Geoderma 268 (2016) 1–6

Contents lists available at ScienceDirect

Geoderma

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

Aug-MIA-SPR/PLS-DA classification of carbonyl herbicides according to levels of soil sorption

Mirlaine R. Freitas, Stephen J. Barigye, Joyce K. Daré, Matheus P. Freitas

Department of Chemistry, Federal University of Lavras, P.O. Box 3037, 37200-000, Lavras, MG, Brazil

ARTICLE INFO

Article history:
Received 12 August 2015
Received in revised form 6 January 2016
Accepted 13 January 2016
Available online 21 January 2016

Keywords:
Carbonyl herbicides
Soil sorption
Aug-MIA-SPR
PLS-DA

ABSTRACT

A major challenge in the design of new herbicides lies in the development of highly active, environmentally friendly compounds. Soil sorption is an ecotoxicological parameter used to probe the prospective environmental fate of persistent organic pollutants, such as some herbicides. This parameter, described in terms of logK_{OC} (the logarithm of the soil/water partition coefficient normalized to organic carbon), is usually estimated using the octanol/water partition coefficient (logP, easily calculated or determined experimentally). However, estimations obtained with the logP are not always accurate. Thus, this work reports the use of molecular descriptors derived from multivariate image analysis of carbonyl herbicides to achieve a predictive classification model based on the partial least squares-discriminant analysis (PLS-DA) method. This model yields 80% accuracy in calibration, 75% in leave-one-out cross-validation and 100% in external validation. In addition, the Y-randomization test reveals that the obtained model is stable from fortuitous correlation, since the accuracy in calibration after shuffling the classes block is only 0.5%. Chemical interpretation in terms of the structural features that affect soil sorption is performed, based on the weights of the selected variables in the classification model. Finally, novel herbicides are rationally designed, based on the inferences arrived at in the structural interpretation experiment and predictions of their qualitative and quantitative soil sorption profiles performed, using the built aug-MIA-SPR and Wang's models, respectively.

1. Introduction

Herbicides play an important role in weed control, but may persist in the environment and, consequently, accumulate in the food chain, causing serious health problems (Corsonlini et al., 2005). While some herbicides accumulate in water, thus affecting principally aquatic organisms through bioconcentration (Kehrig et al., 2011), others interact strongly with soil, presenting high soil sorption, which is described in terms of the logarithm of the soil/water partition coefficient normalized to organic carbon — logK_{OC}. This physicochemical parameter can be estimated using the easily calculable or measurable octanol/water partition coefficient, logP. However, univariate models based on logP for the prediction of logK_{OC} have not shown to be useful in the prediction of the soil sorption of many classes of herbicides (Sabljić et al., 1995; Freitas et al., 2014a, 2014b). Thus, more descriptive variables should be used to estimate/predict the logK_{OC} of chemical structures with herbicidal activity.

In this sense, the multivariate image analysis (MIA) method has emerged in the last decade as a successful source of descriptors, which have been used to model bioactivities and physicochemical parameters of a variety of compounds (Antunes et al., 2008; Nunes and Freitas, 2013a; Goodarzi et al., 2009; Goodarzi and Freitas, 2008, 2009). The descriptors in MIA are pixels (numerical data) of superimposable bidimensional chemical images (of a congeneric series of molecules), whose variance from an image to another (in terms of pixel coordinates representing different groups or substituent positions) explains the fluctuation in the response for a series of compounds. The MIA descriptors have been upgraded from molecules drawn as wireframes (Freitas et al., 2005) to colored and more realistic chemical representations (Nunes and Freitas, 2013b). Such a scheme has been recently used to model the soil sorption of aromatic herbicides, allowing for the comprehension of the structural characteristics responsible for high/low soil sorption (Freitas et al., 2015).

Recently, the so called aug-MIA descriptors have been extended to classification tasks using principal component analysis (PCA), hierarchical cluster analysis (HCA) and partial least squares-discriminant analysis (PLS-DA), giving rise to predictive aug-MIA-SAR models (Duarte et al., 2015). This qualitative analysis can be useful when quantitative response variables are not accessible or significantly inaccurate (in which case averaged values yield relatively high residuals), rendering regression for quantitative analysis impractical. In the present aug-MIA-SPR modeling, carbonyl herbicides (amides, carbamates, thiocarbamates and ureas) were evaluated because of their wide range of applications and the poor

Abbreviations: Aug-MIA-SPR, augmented multivariate image analysis applied to structure–property relationships; PLS-DA, partial least squares-discriminant analysis; logK_{OC}, logarithm of the soil/water partition coefficient normalized to organic carbon; LOOCV, leave-one-out cross-validation.

