAR models [4]. The approaches for creating training and prediction sets range from straightforward random selection, through various clustering techniques [39–42], to the methods of Kohonen map-artificial neural network, now more widely called self organizing maps [22,23], and formal statistical experimental design (factorial and D-Optimal) [43]. The underlying goal at this step is to ensure that both the training and the prediction sets separately span the whole descriptor space occupied by the entire data set, and that the chemical domain in the two data sets is not too dissimilar.

One of the techniques used in this work to select the prediction set is self organizing maps. Through its clustering capabilities, SOM ensures that both sets are homogeneously distributed within the entire area of the descriptor space; in this case the chemicals in both sets, selected to maximize the coverage of the descriptor space (i.e. representativity), represent the structural variety of the studied data set in a balanced way. The selected training chemicals are those with the minimal distance from the centroid of each cell in the top map. In this case, the representative points of the prediction set are close (in the same cell of the top map) to representative points of the training set in the multidimensional structural descriptor space [6,7]. This allows predictions to be made by interpolation and not extrapolation out of the domain of the particular QSAR model [4].

In our statistical approach the splitting methodology was applied *a priori* to all the relevant structural information in the original molecular descriptor sets: the 41 significant principal components of molecular descriptors were calculated from each group of *DRAGON* descriptors. This methodology was applied using a very strong splitting to select the ‘‘minimal training set’’, which covers the information space efficiently: 93 chemicals in the training set and 550 in the prediction set (thus only 14.5% of the chemicals, used for the development of models, will be able to predict about 86% of the available chemicals).

3.3. Selection of the best externally predictive model

Our QSAR approach is statistical, and is based on the genetic algorithm selection of variables modeling the studied response, thus the similarity in the structural information resulting from the various and interchangeable DRAGON molecular descriptors provides a population of similar models with similar predictive power based on different sets of molecular descriptors. All the selected models developed on the training set and used to predict values for the prediction set chemicals have high predictive performance, verified with both internal and external validation. This demonstrates that the structural information included in an informative and representative training set, selected by adequate splitting methodology [5–7], is sufficient, though very reduced, for the best prediction of a validation set six times bigger.

The best predictive model, based on 4 variables, was selected from a population of 100 models of different descriptor typology (no. variables = 1–4). When considering the population of the 60 best four-dimensional models, the range of Q^2 is 0.781–0.823, while the range of Q^2_{ext} is 0.584–0.785 (mean $Q^2_{\mathrm{ext}} = 0.749$). The Q^2_{ext} values confirm the high predictive ability of the majority of the models in the population selected by GA; this guarantees that these models are also reliable for use on new chemicals not employed in the developed model, with the condition that they belong to the same chemical domain. It is very important to highlight that in this population it was again verified that not all the models with high internal predictivity (verified by high values of Q^2_{LOO} and Q^2_{Boot}) have the same high predictive performance on external chemicals (verified by Q^2_{ext}). For instance, three models decrease in predictivity by nearly 0.22, from $Q^2_{\mathrm{LOO}} = 0.80$ to $Q^2_{\mathrm{ext}} = 0.58$, while this difference is smaller, 0.01–0.06, in those models that can be considered really externally predictive.

The crucial problem of the difference between internal and external predictivity of a model has already been stressed by the authors and others in several papers [4,6,7,12,34–38].

The reference model we chose was the one which, from among the models with higher external predictivity, had the smallest difference between internal and external predictivity and the fewest chemicals outside the chemical domain (with high leverage value). The equation and the statistical parameters of the proposed model (in Tables 1 and 2) are:

$$\begin{aligned}\log K_{\mathrm{oc}} &= -2.19(\pm0.30) + 2.10(\pm0.14)\mathrm{VED1}\\ &\quad - 0.34(\pm0.04)\mathrm{nHAcc} - 0.31(\pm0.05)\mathrm{MAXDP}\\ &\quad - 0.33(\pm0.12)\mathrm{CIC0}\\ n(&\text{training set}) = 93 \quad n(\text{prediction set}) = 550\\ R^2 &= 0.82 \quad Q^2 = 0.80 \quad Q^2_{\mathrm{BOOT}} = 0.79 \quad Q^2_{\mathrm{ext}} = 0.78\\ s &= 0.539 \quad \mathrm{RMSE}_{\mathrm{training}} = 0.523 \quad \mathrm{RMSEP}_{\mathrm{LOO}} = 0.554\\ \mathrm{RMSEP}&_{\mathrm{prediction}} = 0.560 \qquad (1)\end{aligned}$$

The model is stable and predictive both internally, as can be verified by the statistical parameters (high value of cross-validation parameters Q^2 and Q^2_{BOOT}, check of Y-scrambling), and externally (similar high value of Q^2_{ext}); the small values of standard deviation errors and the stability of RMSE/RMSEP for both training and prediction sets are additional proof of model predictivity. It is important to note that the standard deviation of the experimental log K_{oc} was calculated [44] to be about 0.44 and that the value of the model (1) is not far from this experimental value: thus, the model is able to predict data with an accuracy similar to experimental error. Fig. 1 shows the regression line of the above proposed model, and highlights (numbering them as in Table S1 of the supporting information listing the chemicals and data predicted by this model) the nine chemicals predicted badly (methylurea (330), 2,3,5-trimethylphenol (358), benfluralin (408), 2,6-dichlorobenzamide (427), 2,6-dinitro-*n*-propyl-trifluoro-*p*-toluidine (432), toxaphene (499), dinitramine (556), and oxyfluoren (591) in the training and trifluralin (394) in the prediction set): the majority of data are underestimated, but this need not be considered highly dramatic, being very near the 3σ line. These response outliers are structurally heterogeneous, thus it is difficult to find a reason for why the model failed to predict them accurately; however, it is important to keep in mind that the quality of the input experimental data could be arguable for some of these chemicals.

