RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2015, **5**, 7547

Coloured chemical image-based models for the prediction of soil sorption of herbicides

Mirlaine R. Freitas, Stephen J. Barigye and Matheus P. Freitas*

Received 9th October 2014
Accepted 16th December 2014

DOI: 10.1039/c4ra12070a

www.rsc.org/advances

Herbicides with high soil sorption profiles constitute important organic pollutants leading to detrimental environmental effects, particularly due to prolonged use. Soil sorption is described in terms of log K_{OC}, the logarithm of the soil/water partition coefficient normalized to organic carbon. This work reports the use of molecular drawings to generate molecular descriptors, which are posteriorly correlated with the log K_{OC} values of a series of herbicides. These images are two-dimensional projections of chemical structures, with their atom sizes drawn to be proportional to the corresponding van der Waals radii and each chemical element assigned a different colour to distinguish atom types. The progressive changes in the molecular structures explain the variance in the corresponding soil sorption. Unlike previous QSPR studies on soil sorption, the series of herbicides employed in the present study included different classes of compounds (carboxylic acids, ethers, phenols, amines, amides and carbamates) guaranteeing a diverse chemical structural space. The obtained Partial Least Squares (PLS) and Multiple Linear Regression (MLR) based models for the log K_{OC} values were found to be robust and with high predictive power. Mechanistic interpretation of the effect of different substituents (bonded to the common structural moiety in the herbicides series) on the log K_{OC} values was performed yielding interesting results. These findings allow greater understanding of the chemical groups (or structural characteristics) responsible for high/low soil sorption, which in turn provides key leads for structural optimization to yield environmentally friendly and equally effective herbicides.

Introduction

Many herbicides have long been known to be persistent organic pollutants, resulting in numerous toxic environmental effects. The persistence of organic compounds in soils or sediments is associated with their soil sorption coefficients described as log K_{OC}, the logarithm of the soil/water–organic carbon partition coefficient, which refers to the affinity of organics to soil or sediment particles. Despite the successful application of molecular topology models to estimate log K_{OC} values for a variety of chemicals,[1] a significant number of QSPR (quantitative structure-property relationship) models used to estimate log K_{OC} values are based on a single parameter, hydrophobicity, described in terms of the logarithm of the octanol/water partition coefficient, log P. Such a relationship does not always achieve good predictions for log K_{OC}, at least for many classes of herbicides, such as amides and triazines.[2] The inclusion of simple molecular descriptors (*e.g.* molecular weight and volume) to the univariate log P model to give a three-parameter modelling of soil sorption has been shown to improve the predictive power of the QSPR models, but only for single classes of herbicides, namely acetanilides and triazines, separately.[3] A more general QSPR model for the simultaneous prediction of the soil sorption of multi-functional herbicide compounds would be useful for providing a wide applications domain. This can be achieved using molecular descriptors that strongly encode the log K_{OC}, such as a two-dimensional molecular shape (since three-dimensional molecular structures have not been shown to play a fundamental role in some QSAR/QSPR studies[4]), atomic size and other atomic properties. The MIA-QSPR (multivariate image analysis applied to QSPR) method has been successfully used in the modelling of a variety of chemical and biological properties. In this approach, the molecular structure images (viewed as an array of pixels) are aligned together with respect to the basic congruent scaffold to yield a multivariate image (MVI). The underlying reasoning for this approach is that the variation in the chemical and biological activity of compounds is a function of the non-congruent groups (or atom-types) in a given dataset.

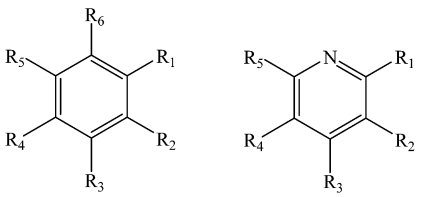
Previous studies using the MIA-QSPR method considered black and white molecular images.[5] However, with the aim of incorporating useful chemical information, extensions of this approach have been implemented using colour schemes carefully defined to correlate with atomic properties of interest which may be chemical, physical, physicochemical or biological, such as electronegativity, atomic hydrophobicity, polarisability, atomic refractivity, covalent radius, *etc.* or using a

