Molecular Diversity (2006) 10: 109–118
DOI: 10.1007/s11030-005-9004-2

Full-length paper

A topological substructural molecular design to predict soil sorption coefficients for pesticides

Maykel Pérez González[1,2,*], Aliuska Morales Helguera[2,3] & Isidro G. Collado[4]
[1]*Unit of Services, Experimental Sugar Cane Station "Villa Clara-Cienfuegos", Ranchuelo, 53100, Villa Clara, Cuba;* [2]*Chemical Bioactives Center, Central University of Las Villas, Santa Clara, 54830 Villa Clara. Cuba;* [3]*Department of Chemistry, Faculty of Chemistry and Pharmacy, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba;* [4]*Department of Organic Chemistry, Faculty of Sciences, Cadiz University, Puerto Real, 11510, Cadiz, Spain*
*(*Author for correspondence, E-mail: mpgonzalez76@yahoo.es; Tel: 53-42-281473, Fax: 53-42-281130)*

Received 10 August 2005; Accepted 19 October 2005

Key words: pesticide, QSAR, soil sorption coefficient, topological indices, TOPS-MODE

Summary

A TOPological Sub-structural MOlecular DEsign (TOPS-MODE) approach was used to predict the soil sorption coefficients for a set of pesticide compounds. The obtained model accounted for more than 85% of the data variance and demonstrated the importance of the dipole moment, the standard distance, the polarizability, and the hydrophobicity in describing the property under study. In addition, we compared this new model to a previous one using different descriptors such as WHIM and molecular connectivity indices. Finally, the TOPS-MODE was used to calculate the contribution of different fragments to the soil sorption coefficient of the compounds studied. The present approximation proved to be a good method for studying the soil sorption coefficient for pesticides, but it could also be extended to other series of chemicals.

Abbreviations: TOPS-MODE, topological sub-structural molecular design approach; WHIM, weighted holistic invariant molecular descriptors; QSPR, Quantitative structure property relationships; QSAR, Quantitative structure activity relationships

Introduction

The sorption of commercial chemicals by soil and sediment plays an important role in the transport and mobility of these chemicals in the environment [1] and may significantly influence their chemical and biological transformation or degradation in the aquatic environment as well. Thus, the measurement or accurate estimation of soil sorption coefficients for hazardous chemicals is of critical importance for evaluating their fate and the resulting potential exposure to such chemicals in the environment, and consequently, for facilitating the whole process of environmental risk assessment.

Since the experimental determination of soil sorption coefficients is both difficult and expensive, several theoretical methods to determine these coefficients have been developed in which different regression equations between structure and parameters including water solubility, octanol-water partition coefficients, or bioconcentration factors have been used.

Despite the extensive experimental work that has been carried out by numerous laboratories for over 40 years, the measurement of soil sorption coefficients are only available for fewer hundreds of chemicals [1–4], which leaves many other chemicals with no reliable sorption coefficients.

Several years ago, more than 200 models for the K_{oc} estimation of non-ionic organic chemicals was collected and reviewed [5]; among these, nearly 80 related to pesticides. Unfortunately, these models are mainly class-specific and were usually obtained with a small number of chemicals. Furthermore, validation and regression diagnostics were often lacking, making it difficult to determine the predictive power and the range of application of a large number of the K_{oc} estimation models that have been published over the last 30 years.

One particularly useful method, the topological substructural molecular design (TOPS-MODE) approach, was developed in the context of *in silico* methods for modeling both the physicochemical properties and the biological activity of chemicals [6, 7].

The successful application of this theoretical approach to the modeling of toxicological and ecotoxicological properties [8–11] has inspired us to perform a more exhaustive study in order to test and validate the applicability of TOPS-MODE in assessing new chemical pesticides and their environmental impact. The selection of a data set of pesticide pollutant compounds is not casual; this property was previously studied by Gramatica et al. using different families of descriptors [12].

We will now show how TOPS-MODE is able to produce a good QSPR model that permits easy structural interpretation of the results in terms of group contributions to soil sorption coefficients.

TOPS-MODE approach

In the present paper, the TOPS-MODE approach was used to obtain molecular descriptors through which the QSAR function was developed. Since the mathematical details of the method have been previously reported [6, 7], we will outline only the fundamental points.