3.4. Structural chemical domain

QSAR models must always be verified for their applicability with regard to chemical domain, in order to produce predicted data that can be considered reliable only for not too structurally dissimilar chemicals. In fact, in the case of structurally dissimilar molecules, the data predicted by the model must be judged as extrapolations. The chemical domain of applicability of the reported model was verified by an analysis of the Williams graph of Fig. 2, in which the standardized residuals and the leverage value (h) are plotted.

In addition to the above listed nine response outliers, three chemicals are outliers for structure (with leverage higher than the warning h value of 0.16) and, for this reason, their predictions must be more correctly considered as extrapolated by the model: chlordecone (258), metasulfron methyl (628), thiameturon methyl (637). Chlordecone, included in the prediction set, has a very peculiar structure, probably not sufficiently represented in the training set; the two pesticides (metasulfron methyl of the prediction set and thiameturon methyl of the training set) have high structural similarity and thus, as expected, similar high leverage values. In this case the data predicted by the model for these influential chemicals are good (lying perfectly along the regression line of Fig. 1), thus they are ‘‘good leverage’’ chemicals, but this situation could not always be verified: indeed the data for high leverage chemicals could be extrapolated wrongly, especially if no high leverage chemical is included in the training set. The presence of the well modeled training set chemical, thiameturon methyl (637), allows also metasulfron methyl (628) of the prediction set to be well predicted.

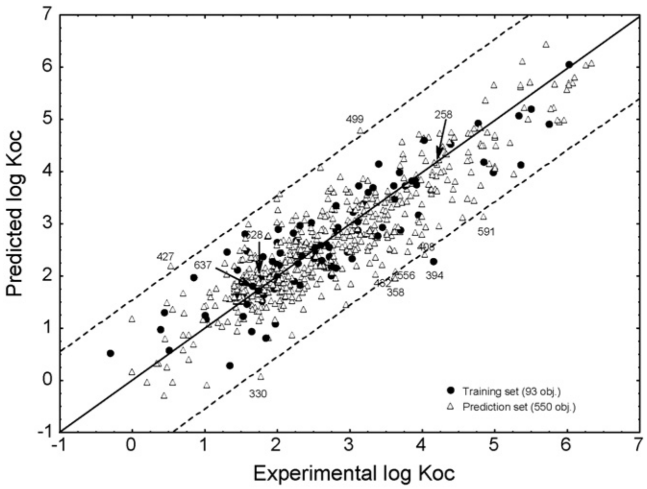

Fig. 1. Regression line for the model (1). The log K_{oc} values for the training and prediction set chemicals are labeled differently, the outliers are numbered as in Table S1 of supporting information. The dotted lines indicate the 3σ interval.

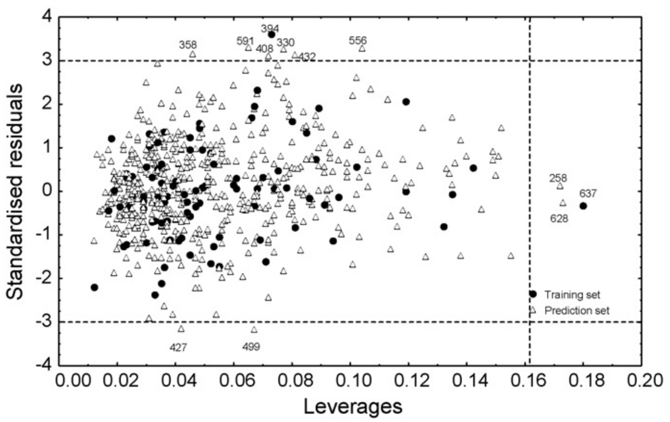

Fig. 2. Williams plot for the model with four variables. The log K_{oc} values for the training and prediction set chemicals are labeled differently, the response outliers and structurally influential chemicals are numbered. The dotted lines are, respectively, the 3σ limit and the warning value of hat ($h^* = 0.16$).

3.5. Mechanistic interpretation of the descriptors

In the above Eq. (1) the descriptors are reported in decreasing order of significance, according to standardized coefficient values (reported below in brackets, after the descriptor symbol).

The best descriptor is the eigenvector coefficient sum from distance matrix (VED1, 1.31) encoding the 2D molecular dimension [45], followed by the number of acceptor atoms for H-bonds (nHAcc, −0.64) and then the maximal electropological positive variation [28] (MAXDP, −0.40); the least relevant is a 2D-descriptor, a complementary information content index [46,47] neighbourhood symmetry of 0 order (CIC0, −0.20) which is related to the differences in the atomic distribution and the molecular dimension of the studied chemicals.