Department of Chemistry, Federal University of Lavras, 37200-000, Lavras, MG, Brazil. E-mail: matheus@dqi.ufla.br

random colour scheme to simply distinguish dissimilar atoms.[6] These schemes are described using the RGB system, in which each individual channel (red, green and blue) varies in intensity from 0 to 255 (thus, the combination of the three channels to give the broad spectrum of colours which varies from 0 – black to 765 – white). The incorporation of colour schemes to the MIA-QSAR method yields the so-called aug-MIA-QSPR (acronym for augmented multivariate image analysis applied to QSPR) method. For instance, with the aim of solely differentiating unlike atoms, colours were randomly allocated to the atoms but their sizes made to be proportional to the corresponding Van der Waals radii. This is known as the aug-MIA-QSPR$_{default}$ approach and has been shown to adequately predict the phytotoxicity of some benzoxazinone herbicides and derivatives.[7] However, in this study, the atom colours were spotlighted to give a three-dimensional insight, causing a variation in the pixel values within the radius of the same atom. The presence of different pixel values for the same atom obliterates the one-to-one correspondence of the colours to the atoms, thus making the interpretation of the ensuing models difficult. It is anticipated that the use of atoms with solid colours should give more interpretable models. On the other hand, in order to make the atomic colour proportional to the electronegativity (which in turn affects bond polarity and, consequently, intermolecular interactions) of carbon, oxygen and fluorine atoms, RGB combinations were selected with pixel values of 250, 350 and 400, respectively, which correlated with the Pauling electronegativities for these atoms, *i.e.* 2.5, 3.5 and 4.0, respectively. This scheme is denominated as the aug-MIA-QSPR$_{colour\text{-}\varepsilon}$.

The aim of the present manuscript is to study the soil sorption profiles of a series of multi-functional herbicides containing only an aromatic ring as the congruent moiety, using the aug-MIA-QSPR$_{default}$ and aug-MIA-QSPR$_{colour\text{-}\varepsilon}$ approaches, respectively. To this end, the statistical methods Partial Least

Table 1 Data set of herbicides with their experimental log K_{OC} values



1, 2, 4-29 **3**

#	Name	R_1	R_2	R_3	R_4	R_5	R_6	log K_{OC}
1	Chlorambem	COOH	Cl	NH_2	H	Cl	H	1.32
2	Dicamba	COOH	OCH_3	Cl	H	H	Cl	0.08
3	Picloram	COOH	Cl	NH_2	Cl	Cl		1.23
4	Bifenox	O(3-COOH, 4-NO_2-Ph)	Cl	H	Cl	H	H	1.86
5	Fluorodifen	O(2-NO_2, 4-CF_3-Ph)	H	H	NO_2	H	H	3.13
6	Dinoseb	OH	NO_2	H	NO_2	H	*sec*-Bu	2.85
7	PCP	OH	Cl	Cl	Cl	Cl	Cl	3.73
8	2,4-D	OCH_2COOH	Cl	H	Cl	H	H	2.20
9	2,4-DB	$OCH_2CH_2CH_2COOH$	Cl	H	Cl	H	H	2.64
10	Dichlorprop	$OCH(CH_3)COOH$	Cl	H	Cl	H	H	3.00
11	MCPA	OCH_2COOH	Me	H	Cl	H	H	2.05
12	MCPB	$OCH_2CH_2CH_2COOH$	Me	H	Cl	H	H	2.73
13	Mecoprop	$OCH(CH_3)COOH$	Me	H	Cl	H	H	1.70
14	2,4,5-T	OCH_2COOH	Cl	H	Cl	Cl	H	1.72
15	Benefin	N(Et) (Bu)	NO_2	H	CF_3	H	NO_2	3.95
16	Butralin	NH(*sec*-Bu)	NO_2	H	*t*-Bu	H	NO_2	3.75
17	Dinitramine	$N(Et)_2$	NO_2	NH_2	CF_3	H	NO_2	3.60
18	Fluchloralin	$N(CH_2CH_2Cl)$ (Pr)	NO_2	H	CF_3	H	NO_2	3.50
19	Isopropalin	$N(Pr)_2$	NO_2	H	*t*-Bu	H	NO_2	4.00
20	Oryzalin	$N(Pr)_2$	NO_2	H	SO_2NH_2	H	NO_2	2.78
21	Trifluralin	$N(Pr)_2$	NO_2	H	CF_3	H	NO_2	4.14
22	Alachlor	$N(CH_2CH_2Cl)$ (CH_2OCH_3)	Et	H	CF_3	H	Et	2.23
23	Butachlor	$N(CH_2CH_2Cl)$ (CH_2OBu)	Et	H	H	H	Et	2.80
24	Metolachlor	$N(COCH_2Cl)$ $(CH(CH_3)CH_2OCH_3)$	Et	H	H	H	Me	2.26
25	Propachlor	$N(COCH_2Cl)$(i-Pr)	H	H	H	H	H	1.90
26	Propanil	NHCOEt	H	Cl	Cl	H	H	2.17
27	Barban	$NHCOOCCCH_2Cl$	H	Cl	H	H	H	2.66
28	Chloropham	$NHCOOCH(CH_3)_2$	H	Cl	H	H	H	2.82
29	Propham	$NHCOOCH(CH_3)_2$	H	H	H	H	H	1.71