E-mail address: matheus@dqi.ufla.br (M.P. Freitas).

http://dx.doi.org/10.1016/j.geoderma.2016.01.013
0016-7061/

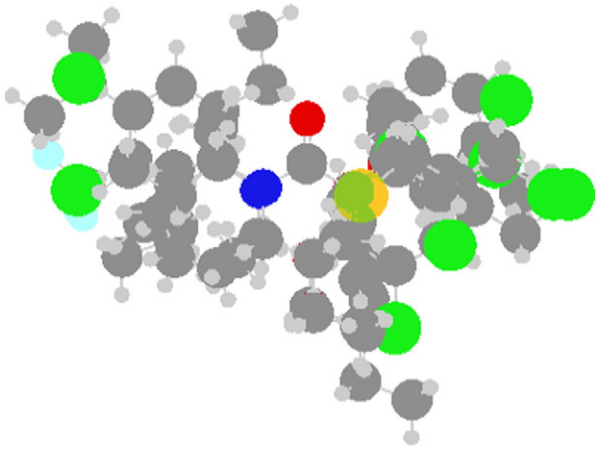


Fig. 1. Superposed chemical images of herbicides used in the aug-MIA-SPR modeling.

correlation of logK_{OC} with logP for some of these compounds (Sabljić et al., 1995). Despite the high predictive ability of some QSPR models (Schüürmann et al., 2006; Wang et al., 2009), the chemical interpretation indispensable to rationally design new herbicide candidates has not been straightforward, which can be easily assessed using our proposed SPR model.

2. Material and methods

A series of 26 carbonyl herbicides pertaining to amides, carbamates, thiocarbamates and ureas was obtained from the literature (Mackay et al., 1997) and the corresponding chemical structures drawn using the GaussView program (Dennington et al., 2008), with the carbonyl group considered as the common basic moiety for 2D alignment. It is worth mentioning that more than 20 compounds are recommended for SPR purposes (Young, 2009) and, therefore, the present dataset fulfills this requirement. Fig. 1 presents the superposed 26 chemical images to illustrate the variance in the chemical space, as well as the carbonyl group as the congruent substructure for alignment purposes. The spheres representing atoms in molecules were designed to be proportional to the respective van der Waals radii and each chemical structure (image) was saved as bitmaps (bmp) files in a workspace of 612 × 441 pixel dimension using the Microsoft Windows Paint application. The images were numerically transformed according to the RGB color system, in which each atom color is the result of a contribution from red (255), green (255) and blue (255) components; thus, the colors vary from black (zero) to white (765, the sum of all three components). This procedure was performed using the Chemoface program (Nunes et al., 2012), giving a 612 × 441 data matrix for each molecule. The 26 matrices were merged to give a 26 × 612 × 441 tridimensional array. Later, the 3D-array was unfolded to yield a [26 × (612 × 441)] matrix.

A step-wise variable selection procedure, considering both relevance (in Shannon's entropy terms) and orthogonality, was applied to the unfolded matrix yielding a lower dimensionality data matrix, capable of capturing the essence of the system using a minimum number of variables and allowing for greater model interpretability. Given the disparity in the logK_{OC} values reported for the same herbicides in the literature, a pattern recognition model using PLS-DA was exploited rather than a quantitative one. This method is based on the transformation of the original variables in latent variables, and these are linearly independent (Chevallier et al., 2006; Neto and Moita, 1998). These latent variables are linear combination of original variables. The preference of PLS-DA to linear discriminant analysis (LDA) is to profit from the score plots obtained with the former as these allow for a straightforward assessment of the contribution of the different variables in the stratification of compounds according to their properties. Later, these variables may be examined for the chemical information codified in terms of the functional groups or substructures relevant for the studied property. Note that when the variables constituting an LDA model are orthogonal, the PLS-DA and LDA techniques yield the same results (Wold et al., 2001).