The nHAcc descriptor, that is related to electronegative atoms of molecules, and MAXDP, related to molecule electrophilicity [28], represent different ways of taking into account the probability of bond formation between chemicals and groundwater: as expected, these descriptors are negative in sign as high affinity for water precludes soil sorption of the chemicals. The other two descriptors are related to molecular size, but their relevance is very different: the more important VED1 has a positive sign, highlighting that the bigger compounds are more sorbed than leached, the less relevant descriptor CIC0, added as the last variable in the nested models, is probably useful only to improve model quality in order to adapt some particular chemicals.

Finally, it is interesting to note that a comparison of the statistical parameters of the different GA populations, based on molecular descriptors with distinct and separate structural information (0-1D, 2D, 3D), highlights that the best models are always based on a combination of 0-1D and 2D descriptors. Indeed, the addition of the more complex 3D descriptors is not useful for improving the modeling of K_{oc}, that is well predicted by simpler structural information.

3.6. Evaluation and prediction sets. Scrambling of sets

Nowadays, in the absence of new reliable data, it is a sound custom at the model development step to carry out statistical external validation by splitting the available data and using only a portion (the more representative and the smallest number possible) for model development, and to apply the obtained model on the external prediction set [4,6,7,12,29–38]. In such an approach the splitting methodologies are based on similarity analysis, therefore the external prediction set of chemicals is, by definition, as structurally similar as possible to the training set chemicals: this allows the same chemical domain to be maintained, as it is obviously impossible to have reliable predictions for chemicals outside the model applicability domain [4,5,12]. However, in this situation, there is a reasonable doubt that the developed models could, obviously, be predictive for chemicals which, even if not included in the training set, are in some way structurally very similar to these compounds. To eliminate this doubt from our model, and to verify if it is applicable to ‘‘completely external’’ chemicals that do not participate not even in the similarity-based splitting, we split the experimental data input into three different sets: (a) a ‘‘completely external’’ evaluation set of 160 chemicals (25% of the total set), selected randomly by activity sampling from the data set ordered by the response value, taking every fourth chemical from the set, (b) a training set of 307 chemicals on which to redevelop the model, selected as usual by SOM (300 epochs, 10 × 10 map) (48% of the total set) and (c) a prediction set of 176 chemicals selected as above by SOM (27% of the total set). In this way set (a), that is based on activity sampling, spans the entire range of the experimental measurements and is numerically representative of the data set, but it cannot represent the entire structural space of the original dataset. Thus, with regard to structural information it is completely unbiased, though, compared to the training set, it could be structurally unbalanced, and there could be more chemicals that fall outside the model applicability domain.

Table 2
List of the QSAR models developed and commented on in this paper, with statistical parameters

No.	Molecular descriptors	No. obj. training	No obj. external set	No. var.	R^2	Q^2_{LOO}	Q^2_{BOOT}	R^2 pred.	RMSE training calculated	RMSE training CV	RMSE predicted	RMSE total
1	VED1, nHAcc, MAXDP, CIC0	93	550	4	0.82	0.80	0.79	0.79	0.523	0.554	0.560	0.559
2	VED1, nHAcc, MAXDP, CIC0	643		4	0.79	0.79	0.79		0.545	0.550		
3	VED1, nHAcc, MAXDP, CIC0	307	*a* = 160 *c* = 176	4	0.78	0.78	0.77	*a* = 0.77 *c* = 0.82	0.536	0.545	*a* = 0.574 *c* = 0.549	
4	VED1, nHAcc, MAXDP, CIC0	176	307	4	0.82	0.81	0.80	0.78	0.547	0.566	0.539	0.550
5	VED1, nHAcc, MAXDP, CIC0	160	307	4	0.80	0.78	0.78	0.75	0.534	0.552	0.572	0.565
6	VED1, nHAcc, MAXDP	93	550	3	0.80	0.78	0.78	0.76	0.546	0.573	0.590	0.588
7	Consensus model	93	550	3	0.82			0.81		0.531	0.532	0.532
8	K_{oc} WIN [16]	93	550		0.78			0.75	0.622		0.635	0.633
9	log K_{ow}	93	550	1	0.78	0.77	0.77	0.77	0.572	0.586	0.570	0.572
10	log S_w	93	550	1	0.81	0.80	0.79	0.78	0.542	0.556	0.561	0.560
11	ARR-log K_{ow}	93	550	2	0.82	0.80	0.80	0.80	0.528	0.547	0.533	0.535
12	ARR-log S_w	93	550	2	0.82	0.81	0.81	0.80	0.519	0.537	0.538	0.538

Also redeveloping the model on this wider training set of 307 chemicals, the variables, selected by GA in the best OLS model (no. 3 in Table 2), are VED1, nHAcc, MAXDP and CIC0, as in the first proposed model (1). This again highlights that the structural information captured by these variables, and condensed in the multivariate model based on them, is really informative of the molecular features related to the soil partition properties, also in a wider and structurally very heterogeneous set of chemicals. As expected, predictive performances on set (a) are the worst, this set being structurally more ‘‘dissimilar’’ than the training set, but even in this case predictivity can still be regarded as good. Analogously, an analysis of the chemical domain of the two external sets (a) and (b) highlights that, the model has more chemicals outside the domain in the evaluation set (a) than in the prediction set (c) (three and zero, respectively).