Squares (PLS) and Multiple Linear Regression (MLR) are used. Posteriorly, a mechanistic interpretation of the MLR-based models in terms of the influence of the different substituents on the soil sorption profiles of molecular structures is performed. The experimental log K_{OC} values used in the modelling were obtained from the literature.[8] It is hoped that this study will provide key insight on the structural characteristics to be taken into account for the design of novel, environmentally friendly herbicides that are effective against undesired plants.

Materials and methods

Despite the fact that there are several factors that may affect the sorption of organic compounds in natural sorbent/water systems,[9] the log K_{OC} values used in this work are an average of experimental data obtained from the literature,[8] which do not vary significantly. A series of 29 compounds are given in Table 1 and the data matrix used for regression with the response variable (log K_{OC}) was obtained as follows. Each chemical image was built using the GaussView program,[10] in which the congruent moiety, an aromatic ring, occupies the same coordinates in a two-dimensional space as the alignment procedure requires. Fig. 1 shows the superimposed images to provide insight on the common nuclei and variant moieties that explain the variance in the log K_{OC} values for the compounds. The atoms in the molecules were coloured according to the default configuration of the program and the numerical value for the colour of each atom (according to the RGB model) made to be proportional to the corresponding Pauling's electronegativity value (Table 2). Each chemical image was saved as a bitmap workspace of 1500 × 900 pixels dimension and then grouped to generate a 29 × 1500 × 900 three-way array, which was unfolded to a 29 × 1 350 000 matrix. This matrix was reduced to *ca.* 29 × 107 000 by eliminating columns without variance, corresponding to congruent moieties of the molecules or blank spaces.

The reduced descriptors matrix was regressed against the log K_{OC} values using PLS regression. The theoretical structure of this method has been extensively explained in the literature.[11] The key advantage of PLS as a multivariate statistical tool is that it permits the analysis of high dimensionality data matrices of strongly collinear variables, through the derivation of orthogonal projections of the original data matrix. Nonetheless, mechanistic interpretations of PLS models are more complex compared to models built with the original variables. In this sense, the MLR-based models are built as well particularly with the aim of elucidating the contribution of different substituents to the modeled property. However, the MIA-QSAR method yields high dimensionality matrices, given that each pixel coordinate in the MVI represents a variable (instances are pixel values in the same coordinate for the images that constitute the MVI), which makes the exhaustive exploration of all linear combinations for MLR models extremely tedious. In this sense, two dimensionality reduction strategies were adapted: (1) an information-theoretic filter based on Shannon's entropy (SE) was applied. This is a variability measure that quantifies the distribution of instances in discrete intervals, in the sense that information-rich variables yield instance distributions over the entire variable range (all discrete intervals) and thus have higher SE values, while redundant variables have low SE. For details see ref. 12 and 13. In this sense, variables presenting less than 10% of the maximum SE [$SE_{max} = \log_2$(number of discrete intervals)] were discarded. (2) Posteriorly, an *x*/*x* correlation filter was applied to the data matrix, in that for a pair of variables with a correlation coefficient >0.96, one is selected. With these two filters, 354 molecular descriptors (MDs) were retained. Posteriorly, regressions with a 3 variable model size were built using the statistical method Multiple Linear Regression coupled with Genetic Algorithm (MLR-GA) as the search strategy, and the Leave-One-Out Cross validation (LOO_{CV}) parameter as the optimization function. Consequently, the best models according to the LOO_{CV} parameter were retained for later validation.

The prediction ability of the aug-MIA-QSPR models was evaluated using external validation statistical parameters

Table 2 Coloured pixels numerically described according to the RGB (red-green-blue) system, for each atom within the series of molecules, using colours proportional to the Pauling's electronegativity (ε) scale and the default GaussView configuration

		Colours proportional to ε		Default colours	
Atom/type	ε	Colour	Pixel	Colour	Pixel
H	2.1	Charcoal	210	Silver	612
C	2.5	Teal	250	Gray	426
N	3.0	Persian blue	300	Blue	279
O	3.5	Scarlet	350	Red	229
F	4.0	Turquoise	400	Electric blue	688
S	2.5	Olive	250	Gold	493
Cl	3.0	Green	300	Green	289
Chemical bond	—	Light gray[a]	615	Light gray	615
Blank space	—	White	765	White	765

[a] Colour for chemical bond arbitrarily selected.