Based on the average logK_{OC} values obtained from the literature (Mackay et al., 1997; Kenaga and Goring, 1980; Liu and Qian, 1995),

Table 1
Herbicides and the respective classification according to the soil sorption. Compounds marked with asterisk (*) pertain to the test set (external validation) and were selected through Kennard–Stone sampling.

Cpd no.	Herbicide	Class	Cpd no.	Herbicide	Class
1	Chlortoluron	2	14	Butachlor	2
2	Diuron	2	15	Metolachlor	2
3	Fenuron	1	16	Propachlor*	2
4	Fluometuron	2	17	Propanil*	2
5	Isoproturon	2	18	Diphenamid*	2
6	Monuron	1	19	Pronamide	2
7	Neburon	3	20	Butylate	2
8	Monolinuron	2	21	Diallate	3
9	Linuron	2	22	Triallate*	3
10	Barban	3	23	EPTC*	2
11	Chlorpropham	2	24	Pebulate	2
12	Propham	1	25	Molinate	2
13	Alachlor	2	26	Vernolate*	2

Table 2
Statistical results for the aug-MIA-SPR/PLS-DA modeling.

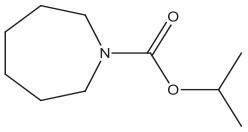
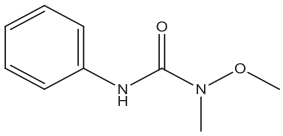
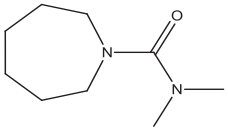
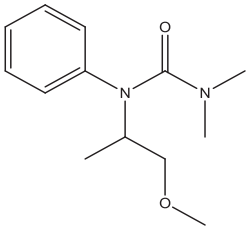
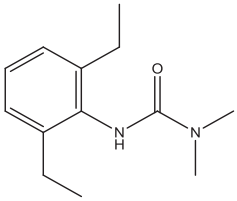
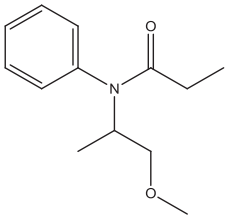
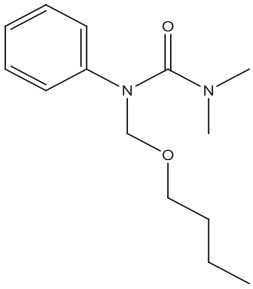
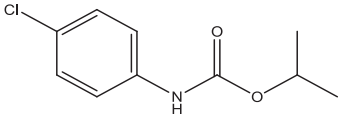
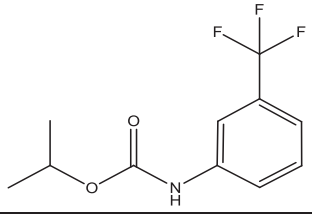
Parameter	Value
Variables	3
Calibration success (%)	80
Y-randomization success (%)	0.5
Leave-one-out success (%)	75
External validation success (%)	100

Table 3
Selected variables and the corresponding pixel values according to the RGB color system encoding atom types (289 = chlorine; 426 = carbon; 612 = hydrogen; 765 = blank space).

Cpd no.	X3100	X4543	X4148	Class[a]
1	765	765	289	2
2	765	765	289	2
3	765	765	765	1
4	765	765	765	2
5	765	765	765	2
6	765	765	765	1
7	765	426	289	3
8	765	765	765	2
9	765	765	289	2
10	765	765	289	3
11	765	765	289	2
12	765	765	765	1
13	765	765	765	2
14	765	765	765	2
15	765	765	765	2
16	765	765	765	2
17	765	765	289	2
18	765	765	765	2
19	426	289	765	2
20	612	426	612	2
21	765	426	765	3
22	765	426	765	3
23	612	426	612	2
24	612	426	765	2
25	612	765	765	2
26	612	426	612	2

[a] Classes: 1 = low soil sorption; 2 = medium soil sorption; 3 = high soil sorption.

Table 4
Prediction of soil sorption profile based on the aug-MIA-SPR PLS-DA and Wang's quantitative model for logK_{OC}.