Briefly, this method codifies the molecular structure by means of the edge adjacency matrix **E** (likewise called bond adjacency matrix **B**). The **E**, or **B**, matrix is a square table of order m (the number of chemical bonds in the molecule). The elements of such a matrix (e_{ij}) are equal to 1 if bonds i and j are adjacent (this presupposes the existence of an atom that participates either in bond i or bond j); otherwise, they are equal to 0. In order to codify information related to the heteroatom, the TOPS-MODE approach replaces **B** with **B** ($\mathbf{w}_{ij}$) weighted matrices. The weights ($\mathbf{w}_{ij}$) are chemically meaningful numbers such as bond distances, bond dipoles, bond polarizabilities, or even mathematical expressions involving atomic weights such as hydrophobicity or van der Waals radii [8–11]. These weights are introduced in the main diagonal of matrix B ($\mathbf{w}_{ij}$). Afterwards, the spectral moments of this matrix may be used as molecular fingerprints in QSAR studies in order to codify the molecular structure. By definition, the expression "spectral moments" must be understood as the sum of the elements in the natural powers of **B** ($\mathbf{w}_{ij}$). This means that the spectral moment of order k (μ_k) is the sum of the main diagonal elements (e_{ii}) of matrix **B** $(\mathbf{w}_{ij})^{\mathbf{k}}$. In the present work the **B** ($\mathbf{w}_{ij}$) matrix was weighted in the main diagonal with the standard bond distance, standard dipole moments, atomic polarizability, and atomic hydrophobicity, as shown in Table 1.

Such a parameter μ_1 equals the sums of atom molar refractivity, bond dipoles, or bond distances in the molecule according to each selected case. The calculation of μ_k was carried out with the software package ModesLab 1.0 [13, 14].

Table 1. Definition of the different weighting bonds used in the current work.

Weighting bonds	Definition[a]
Distance	Standard bond distances
Dipole	Bond parameters computed with relative electronegativity
Polarizability	Bond parameters computed with polarizability
Hydrophobicity	Bond parameters computed with hydrophobicity

[a]See ref. [17] for a more complete definition of bond parameters.

Computation of fragment contributions

Each of the μ_k spectral moments contains structural information about the molecules that can be directly obtained by a computational approach [13]. The first step in this approach is to select the substructures whose contribution to the moments is to be determined, in this case the selection of the fragments was carry out taking into account the principal sub-structures in the training set and according to the functional group and previous knowledge of the substructures that should be important in this type of activity. Then all the fragments (sub graphs) contained in the corresponding substructure are generated and the spectral moments for both the substructure and all its fragments are calculated. The contribution of the substructure of the spectral moments is thus obtained as the difference between the spectral moments of the substructure and all its fragments. Once the contributions of the various pertinent structural fragments have been established, describing the property under study is simply a matter of substituting these contributions into the quantitative model developed.

Data set and computational strategies

The experimental soil sorption coefficients log K_{oc} of the 143 pesticides used as the training set were taken from the paper of Gramatica et al. [12]. Moreover, the log K_{oc} values for other chemicals used by the same authors were used as an external validation set. The names of the chemicals used in each set, as well as their predicted and observed activities are shown in Tables 2 and 3.

The TOPS-MODE computer software [13, 14] was employed to calculate the molecular descriptors. The standard dipole moments, standard bond distances, atomic polarizability, and atomic hydrophobicity were used as bond weights for differentiating heteroatoms [6].

In general, 15 spectral moments were calculated for each of the studied schemes, making a total number of 60 descriptors, of them descriptors with constant or near constant values were discarded.

For the remaining descriptors, a pairwise correlation analysis was performed for eliminating the collinearity among them. The procedure consists of elimination of one of the descriptors from each pair with the modulus of the correlation coefficients higher than a predefined value R_{max}. The procedure must be carried out with care. Indeed, let $R_{ij} = R(d_i, d_j)$ be the correlation coefficient between descriptors d_i and d_j. Then from $R_{ij} > R_{max}$ and $R_{jk} > R_{max}$ does not follow that $R_{ik} > R_{max}$. So in this case, if d_j is eliminated, d_k must be retained.

In this work, we have used the following algorithm of the pairwise correlation analysis.

1. Sort descriptors by variance and exclude all descriptors with the variance lower than the predefined value. Let D be the descriptor with the highest variance.

Table 2. Observed, predicted, and residual Log K_{oc} of pesticides used in the training series.