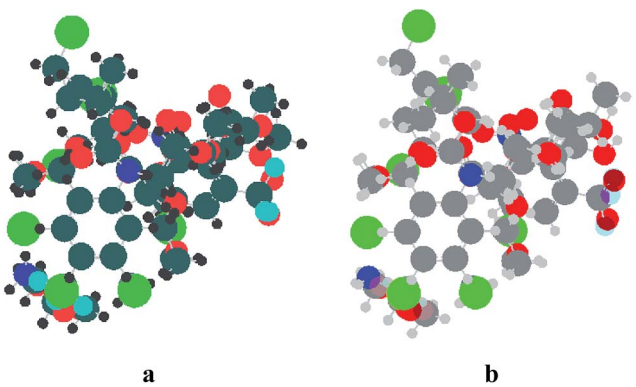


Fig. 1 Superimposed chemical structures used to generate descriptors for the aug-MIA-QSPR models. (a) Chemical structures with atoms coloured according to their Pauling's electronegativity values; (b) Chemical structures with atoms coloured according to the default style of the GaussView program.

Table 3 Statistical results of the aug-MIA-QSPR models

	PLS		MLR	
Parameter	aug-MIA-QSPR$_{\text{colour-}\varepsilon}$	aug-MIA-QSPR$_{\text{default}}$	aug-MIA-QSPR$_{\text{colour-}\varepsilon}$	aug-MIA-QSPR$_{\text{default}}$
LV	4	2		
r^2	0.95	0.76	0.84	0.79
RMSEC	0.18	0.41	0.32	0.46
q^2	0.60	0.59	0.76	0.67
RMSECV	0.54	0.53	0.39	0.59
r_{test}^2	0.68	0.74	0.68	0.71
RMSEP	0.67	0.63	0.52	0.59
r_{m}^2	0.61	0.67	0.63	0.69
$^{\text{c}}r_{\text{P}}^2$	0.32	0.56	0.76	0.71
Outliers	**2** and **7**	**2** and **7**	**1**, **2**, **3** and **7**	**18**, **20** and **22**
Test set	**8**, **10**, **11**, **12**, **21** and **28**	**8**, **10**, **11**, **12**, **21** and **28**	**10**, **15**, **20**, **23**, **26** and **28**	**3**, **4**, **10**, **11**, **17** and **23**

RMSEP (root mean square error of prediction) and r_{test}^2 (correlation coefficient for test). Additionally, *y*-randomization tests (mean of 10 repetitions) were carried out to attest the robustness of the models (statistically evaluated using RMSEC$_{\text{rand}}$ and r_{rand}^2). Other statistical parameters proposed by Roy *et al.*,[14–17] namely modified $r^2(r_{\text{m}}^2)$ and corrected penalized $r^2(^{\text{c}}r_{\text{P}}^2)$, were also used for validation purposes. While r_{m}^2 determines the proximity between the observed and predicted property for the test set (values $\geq$ 0.5 are recommended), the $^{\text{c}}r_{\text{P}}^2$ parameter gives insight into the statistical difference between r^2 and r_{rand}^2 (values $\geq$0.5 are recommended). Outlier diagnostics were performed on the basis of leverage analysis and studentized residuals. The treatment of the molecular images to generate the MDs for the aug-MIA-QSPR$_{\text{default}}$ and aug-MIA-QSPR$_{\text{colour-}\varepsilon}$ approaches, respectively, as well as the posterior modelling and validation procedures were performed using the Chemoface program[18] implemented using the Matlab platform.[19]

Results and discussion

In the first experiment, PLS-based models were built from the entire data set matrix, obtained according to the aug-MIA-QSPR$_{\text{colour-}\varepsilon}$ (Fig. 1a) and aug-MIA-QSPR$_{\text{default}}$ (Fig. 1b) approaches, respectively. In both cases, compounds **2** and **7** were found to be outliers and were therefore excluded. The external validation set was chosen using Kennard-Stone sampling and was composed of compounds **8**, **10**, **11**, **12**, **21** and **28**. The PLS models obtained with 4 latent variables (LVs) for the aug-MIA-QSPR$_{\text{colour-}\varepsilon}$ and 2 LVs for the aug-MIA-QSPR$_{\text{default}}$ approaches, respectively, were generally satisfactory (Table 3 and Fig. 2), despite the unsatisfactory $^{\text{c}}r_{\text{P}}^2$ for the first and the line shift for the test set in the experimental $\times$ predicted log K_{OC} plot, which does not meet the validation criteria established by Chirico and Gramatica.[20] The problem with $^{\text{c}}r_{\text{P}}^2$ suggests that the aug-MIA-QSPR$_{\text{colour-}\varepsilon}$ model has been overfitted. In addition, the PLS-based models are more difficult to interpret in terms of the influence of the variables on the modelled activity due to the large number of (collinear) descriptors employed.