No.	Chemical structure	Aug-MIA-SPR model[a]	Wang's model[b]
1		Low	1.595
2		Low	1.611
3		Low	1.668
4		Low	1.897
5		Low	1.988
6		Low	2.025
7		Low	2.079
8		Low	2.087
9		Low	2.123

(continued on next page)

Table 4 (*continued*)

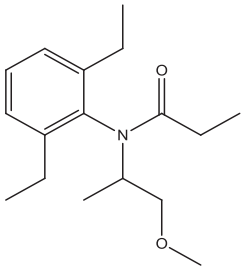
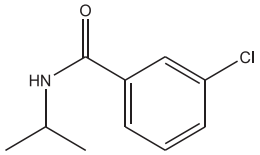
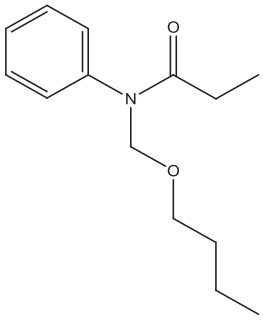
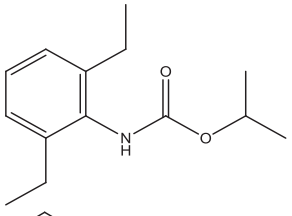
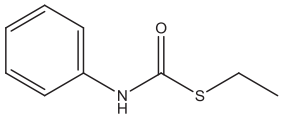
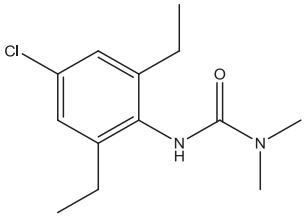
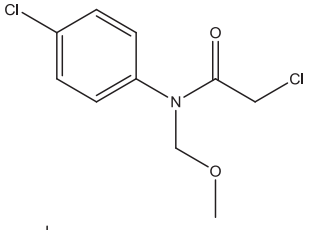
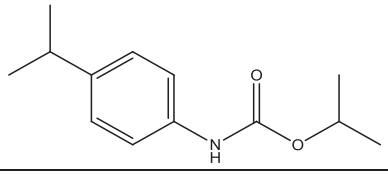
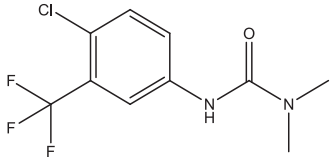
No.	Chemical structure	Aug-MIA-SPR model[a]	Wang's model[b]
10		Low	2.167
11		Low	2.191
12		Low	2.193
13		Low	2.228
14		Low	2.232
15		Low	2.260
16		Low	2.281
17		Low	2.305

Table 4 (*continued*)

No.	Chemical structure	Aug-MIA-SPR model[a]	Wang's model[b]
18		Low	2.323

[a] Low (class 1).
[b] MLR-based model for logK_{OC}.

the following classes (levels of soil sorption) were assigned for the 26 herbicides:

- Low soil sorption: logK_{OC} ≤ 2
- Medium soil sorption: 2 < logK_{OC} ≤ 3
- High soil sorption: logK_{OC} > 3

3. Results and discussion

The chemical structures of 26 carbonyl herbicides in Table 1 pertaining to amide, carbamate, thiocarbamate and urea groups were drawn and analyzed according to the procedure described earlier, using the soil sorption levels low (1) medium (2) and high (3) for classification purposes. Three descriptors (**X3100**, **X4543** and **X4148**) were filtered from the pool of variables generated for the dataset, representing pixels that encode chemical information on the herbicide chemical structures. The quality of the PLS-DA model obtained with the aug-MIA descriptors was evaluated in terms of the accuracy in calibration (%), *i.e.* the percentage of correct classification in the calibration step. The model was validated using leave-one-out cross-validation (LOOCV) and external validation (for compounds not included in the calibration set) procedures, according to the recommendation from the literature (Golbraikh and Tropsha, 2002). The Y-randomization test was performed to attest that the calibration model was not a result of fortuitous correlation. The statistical results are shown in Table 2.

The predictability of the model was attested by the accuracy in calibration and validation (both LOOCV and external), which were above 80% and 60%, respectively, while the negligible value for the Y-randomization test (0.5%) assures that the good calibration performance was not due to chance correlation.

A score plot using the first two, most informative latent variables was capable of clustering compounds with similar soil sorption profiles (Fig. 2). Herbicides with high soil sorption are found at the bottom of the score plot (negative scores in latent variable 2 — LV2), while those with medium and low soil sorption have positive scores in LV2, with the latter presenting more negative scores in LV1. The inconsistencies relative to the samples with medium soil sorption classified within the group of samples with low and high soil sorption are possibly due to the mean square error of the logK_{OC} values used in defining the classes and thus these clusters are more representative of the classes: low/medium, medium and medium/high.