From Table 2 it is also possible to verify, using the reported statistical parameters, that good, externally predictive models (nos. 4 and 5) are obtained on the same variables, even if scrambling is performed between the training and the different prediction/evaluation sets.

3.7. Full model

It is interesting to note that when genetic algorithms are applied to all the DRAGON descriptors with correlation less than 0.97 (1079 descriptors) in the modelling of all the available chemicals (643), a population of good and similarly predictive four-dimensional models is also obtained. MAXDP and NHAcc, already found relevant in our previous modelling of K_{oc} of pesticides [28], are present in all the models and always negative in sign, giving incontrovertible proof that these two descriptors are highly informative also of the soil partitioning properties of this wide set of heterogeneous chemicals. Features of size are captured in this population of models by different typologies of descriptors (topological, information indexes, etc.).

A final full model (model no. 2 in Table 2) developed on all the available data can be proposed, having verified that the four above selected descriptors are also well able to predict completely new chemicals in each verified splitting.

The equation and the statistical parameters of the full model are:

$$\log K_{\mathrm{oc}} = -1.92(\pm 0.11) + 2.07(\pm 0.06)\mathrm{VED1} - 0.31(\pm 0.01)\mathrm{nHAcc} - 0.31(\pm 0.02)\mathrm{MAXDP} - 0.39(\pm 0.05)\mathrm{CIC0} \quad (2)$$

$$n = 643 \quad R^2 = 0.79 \quad Q^2 = 0.79 \quad Q^2_{\mathrm{BOOT}} = 0.79$$

$$s = 0.547 \quad \mathrm{RMSE} = 0.545 \quad \mathrm{RMSEP_{LOO}} = 0.550$$

In this full model (model 2) there are only three response outliers (2,3,5-trimethylphenol (358), 2,6-dichlorobenzamide (427), predicted as the worst, and toxaphene (499)), compared to ten outliers in Model 1, and two structurally influential chemicals (with high leverage value), metasulfron methyl (628) and thiameturon methyl (637), compared to three in model 1. Structural information obtained when the domain of chemicals was enlarged during model development allowed the improvement of the modeling of seven chemicals (predicted badly by model 1) and allowed the inclusion of chlordecone (258) (high leverage in model 1) in the structural domain. Thus the great majority of chemicals were calculated correctly and are within the applicability domain of the proposed full model (model 2).

3.8. Consensus modeling

As mentioned above, the application of the genetic algorithm-variable subset selection procedure provides a large set of possible models with nearly equivalent predictive performance. The models are based on a variety of descriptors, reflecting the different aspects of molecular structure. Any individual QSAR model may overemphasize some aspects, underestimate others, or completely ignore many important features. Thus, it seems reasonable that a consensus QSAR model, which can be derived by calculating an average for representative individual models based on different descriptors, might provide better predicted data than the majority of individual models, and might, overall, take into account the more peculiar aspects of some particular structure [7,48–51].

For the comparison of different QSAR models it is useful to examine their variability in predicting responses [7,50]. The choice of the best and more representative models from among the 48 models with $Q^2_{\mathrm{ext}} > 75\%$ was done by principal component analysis (PCA) and multidimensional scaling (MDS) of the model residuals. The most different models (e.g. with dissimilar residual profiles) are far apart in the multivariate graphs, while similar models (e.g. giving similar predicted values) are clustered and redundant. On the basis of the different structural descriptions, an average/consensus model can be derived from dispersed models corresponding to different prediction schemes.

A consensus prediction for $\log K_{\mathrm{oc}}$ is calculated by averaging the predicted values from 10 individual models (listed in Table 3), selected from the three-dimensional model population obtained on the SOM split (93 chemicals in the training set and 550 in the prediction set) by applying the above described procedure; this is obviously an arbitrary choice and

Table 3
Models used for the consensus model and their statistical parameters

Descriptors	Q^2	R^2	SDEP	SDEC	Q^2_{BOOT}	Q^2_{ext}
VED1, nHAcc, MAXDP	0.78	0.80	0.573	0.547	0.78	0.76
VED1, IC2, nHAcc	0.75	0.77	0.618	0.589	0.74	0.72
VED1, R4u, nHAcc	0.74	0.77	0.622	0.595	0.74	0.70
VED1, nHAcc, T(O..F)	0.74	0.76	0.624	0.597	0.68	0.69
Mv, VED1, nHAcc	0.74	0.76	0.630	0.603	0.73	0.70
VED1, GGI5, nHAcc	0.74	0.76	0.632	0.598	0.73	0.69
H3p, MAXDP, HOMT	0.73	0.76	0.633	0.606	0.73	0.70
VED1, MAXDP, nCb-	0.71	0.74	0.660	0.629	0.71	0.70
Mv, VED1, MAXDP	0.69	0.72	0.682	0.650	0.69	0.71
H3p, VED1, MAXDP	0.68	0.71	0.696	0.663	0.68	0.69
Consensus model		0.82		0.531		0.80

other choices could be possible. It is interesting to note that the model population is highly homogeneous, providing additional proof of the supremacy of some, already highlighted, descriptors in the modeling of K_{oc}.

In addition to the descriptors of the best model (VED1, nHAcc, MAXDP) present in most of the models, the descriptors selected in the consensus modeling are those most abundant in the whole GA-population and thus the most relevant, in their alternative combinations, in modeling the response.