In the second experiment, a search for MLR-based models for the log K_{OC} of the studied dataset was performed and the best regressions, according to the aforementioned validation statistical parameters, retained. The equations for the best MLR models according to the aug-MIA-QSPR$_{\text{colour-}\varepsilon}$ and aug-MIA-QSPR$_{\text{default}}$ approaches, respectively, are given below (see eqn (1) and (2)). From a statistical perspective, both MLR models were robust (assessed by the cross-validation procedure), predictive (evaluated by external validation performance) and not prone to chance correlation (measured by the *y*-

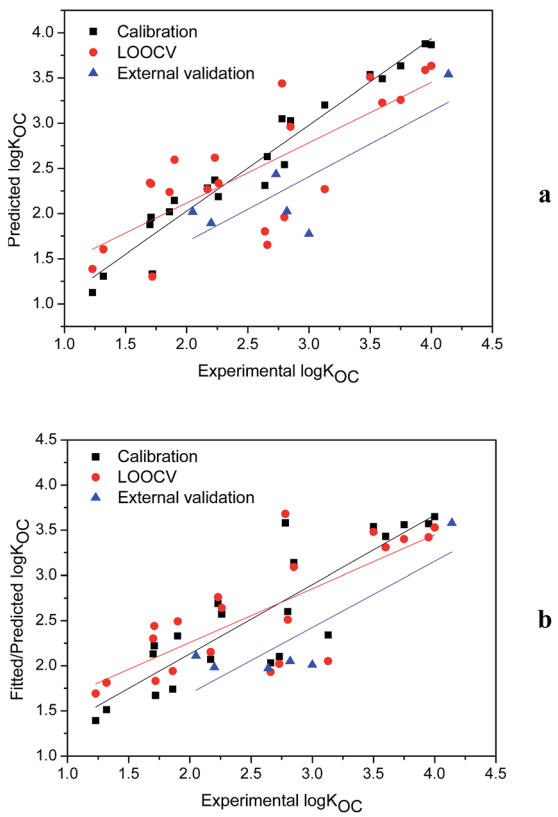


Fig. 2 Plot of experimental *versus* predicted log K_{OC} values using PLS and the complete set of descriptors from (a) aug-MIA-QSPR$_{\text{colour-}\varepsilon}$ and (b) aug-MIA-QSPR$_{\text{default}}$.

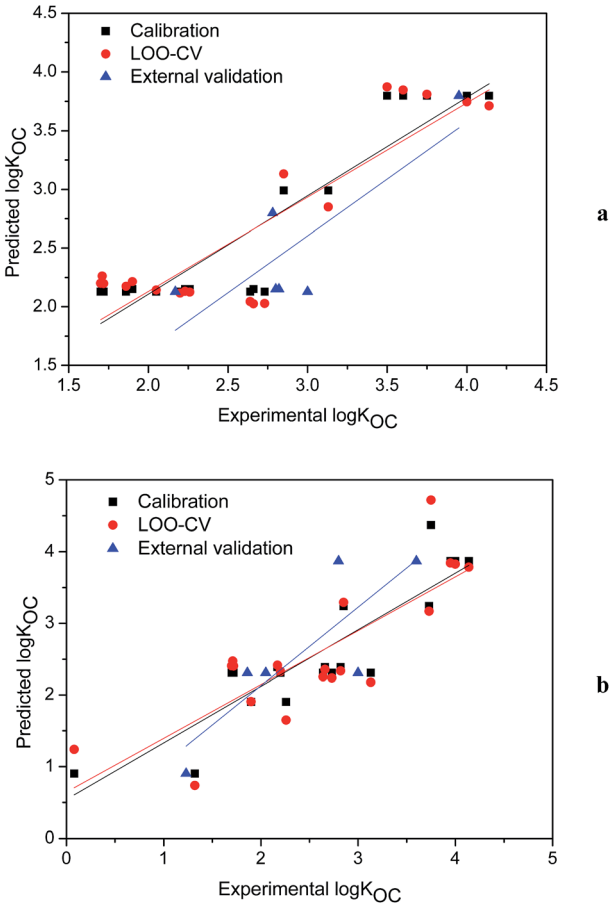


Fig. 3 Plot of experimental *versus* predicted log K_{OC} values using MLR and selected descriptors from the (a) aug-MIA-QSPR$_{colour\text{-}\varepsilon}$ and (b) aug-MIA-QSPR$_{default}$ models.