In order to gain greater insight into the chemical characteristics responsible for the modeled soil sorption profiles, a close inspection of Table 3 (the data matrix) was performed to map the correspondence of the different structural moieties to the selected variables (**X3100**, **X4543** and **X4148**), as a means of assessing the contribution of the different chemical groups to the lowering or enhancing the soil sorption. From this analysis, herbicides with low soil sorption contain high pixel values (765 — blank space) in all three variables (coordinates); therefore, small groups at the positions shown in Fig. 3 lead to decreasing soil sorption. On the other hand, small pixel values (289 corresponding to chlorine in the **X4148** coordinate and 426 corresponding to carbon of long alkyl chains in the **X4543** coordinate) lead to increased soil sorption. In fact the removal of the chlorine atom from compound 7 (occupant of pixel position **X4148**) and the last 2 carbon atoms of the alkyl chain (containing the variable **X4543**) yields compound 6, representative of the passage from high to low soil sorption. While the similarity between lipophilicity of halogenated benzenes and natural sorbents is well known, the effect of chlorine on the specific ring position corresponding to **X4148** has not been reported to the best of our knowledge; on the other hand, the literature has indicated that the basic character of the *N*-alkyl chains (matching the **X4543** coordinate) favors the protonation of amine herbicides, thus enabling subsequent ion exchange reactions with soil organic colloids, which explains the high logK_{OC} values for related chemicals (Site, 2001).

3.1. Rational design of novel herbicides and validation

Based on the inferences obtained in the structural interpretation of the built PLS-DA model, 18 chemical structures were proposed, as a rational combination of the structural characteristics considered as desirable for low and low/medium soil sorption. These chemical structures

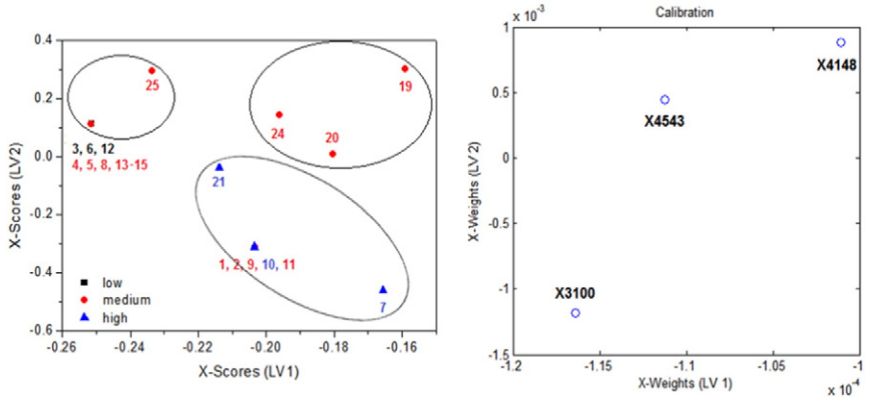


Fig. 2. Scores and loading plots obtained in the aug-MIA-SPR modeling of carbonyl herbicides.