A comparison of the statistical parameters calculated for the consensus model with those obtained from individual models highlights the superiority of the consensus model (Table 3). Fitting ability ($R^2 = 0.82$) and predictive ability verified by $R^2_{ext} = 0.80$ are better than for any individual model.

Fig. 3 shows the regression line plotting the values predicted by the consensus model and the experimental data. Note that only two chemicals have a range value (Δ between maximum and minimum predicted value by the 10 models) higher than 4 log units: mirex and fluazipop-butyl. We verified that these two chemicals are those well beyond the chemical domain of the consensus model, thus their predicted data, being extrapolated by the model, must be considered unreliable, even if now apparently well predicted. In addition, the experimental data of these two chemicals are considered, according to the literature [32,52], to be not well defined and uncertain.

To verify the chemical domain of the new consensus model and the distribution of the studied chemicals in this new multidimensional space, the chemicals are plotted in a principal components 3D-graph, obtained by PCA of all the molecular descriptors selected in the models of Table 3, that are used in the consensus model (Fig. 4).

From this PCA plot it is evident that, as expected, chemicals with smaller leverage values (less influential) are grouped more towards the central part of the graph, while the more influential chemicals (with higher leverage values) are more isolated and lie mainly at the border of the molecular descriptors domain.

It is also interesting to note that all the models selected for consensus modeling predicted some chemicals with practically the same K_{oc}, while for other chemicals K_{oc} was predicted differently by each selected model (see the Δ value among the predictions of the selected models in Table S1 in the supporting information, Δ range: 0.19–4.31 log units). In the former case the chemicals can be defined as ‘‘prediction safe’’, while in the latter they can be defined as ‘‘prediction sensitive’’, being selectively related to the structural information included in the different models. The range of predicted values between the models can be used to highlight ‘‘prediction sensitive’’ chemicals and therefore those of greatest concern (with higher Δ). If K_{oc} is detected experimentally, or there is future verification, then ‘‘prediction sensitive’’ compounds could be useful to choose between several possible QSAR models.

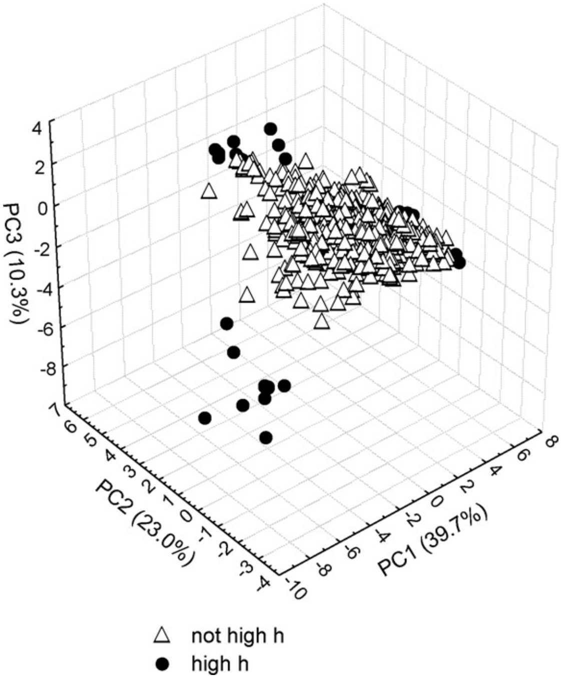

Fig. 4. Principal component analysis on the selected molecular descriptors for the consensus model (PC1–PC2: EV% = 73%). Influential chemicals with higher mean leverage values (h) are highlighted differently (●).

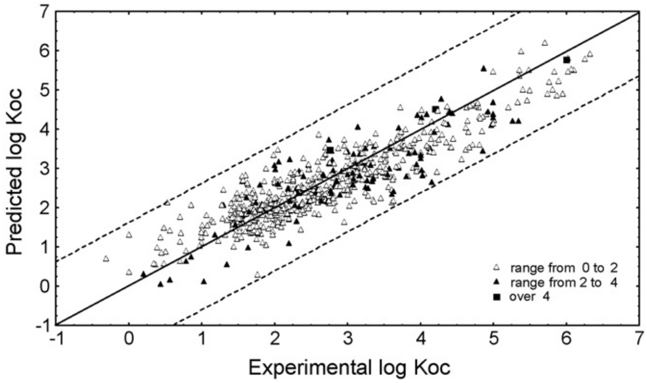

Fig. 3. Regression line of the consensus model (K-ANN splitting). The log K_{oc} values for the different range (Δ between maximum and minimum predicted value by the 10 models) are labelled differently. The dotted lines indicate the 3σ interval.

Table 2 allows the statistical parameters of the models developed in this work (nos. 1–7) to be compared with the K_{oc} WIN model (no. 8) (from EPISuite) [16], which appears the poorer predictive model (lower value of R^2 and higher RMSE for the prediction set chemicals).

In addition, as it is well known [1,9,32] that log K_{ow} and log S_w are, in general, good descriptors of K_{oc}, we applied these parameters to our data set modeling, and verified the good quality of the obtained models (nos. 9 and 10 in Table 2), comparable with the models based on theoretical descriptors proposed in this paper.