Table 4 Selected variables for the aug-MIA-QSPR$_{colour\text{-}\varepsilon}$ and aug-MIA-QSPR$_{default}$ models

	aug-MIA-QSPR$_{colour\text{-}\varepsilon}$			aug-MIA-QSPR$_{default}$			
Cpd	X5218	X11516	X13011	X164	X12570	X12582	log K_{OC}
1[a]	—	—	—	229	615	765	1.32
2[a]	—	—	—	229	615	765	0.08
3[a]	—	—	—	229	615	765	1.23
4	765	300	300	765	765	765	1.86
5	765	612	300	765	765	765	3.13
6	765	612	300	765	612	612	2.85
7[a]	—	—	—	765	612	612	3.73
8	765	300	300	765	765	765	2.20
9	765	300	300	765	765	765	2.64
10	765	300	300	765	765	765	3.00
11	765	300	300	765	765	765	2.05
12	765	300	300	765	765	765	2.73
13	765	300	300	765	765	765	1.70
14	765	300	300	765	765	765	1.72
15	350	612	250	426	426	612	3.95
16	350	612	250	765	426	426	3.75
17	350	612	250	426	426	612	3.60
18[b]	350	612	250	—	—	—	3.50
19	350	612	250	426	426	612	4.00
20[b]	350	250	250	—	—	—	2.78
21	350	612	250	426	426	612	4.14
22[b]	765	765	765	—	—	—	2.23
23	765	765	765	426	426	612	2.80
24	765	765	765	426	426	426	2.26
25	765	765	765	426	426	426	1.90
26	765	300	300	426	612	765	2.17
27	765	765	765	426	612	765	2.66
28	765	765	765	426	612	765	2.82
29	765	765	765	426	612	765	1.71

[a] Outliers for the aug-MIA-QSPR$_{colour\text{-}\varepsilon}$ model. [b] Outliers for the aug-MIA-QSPR$_{default}$ model.

randomization test), see Table 3 for values of these validation parameters. However, four outliers were observed with the aug-MIA-QSPR$_{colour\text{-}\varepsilon}$ model, including all three carboxylic acid derivatives with the COOH group directly bonded to the aromatic ring. In addition, the experimental *versus* predicted log K_{OC} plot obtained from this model clearly delineated block-wise separations (Fig. 3), which does not accurately describe the distribution of log K_{OC} values in the series of compounds used in the present study. An explanation for this trend points to a weakness inherent in using Pauling's electronegativity scale as a criterion for selecting the colours for different atoms, in the sense that while the influence of different atoms on the molecular soil sorption profiles is not necessarily the same, this scale assigns equal values for different atom pairs (*e.g.* N and Cl, and C and S), thus yielding identical predictions for different chemical structures. It is therefore concluded that electronegativity, as a chemical property, does not explain the variance of log K_{OC} values for this series of compounds. Posteriorly, the MLR model for the aug-MIA-QSPR$_{default}$ approach (based on molecular structure images in which the atom sizes are proportional to the Van der Waals radii) is analysed. Table 4 shows the variables contained in eqn (2). The components for each variable are pixel values in identical coordinates for images that constitute the MVI for the compound dataset.

aug-MIA-QSPR$_{colour\text{-}\varepsilon}$

$$\log K_{OC} = 3.3531 - 0.0016 \times X5218 + 0.0028 \times X11516 - 0.0027 \times X13011 \quad (1)$$

aug-MIA-QSPR$_{default}$

$$\log K_{OC} = 1.3974 + 0.0073 \times X164 - 0.0166 \times X12570 + 0.0106 \times X12582 \quad (2)$$

The MLR-based aug-MIA-QSPR$_{default}$ model presented three outliers (**18**, **20** and **22**); in fact, **18** and **22** bear a structural resemblance to one another, while **20** has a unique substituent