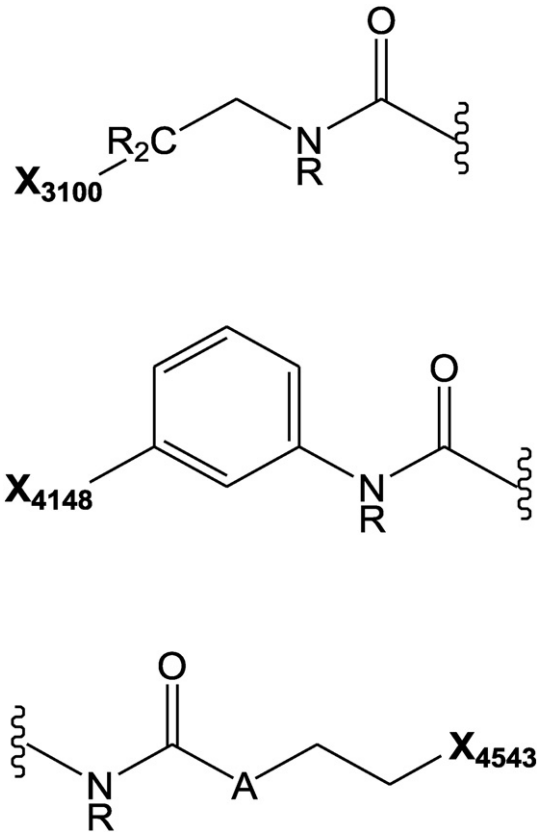


Fig. 3. Substituent positions (coordinates **X3100**, **X4148** and **X4543**) affecting the soil sorption of carbonyl herbicides, according to the aug-MIA-SPR/PLS-DA model.

were screened using the built PLS-DA model and 18 compounds (100%) were classified as of low soil sorption profile (Table 4). Additionally, in order to corroborate the predictions performed, the logK_{OC} values of the proposed structures were computed using Wang's quantitative Multiple Linear Regression (MLR)-based model (Wang et al., 2015), and the results compared. Table 4 shows the chemical structures predicted by both the aug-MIA-SPR and Wang's models, as members of the low soil sorption class and as of low/medium logK_{OC} values, respectively. The low soil sorption profile of the proposed structures suggests that these constitute a promissory set of eco-friendly chemicals to count on in agricultural weed control. Altogether, these results demonstrate the usability of the aug-MIA-SPR approach as a quick, simple and interpretable method, relevant in the rational design of organic chemical compounds of interest.

4. Conclusions

The utility of graphical representations of bidimensional chemical structures of carbonyl herbicides in encoding relevant structural information has been demonstrated. It was possible to precisely classify amide, carbamate, thiocarbamate and urea herbicides according to levels of soil sorption using the aug-MIA-SPR/PLS-DA technique. The interpretability of the built model in terms of the chemical characteristics that favor decreased soil sorption allowed for the rational design of 18 novel herbicides with the desired soil sorption profile. The soil sorption profile for these chemicals was posteriorly predicted using the built model and further corroborated employing a quantitative model reported in the literature. Synthesis and experimental evaluation of the physicochemical properties of these compounds constitute forthcoming tasks. It may therefore be inferred that the aug-MIA-SPR technique constitutes as an important method, useful in the rational design of environmentally friendly herbicides.

Acknowledgments

The authors are thankful to FAPEMIG (grant number APQ-00383-15) for the financial support as well as to CAPES/Rede Mineira de Química (to M.R.F.) and CNPq (to S.J.B. and M.P.F.) for the fellowships. This work is a collaboration research project of members of the Rede Mineira de Química (RQ-MG) supported by FAPEMIG (Project: CEX-RED-00010-14).