It appears that models based on these physico-chemical properties are simpler, but it is important to remember that the experimental data of these properties are not always available. Furthermore, their predicted data could be subject to high variability due to the selected QSAR calculation method and, as we have already pointed out [53], they look like single descriptors, but are, in reality, the condensation of all the structural information represented by the fragments and the

Table 4
Comparison of mean residuals of the different QSAR models

No.	Models	Training	% with residual <0.5	% with residual <1	Mean residual	Mean residual training	Mean residual validation	Max residual	No. comp. res. >1.5
1	VED1, nHAcc, MAXDP, CIC0	93	66	92	0.430	0.420	0.431	1.87	12
2	Consensus (10 modelli) 3D	93	66	92	0.423	0.423	0.423	1.60	1
3	K_{oc} WIN (EPISuite)		70	89	0.475	0.438	0.452	2.94	34
4	log K_{ow}	93	67	91	0.434	0.470	0.428	1.92	13
5	log S_w	93	66	92	0.425	0.423	0.425	2.08	11
6	ARR-log K_{ow}	93	67	94	0.415	0.437	0.411	2.34	5
7	ARR-log S_w	93	69	94	0.410	0.411	0.409	1.99	9

correction factors used for their determination. In any case, we verified that adding a theoretical descriptor of aromaticity (ARR) to these parameters improves the corresponding models that are then only slightly worse than the consensus model, which always remains the best. It is also interesting to verify that the information provided by ARR also affects the chemical domain of the two models based on physico-chemical properties. In fact, while the log K_{ow} model (no. 9 in Table 2) has 22 chemicals outside the domain (7 response outliers and 19 structural outliers with high leverage, 3 chemicals being both), the model based on log K_{ow} plus ARR (no. 11 in Table 2) has 6 response outliers and the number of structural outliers drops to only 9 compounds. Similarly, 26 chemicals lie outside the log S_w model (no. 10 in Table 2) domain, these decreasing to 14 in the log S_w plus ARR model (no. 12 in Table 2). Also in this case the addition of the ARR descriptor leads to the very evident reduction of structural outliers, from 20 to only 8 chemicals. The majority of the outliers in these models are persistent organic pollutants (POPs), i.e. PCB, PAH, and some pesticides (always mirex). It is important to stress here that all our models (nos. 1–7 proposed in this paper in Table 2) have significantly fewer outliers, particularly for structure, than the compared models.

3.9. Comparison of residuals

Additional interesting results on the different predictive quality of some models can be obtained by comparing the mean residuals between the experimental and predicted log K_{oc} values using different models (Table 4). In general, it is important to note that all the reported models have mean residuals very similar to the calculated standard deviation of the experimental values of log K_{oc}, namely about 0.44 [44]. Again the best predictive model is the consensus model (no. 2 in Table 4), which has mean residuals that are practically the same in the training and prediction chemicals, indicating stability and generalizability for prediction; this model also has the lowest maximum residual (1.60) and only 1 chemical (2,6-Dichlorobenzamide (427)) with residual >1.5 log units. The worst model is K_{oc} WIN (no. 3) with the highest mean and maximum residuals, and the highest number of chemicals with residual >1.5 (34 chemicals). The model of Eq. (1) and log K_{ow}- or log S_w-based models (nos. 4 and 5) are very similar in their performances. Once again there was confirmation of model improvement by adding the aromaticity descriptor (ARR): very few compounds have residuals >1.5. The apparently best model of Tao et al. [8] (no. 3 in Table 1) has the lowest mean residual for training chemicals (0.358) but the highest for prediction set chemicals (0.468), thus it can be considered the best in fitting but the worst in prediction; in fact the use of 86 fragments and 19 structural factors as descriptors in the models leads one to fear an overfitted model.

3.10. K_{oc} for POPs

Baker et al. [52,54] demonstrated that log K_{ow} is not a strong predictor of log K_{oc} for chemicals with log K_{ow} >6–7, in particular for persistent organic pollutants (POPs), while theoretical descriptors like, for instance, connectivity indexes, have good correlation with this soil sorption coefficient. We verified the applicability of our models (no. 1 and consensus no. 7 in Table 2) for the K_{oc} prediction of some POPs, with a log K_{ow} value ranging from 7 to 13 (for instance, polybromodiphenylethers (PBDE) and some fluorinated compounds): the preliminary results are very satisfactory for polybromocompounds, but not for fluorinated. The details of this study will be dealt with in a later paper (presently in preparation).

4. Conclusions

In summary, multivariate linear QSAR models have been proposed to predict the soil sorption partition coefficient of a comprehensive set of heterogeneous chemicals. The OLS models are developed by a genetic algorithm selection of theoretical molecular descriptors from among a wide set of theoretical molecular descriptors. The proposed models have good stability, robustness and predictivity when verified by internal validation (cross-validation by LOO and Bootstrap) and also external validation. To have ‘‘external’’ chemicals not used in the model development, the original data set is split into training and prediction sets by the SOM approach, but also randomly in order to avoid the bias of structural similarity. The chemical applicability domain of the studied models and the reliability of the predictions are always verified by the leverage approach.

The selected molecular descriptors have a clear mechanistic meaning: they are related to both the molecular size of the chemical and its electronic features relevant to soil partitioning as well as to the chemical's ability to form hydrogen bonds with water. A selection of different models from the GA-model population allows the proposal of Consensus predictions that, compared with published models and EPISuite predictions, are always among the best. The proposed models have been proved to fulfil the fundamental points set down by OECD principles [10,11] for regulatory QSAR acceptability and could be used reliably as scientifically valid models in the REACH program.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2006.06.005.