(the $-SO_2NH_2$ group at R_4). The statistical parameters obtained using the remaining data set were similar to those obtained using the aug-MIA-QSPR$_{\text{colour-}\varepsilon}$ model, with the advantage of obtaining a more linear rather than block-wise distribution between the experimental and predicted log K_{OC} values (Fig. 3). This result suggests that the atomic Van der Waals radius has a greater impact on the modelled property. Previous studies have demonstrated that good correlation exists between the Van der Waals radii and the non-specific behaviour of chemical compounds such as hydrophobicity, which is in turn known to correlate with the log K_{OC} value of molecules, thus justifying the obtained result.[21] Consequently, the influence of the chemical structures on the soil sorption in the series of herbicides can be meaningfully evaluated using the aug-MIA-QSPR$_{\text{default}}$ model. Bearing in mind that each variable represents a particular pixel coordinate in the MVI, a simple inductive analysis for each variable contained in eqn (2) is performed to identify these coordinates and most importantly the different atoms contained in these coordinate positions, for example the variable X161 has the pixel values [229, 229, 229] for the molecular structure images 1, 2 and 3 in the MVI (see Table 4), and according to Table 2, the pixel value 229 corresponds to the oxygen atom, and therefore the coordinate (variable) X161 represents a substituent position with oxygen atoms for compounds 1, 2 and 3, and from Table 1, it is evident that this substituent position is non-other than R1 (note that although this illustrative inductive analysis is limited to the first three compounds, all values for the suggested pixel coordinate must be in agreement with the images that constitute the MVI). Identifying the belongingness of the pixel coordinates (variables) in eqn (2) with respect to the substituents of the basic molecular scaffold enables the mechanistic interpretation of the regression model.

It is interesting to note that all three of the variables in the aug-MIA-QSPR$_{\text{default}}$ model (eqn (2)) correspond to pixel coordinates surrounding R_1; thus, the variance in this substituent explains the distribution of log K_{OC} values for the series of herbicides. This is an important finding because it indicates that soil sorption of a congeneric herbicide can be controlled by structural modifications in the R_1 substituent position, while changes in its phytotoxicity may be obtained by modifications of substituents in other ring positions, which do not affect log K_{OC}. According to the variable coefficients in eqn (2), higher pixel values in X12570 are related to lower soil sorption of herbicides, while the opposite is expected for X12582 and X164 variables. As can be observed in Table 4, the variable X164 contains pixel values for: the carbonyl oxygen (229) in *e.g.* compound **1**, carbon (426) atoms of *N*-alkyl chains in *e.g.* compound **15**, and blank spaces (765) *e.g.* compound **4**, due to the absence of substituents at this pixel coordinate for these molecular structure images. Since lower X164 values (*i.e.* 229) are related to chemicals with a lower soil sorption, the carboxyl group directly bonded to the aromatic ring (benzoic and picolinic acid derivatives) is considered to be an important requisite for the design of novel herbicides with low soil sorption. This result is consistent with other reports in the literature, in which the low soil sorption of carboxylic herbicides has been attributed to the repulsion between the surface negative charge of soil organic matter and the anionic carboxylate, as well as their high solubility in aqueous medium.[22] Additionally, pixels for carbon atoms (426) forming part of amide/carbamate moieties are generally associated with low log K_{OC} values, *e.g.* compound **25** (see Table 4), suggesting the importance of these functional groups in the design of compounds with a low soil sorption profile. A previous study on acetanilide herbicides suggested high water solubility as the key factor in explaining the low soil sorption.[23] Similar reasoning may be extrapolated to carbamates as well, although competition of water molecules for the sorption sites has been indicated as another important factor.[22] The variable X12570 contains pixel values for: carbon atoms (446) forming part of *N*-alkyl chains, *e.g.* compound **15**, hydrogen (612) *e.g.* compound **6**, bonds (615) *e.g.* compound **1** and blank spaces (765) *e.g.* compound **9**. As can be observed in Table 4, unlike X164, low pixel values for X12570 (*i.e.* 446) correspond to compounds with higher soil sorption (see also eqn (2), low X12570 values result in an increase in log K_{OC}). Therefore, since these low pixel values are for carbons forming part of the *N*-alkyl chains, it is deduced that this substituent type is important for high log K_{OC} values. Studies reported in the literature have indicated that the basic character of the *N*-alkyl chains favors the protonation of amine herbicides, thus enabling subsequent ion exchange reactions with soil organic colloids, which explains the high log K_{OC} values for these chemicals.[22] As for the variable X12582, it is observed from Table 4 that in particular, compounds with high soil sorption possess pixels for hydrogen (612) atoms forming part of *N*-alkyl chains *e.g.* compound **21** and OH groups *e.g.* compound **7**, and carbon (426) also belonging to the *N*-alkyl chain (compound **16**). Hydrogen bonding has been cited as the mechanism by which phenolic compounds are sorbed onto soil surfaces.[22] Altogether the following may be deduced with regards to the soil sorption profile of the studied compound dataset: (a) carboxyl, carbamate and amide substituent groups in R1 favor low soil sorption while (b) *N*-alkyl and hydroxyl moieties are important for high log K_{OC} values.