References

Antunes, J.E., Freitas, M.P., Rittner, R., 2008. Bioactivities of a series of phosphodiesterase type 5 (PDE-5) inhibitors as modelled by MIA–QSAR. Eur. J. Med. Chem. 43, 1632–1638.
Chevallier, S., Bertrand, D., Kohler, A., Courcoux, P., 2006. Application of PLS-DA in multivariate image analysis. J. Chemom. 20, 221–229.
Corsonlini, S., Ademollo, N., Romeo, T., Greco, S., Focardi, S., 2005. Persistent organic pollutants in edible fish: a human and environmental health problem. Microchem. J. 79, 115–123.
Dennington, R.D., Keith, T.A., Millam, J.M., 2008. GaussView 5.0. Gaussian, Inc., Wallingford.
Duarte, M., Barigye, S., da Mota, E., Freitas, M., 2015. Computational modelling of the antischistosomal activity for neolignan derivatives based on the MIA-SAR approach. SAR QSAR Environ. Res. 26, 205–216.
Freitas, M.P., Brown, S.D., Martins, J.A., 2005. MIA-QSAR: a simple 2D image-based approach for quantitative structure–activity relationship analysis. J. Mol. Struct. 738, 149–154.
Freitas, M.R., Freitas, M.P., Macedo, R.L.G., 2014b. Aug-MIA-QSPR modeling of the soil sorption of carboxylic acid herbicides. Bull. Environ. Contam. Toxicol. 93, 489–492.
Freitas, M.R., Matias, S.V.B.G., Macedo, R.L.G., Freitas, M.P., Venturin, N., 2014a. Three-parameter modeling of the soil sorption of acetanilide and triazine herbicide derivatives. Bull. Environ. Contam. Toxicol. 92, 143–147.
Freitas, M.R., Barigye, S.J., Freitas, M.P., 2015. Coloured chemical image-based models for the prediction of soil sorption of herbicides. RSC Adv. 5, 7547–7553.
Golbraikh, A., Tropsha, A., 2002. Beware of q^2! J. Mol. Graph. Model. 20, 269–276.
Goodarzi, M., Freitas, M.P., 2008. Predicting boiling points of aliphatic alcohols through multivariate image analysis applied to quantitative structure–property relationships. J. Phys. Chem. A 112, 11263–11265.
Goodarzi, M., Freitas, M.P., 2009. Prediction of electrophoretic enantioseparation of aromatic amino acids/esters through MIA-QSPR. Sep. Purif. Technol. 68, 363–366.
Goodarzi, M., Freitas, M.P., Ramalho, T.C., 2009. Prediction of 13C chemical shifts in methoxyflavonol derivatives using MIA-QSPR. Spectrochim. Acta A 74, 563–568.
Kehrig, H.A., Malm, O., Palermo, E.F.A., Seixas, T.G., Baêta, A.P., Moreira, I., 2011. Bioconcentração e biomagnificação de metilmercúrio na Baía de Guanabara, Rio de Janeiro. Quim Nova 34, 377–384.
Kenaga, E.E., Goring, C.A.I., 1980. Relationship between water solubility, soil sorption, octanol–water partitioning, and bioconcentration of chemicals in biota. In: Eaton, J.G., Parrish, P.R., Hendricks, A.C. (Eds.), Aquatic ToxicologyASTM STP 707. American Society for Testing and Materials, Philadelphia, pp. 78–115.
Liu, J., Qian, C., 1995. Hydrophobic coefficients of *s*-triazine and phenylurea herbicides. Chemosphere 31, 3951–3959.
Mackay, D., Shiu, W.-Y., Ma, K.-C., 1997. Illustrated handbook of physical–chemical properties and environmental fate for organic chemicals. Lewis Publishers, New York.
Neto, J.M.M., Moita, G.C., 1998. Uma introdução à análise exploratória de dados multivariados. Quim. *Nova* 21, 467–469.
Nunes, C.A., Freitas, M.P., 2013a. Aug-MIA-QSPR on the modeling of sweetness values of disaccharide derivatives. LWT Food Sci. Technol. 51, 405–408.
Nunes, C.A., Freitas, M.P., 2013b. Introducing new dimensions in MIA-QSAR: a case for chemokine receptor inhibitors. Eur. J. Med. Chem. 62, 297–300.
Nunes, C.A., Freitas, M.P., Pinheiro, A.C.M., Bastos, S.C., 2012. Chemoface: a novel free user-friendly interface for chemometrics. J. Braz. Chem. Soc. 23, 2003–2010.
Sabljić, A., Güsten, H., Verhaar, H., Hermens, J., 1995. QSAR modelling of soil sorption. Improvements and systematics of logK_{OC} vs. logK_{OW} correlations. Chemosphere 31, 4489–4514.
Schüürmann, G., Ebert, R.-U., Kühne, R., 2006. Prediction of the sorption of organic compounds into soil organic matter from molecular structure. Environ. Sci. Technol. 40, 7005–7011.
Site, A.D., 2001. Factors affecting sorption of organic compounds in natural sorbent/water systems and sorption coefficients for selected pollutants. A review. J. Phys. Chem. Ref. Datas 30, 187–439.
Wang, Y., Chen, J., Li, X., Wang, Y., Chen, L., Zhu, M., Yu, H., Kühne, R., Schüürmann, G., 2009. Estimation of soil organic carbon normalized sorption coefficient (K_{oc}) using least squares-support vector machine. QSAR Comb. Sci. 28, 561–567.
Wang, Y., Chen, J., Yang, X., Lyakurwa, F., Li, X., Qiao, X., 2015. In silico model for predicting soil organic carbon normalized sorption coefficient (K_{OC}) of organic chemicals. Chemosphere 119, 438–444.
Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58, 109–130.
Young, D.C., 2009. Computational Drug Design: A Guide for Computational and Medicinal Chemists. Wiley, New York.