References

[1] A. Sabljic, H. Güsten, H. Verhaar, J. Hermens, QSAR modeling of soil sorption. improvements and systematics of log K_{oc} vs. log K_{ow} correlations, Chemosphere 31 (1995) 4489–4514.

[2] B.M. Gawlik, N. Sotiriou, E.A. Feicht, S. Schulte-Hostede, A. Kettrup, Alternatives for the determination of the soil adsorption coefficient, Koc, of non-ionic organic compounds—a review, Chemosphere 34 (1997) 2525–2551.

[3] W.J. Doucette, Quantitative structure-activity relationships for predicting soil-sediment sorption coefficients for organic chemicals, Environ. Toxicol. Chem. 22 (2003) 1771–1788.

[4] A. Tropsha, P. Gramatica, V.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) 69–76.

[5] L. Eriksson, E. Johansson, M. Müller, S. Wold, On the selection of the training set in environmental QSAR analysis when compounds are clustered, J. Chemom. 14 (2000) 599–616.

[6] A. Golbraikh, M. Shen, Z. Xiao, Y.-D. Xiao, K.-H. Lee, A. Tropsha, Rational selection of training sets for the development of validated QSAR models, J. Comput. Aided Mol. Des. 17 (2003) 241–253.

[7] 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.

[8] S. Tao, H. Piao, R. Dawson, X. Lu, H. Hu, Estimation of organic carbon normalized sorption coefficient (KOC) for soils using the fragment constant method, Environ. Sci. Technol. 33 (1999) 2719–2725.

[9] J. Huuskonen, Prediction of soil sorption coefficient of a diverse set of organic chemicals from molecular structure, J. Chem. Inf. Comput. Sci. 43 (2003) 1457–1462.

[10] OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationship models http://www.oecd.org/document/23/0,2340,en_2649_201185_33957015_1_1_1_1,00.html (accessed 28 April 2006).

[11] Web site of the QSAR Group, Joint Research Center, European Chemical Bureau, Ispra, Italy, http://ecb.jrc.it/QSAR/ (accessed 28 April 2006).

[12] P. Gramatica, Evaluation of Different Statistical Approaches to the Validation of Quantitative Structure–Activity Relationships, ECVAM, JRC, Ispra, 2004 (http://ecb.jrc.it/QSAR/under QSARs/documents/public access).

[13] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, DRAGON—software for the calculation of molecular descriptors, in: Version 5.3 for Windows, 2005.

[14] HYPERCHEM, Release 7.03 forWindows, 2002. in: Molecular Modeling System, Hypercube, Inc., Gainesville, FL, USA.

[15] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, Germany, 2000.

[16] EPI Suite ver.3.12, 2000. U.S.EPA: http://www.epa.gov/opptintr/exposure/docs/EPISuitedl.htm.

[17] MOBY DIGS—software for multilinear regression analysis and variable subset selection by genetic algorithm, in: Version 1 for Windows, 2005, Talete srl, Milan, Italy.

[18] R. Leardi, R. Boggia, M. Terrile, Genetic algorithms as a strategy for feature selection, J. Chemom. 6 (1992) 267–281.

[19] L. Eriksson, J. Jaworska, A. Worth, M. Cronin, R.M. McDowell, P. Gramatica, Methods for reliability, uncertainty assessment, and applicability evaluations of regression based and classification QSARs, Environ. Health Perspect. 111 (10) (2003) 1361–1375.

[20] SCAN—software for chemometric analysis, Release 1.1 for Windows 1995, Minitab, USA.

[21] A.C. Atkinson, Plots, Transformations and Regression, Clarendon Press, Oxford, 1985.

[22] J. Zupan, M. Novic, I. Ruisánchez, Kohonen and counter propagation artificial neural networks in analytical chemistry, Chemom. Int. Lab. Syst. 38 (1997) 1–23.

[23] J. Gasteiger, J. Zupan, Neural networks in chemistry, Angew. Chem. Int. Ed. Engl. 32 (1993) 503–527.

[24] KOALA-Software for Kohonen Artificial Neural Networks, by R., Todeschini, V., Consonni, A., Mauri, Rel. 1.0 for Windows, 2001. Milan, Italy.

[25] L.M. Shi, H. Fang, W. Tong, J. Wu, R. Perkins, R.M. Blair, W.S. Branham, S.L. Dial, C.L. Moland, D.M. Sheehan, QSAR Models using a large diverse set of estrogens, J. Chem. Inf. Comput. Sci. 41 (2001) 186–195.

[26] STATISTICA, Rel. 6 for Windows, 2001, StatSoft, Inc., USA.

[27] S. Tao, X. Liu, Estimation of organic carbon normalized sorption coefficient (Koc) for soil by topological indices and polarity factors, Chemosphere 39 (1999) 2019–2034.

[28] 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.

[29] P.L. Andersson, U. Maran, D. Fara, M. Karelson, J.L.M. Hermens, General and class specific methods for prediction of soil sorption using various physicochemical descriptors, J. Chem. Inf. Comput. Sci. 42 (2002) 1450–1459.