Conclusions

Coloured bidimensional projections of the chemical structures of a variety of herbicides have been used to encode the corresponding soil sorptions in terms of log K_{OC}. The analysis of colour pixels representing the chemical elements indicated that this physical property can be described by modifications limited only to the substituents attached to the congruent aromatic ring in the R_1 position, while other properties, such as phytotoxicity, may be varied by modifying the remaining substituents. In general, benzoic and picolinic acid herbicides present lower soil sorption, while *N*-alkylated aniline derivatives, followed by phenols, amides, carbamates and ethers, respectively, are expected to accumulate in the soil to a higher extent, and as a result this increases the possibility of harmful environmental and health effects. Thus, the design and synthesis of novel herbicides can be driven by considering the outcomes of this work in order to obtain efficient and safer compounds.

Acknowledgements

The authors are grateful to Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) for the financial support for this research, as well as to Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (PNPD/CAPES, Rede Mineira de Química to M.R.F.) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, to S.J.B. and M.P.F.) for the fellowships.

Notes and references

1 A. Sabljić, *Environ. Sci. Technol.*, 1987, **21**, 358–366.
2 A. Sabljić, H. Güsten, H. Verhaar and J. Hermens, *Chemosphere*, 1995, **31**, 4489–4514.
3 M. R. Freitas, S. V. B. G. Matias, R. L. G. Macedo, M. P. Freitas and N. Venturin, *Bull. Environ. Contam. Toxicol.*, 2014, **92**, 143–147.
4 M. P. Freitas and T. C. Ramalho, *Cienc. Agrotecnol.*, 2013, **37**, 485–494.
5 M. P. Freitas, S. D. Brown and J. A. Martins, MIA-QSAR: a simple 2D image-based approach for quantitative structure–activity relationship analysis, *J. Mol. Struct.*, 2005, **738**, 149–154.
6 C. A. Nunes and M. P. Freitas, *Eur. J. Med. Chem.*, 2013, **62**, 297–300.
7 M. R. Freitas, S. V. B. G. Matias, R. L. G. Macedo, M. P. Freitas and N. Venturin, *J. Agric. Food Chem.*, 2013, **61**, 8499–8503.
8 D. Mackay, W.-Y. Shiu and K.-C. Ma, *Illustrated Handbook of Physical–Chemical Properties and Environmental Fate for Organic Chemicals*, Lewis Publishers, New York, 1997.
9 A. D. Site, *J. Phys. Chem. Ref. Data*, 2001, **30**, 187–439.
10 R. D. Dennington II, T. A. Keith and J. M. Millam, *GaussView 5.0*, Gaussian, Inc., Wallingford, 2008.
11 W. Svante, M. Sjöström and L. Eriksson, *Chemom. Intell. Lab. Syst.*, 2001, **58**, 109–130.
12 J. W. Godden, F. L. Stahura and J. Bajorath, *J. Chem. Inf. Comput. Sci.*, 2000, **40**, 796–800.
13 S. J. Barigye, Y. Marrero-Ponce, Y. Martínez-López, F. Torrens, L. M. Artiles-Martínez, R. W. Pino-Urias and O. Martínez-Santiago, *J. Comput. Chem.*, 2013, **34**, 259–274.
14 I. Mitra, A. Saha and K. Roy, *Mol. Simul.*, 2010, **36**, 1067–1079.
15 K. Roy, I. Mitra, S. Kar, P. K. Ojha, R. N. Das and H. Kabir, *J. Chem. Inf. Model.*, 2012, **52**, 396–408.
16 P. P. Roy, S. Paul, I. Mitra and K. Roy, *Molecules*, 2009, **14**, 1660–1701.
17 P. K. Ojha, I. Mitra, R. N. Das and K. Roy, *Chemom. Intell. Lab. Syst.*, 2011, **107**, 194–205.
18 C. A. Nunes, M. P. Freitas, A. C. M. Pinheiro and S. C. Bastos, *J. Braz. Chem. Soc.*, 2012, **23**, 2003–2010.
19 *Matlab*, Mathworks, Inc., Natick, 2007.
20 N. Chirico and P. Gramatica, *J. Chem. Inf. Model.*, 2012, **52**, 2044–2058.
21 I. Moriguch, Y. Kanada and K. Komatsu, *Chem. Pharm. Bull.*, 1976, **24**, 1799–1806.
22 A. Delle-Site, *J. Phys. Chem. Ref. Data*, 2001, **30**, 187–439.
23 Q. Wang, W. Yang and W. Liu, *Pestic. Sci.*, 1999, **55**, 1103–1108.