[30] J. Huuskonen, Prediction of soil sorption coefficient of organic pesticides from the atom-type electrotopological state indices, Environ. Toxicol. Chem. 22 (2003) 816–820.

[31] E.J. Delgrado, J.B. Alderete, A.J. Gonzalo, A simple QSPR model for predicting soil sorption coefficients of polar and nonpolar organic compounds from molecular formula, J. Chem. Inf. Comput. Sci. 43 (2003) 1928–1932.

[32] I. Kahn, D. Fara, M. Karelson, U. Maran, P.L. Andersson, QSPR treatment of the soil sorption coefficients of organic pollutants, J. Chem. Inf. Model. 45 (2005) 94–105.

[33] S. Wold, L. Eriksson, Chemometric Methods in Molecular Design, VCH, Germany, 1995.

[34] A. Golbraikh, A. Tropsha, Beware of q^2! J. Mol. Graph. Model. 20 (2002) 269–276.

[35] H. Kubinyi, Good practice in QSAR model validation, Am. Chem. Soc. 227 (2004) 1027 (Abstracts).

[36] T. Oberg, A QSAR for baseline toxicity: validation, domain of application and prediction, Chem. Res. Toxicol. 17 (2004) 1630–1637.

[37] T. Oberg, A QSAR for the hydroxyl radical reaction rate constant: validation, domain of application and prediction, Atm. Environ. 39 (2005) 2189–2200.

[38] G.G. Cash, B. Anderson, K. Mayo, S. Bogaczyk, J. Tunkel, Predicting genotoxicity of aromatic and heteroaromatic amines using eelctrotopological state indices, Mutat. Res. 585 (2005) 170–183.

[39] T. Pötter, H. Matter, Random or rational design? Evolution of diverse compound subsets from chemical structure data set, J. Med. Chem. 41 (1998) 478–488.

[40] F.R. Burden, D.A. Winkler, Robust QSAR models using Bayesian regularized neural networks, J. Med. Chem. 42 (1999) 3183–3187.

[41] F.R. Burden, M.G. Ford, D.C. Whitley, D.A. Winkler, Use of automatic relevance determination in QSAR studies using Bayesian Regularized neural networks, J. Chem. Inf. Comput. Sci. 40 (2000) 1423–1430.

[42] A. Golbraikh, A. Tropsha, Predictive QSAR modeling based on diversity sampling of experimental datasets for the training set selection, J. Comput. Aided Mol. Des. 16 (2002) 357–369.

[43] L. Eriksson, E. Johansson, Multivariate design and modeling in QSAR, Tutorial Chemom. Int. Lab. Syst. 34 (1996) 1–19.

[44] H. Lohninger, Estimation of soil partition coefficients of pesticides from their chemical structure, Chemosphere 29 (1994) 1611–1626.

[45] A.T. Balaban, D. Ciubotariu, M. Medeleanu, Topological indices and real number vertex invariants based on graph eigenvalues or eigenvectors, J. Chem. Inf. Comput. Sci. 31 (1991) 517–523.

[46] D. Bonchev, Information Theoretic Indices for Characterization of Chemical Structures, RSP/Wiley, Chichetser (UK), 1983.

[47] V.R. Magnuson, D.K. Harriss, S.C. Basak, in: R.B. King (Ed.), Studies in Physical and Theoretical Chemistry, Elsevier, Amsterdam, The Netherlands, 1983, pp. 178–191.

[48] J.R. Votano, M. Parham, L.H. Hall, L.B. Kier, S. Oloff, A. Tropsha, Q.A. Xie, W. Tong, Three new consensus QSAR models for the prediction of Ames genotoxicity, Mutagenesis 19 (2004) 365–377.

[49] A.H. Asikainen, J. Ruuskanen, K.A. Tuppurainen, Consensus kNN QSAR: a versatile method for predicting the estrogenic activity of organic compounds in silico. A comparative study with five estrogen receptors and a large, diverse set of ligands, Environ. Sci. Technol. 38 (2004) 6724–6729.

[50] J.J. Sutherland, D.F. Weaver, Development of quantitative structure–activity relationships and classification models for anticonvulsant activity of hydantoin analogues, J. Chem. Inf. Comput. Sci. 43 (2003) 1028–1036.

[51] N. Baurin, J.C. Mozziconacci, E. Arnoult, P. Chavatte, C. Marot, L. Morin-Allory, 2D QSAR consensus prediction for high-throughput virtual screening. An application to COX-2 inhibition modeling and screening of the NCI database, J. Chem. Inf. Comput. Sci. 44 (2004) 276–285.

[52] J.R. Baker, J.R. Mihelcic, A. Sabljic, Reliable QSAR for estimating Koc for persistent organic pollutants: correlation with molecular connectivity indices, Chemosphere 45 (2001) 213–221.

[53] P. Gramatica, F. Villa, E. Papa, Statistically validated QSARs and theoretical descriptors for the modelling of the aquatic toxicity of organic chemicals in *Pimephales promelas* (Fathead Minnow), J. Chem. Inf. Model. 45 (2005) 1256–1266.

[54] J.R. Baker, J.R. Mihelcic, E. Shea, Estimating Koc for persistent organic pollutants: limitations of correlations with Koc, Chemosphere 41 (2000) 813–817.