Number	Name	Obs. (Log K_{oc})	Pred. (Log K_{oc})	Residual (Log K_{oc})
1	2-Chlorophenylurea	1.61	1.71	−0.10
2	2-Fluorophenylurea	1.32	1.42	−0.10
3	3,4-Dichlorophenylurea	2.53	2.13	0.40
4	3-Bromophenylurea	2.12	1.96	0.16
5	3-Chloro-4-methoxyphenylurea	2.00	1.89	0.11
6	3-Chlorophenylurea	2.01	1.71	0.30
7	3-Fluorophenylurea	1.77	1.42	0.35
8	3-Methyl-4-fluorophenylurea	1.75	1.60	0.15
9	3-Methyl-4-bromophenylurea	2.37	2.14	0.23
10	3-Methylphenylurea	1.56	1.47	0.09
11	3-Trifluoromethylphenylurea	1.98	1.92	0.06
12	4-Bromophenylurea	2.06	1.96	0.10
13	4-Fluorophenylurea	1.52	1.42	0.10
14	4-Phenoxyphenylurea	2.56	2.49	0.07
15	Acetochlor	2.32	2.89	−0.57
16	Alachlor	2.28	2.76	−0.48
17	Aldicarb	1.50	1.98	−0.48
18	Aldicarb sulfone	0.42	0.23	0.19
19	Aldrin	4.69	4.71	−0.02
20	Ametryn	2.59	2.42	0.17
21	Atrazine	2.24	2.06	0.18
22	Azinphos methyl	2.28	2.70	−0.42
23	Benfluralin	3.99	3.48	0.51
24	Benomyl	2.71	2.81	−0.10
25	Butachlor	2.86	3.38	−0.52
26	Butralin	3.98	3.14	0.84
27	Butylate	2.11	2.48	−0.37
28	Butyl-N-phenylcarbamate	2.26	2.20	0.06
29	Carbaryl	2.40	2.42	−0.02
30	Carbendazim (MBC)	2.35	1.98	0.37
31	Carbofuran	1.75	2.51	−0.76
32	Carbophenothion	4.66	4.27	0.39
33	Chlorbromuron	2.70	2.67	0.03
34	Chlordane	5.15	4.97	0.18
35	Chlorfenvinphos (cis)	2.47	3.26	−0.79
36	Chlorfenvinphos (trans)	2.47	3.26	−0.79
37	Chlorotoluron	2.02	2.26	−0.24
38	Chloroxuron	3.55	3.30	0.25
39	Chlorpropham	2.53	2.48	0.05
40	Chlorpyrifos	3.70	3.70	0.00
41	Chlorpyrifos methyl	3.52	3.19	0.33
42	Crotoxyphos (trans)	2.00	2.29	−0.29
43	Cyanazine	2.28	2.30	−0.02
44	Cycloate	2.54	2.47	0.07
45	Diallate (cis)	3.28	2.80	0.48
46	Diallate (trans)	3.28	2.80	0.48
47	Diazinon	2.75	2.97	−0.22
48	Dicrotophos (cis)	1.66	0.91	0.75
49	Dieldrin	4.55	4.94	−0.39
50	Dimethoate	1.20	1.73	−0.53
51	Dinitramine	3.63	3.09	0.54

(*Continued on next page*)

Table 2. (*Continued*)

Number	Name	Obs. (Log K_{oc})	Pred. (Log K_{oc})	Residual (Log K_{oc})
52	Dipropetryn	3.07	2.83	0.24
53	Disulfoton	3.22	3.03	0.19
54	Diuron	2.40	2.50	−0.10
55	Endosulfan	4.13	3.97	0.16
56	EPTC	2.38	2.07	0.31
57	Ethion	4.06	3.92	0.14
58	Ethoprophos	1.80	1.87	−0.07
59	Ethyl-N-phenylcarbamate	1.82	1.85	−0.03
60	Fenamiphos	2.51	2.86	−0.35
61	Fenitrothion	2.63	2.44	0.19
62	Fensulfothion	2.52	3.07	−0.55
63	Fenuron	1.40	1.65	−0.25
64	Fluchloralin	3.55	3.49	0.06
65	Fluometuron	2.00	2.30	−0.30
66	Fonofos	3.44	3.09	0.35
67	Imazalil	3.73	3.37	0.36
68	Ipazine	2.91	2.50	0.41
69	Isazophos	2.01	2.73	−0.72
70	Lindane	3.00	3.16	−0.16
71	Linuron	2.70	2.42	0.28
72	Malathion	3.07	3.03	0.04
73	Metalaxyl	1.57	–	–
74	Methiocarb	2.32	2.71	−0.39
75	Methomyl	1.30	1.31	−0.01
76	Methoxychlor	4.90	4.68	0.22
77	Methyl-N-(3,4-dichlorophenyl)carbamate	2.74	2.44	0.30
78	Methyl-N-(3-chlorophenyl)carbamate	2.15	2.03	0.12
79	Methyl-N-phenylcarbamate	1.73	1.59	0.14
80	Metobromuron	2.10	2.26	−0.16
81	Metolachlor	2.46	2.92	−0.46
82	Metoxuron	1.72	2.27	−0.55
83	Metribuzin	1.71	1.91	−0.20
84	Mevinphos (cis)	0.85	0.85	0.00
85	Mevinphos (trans)	0.85	0.85	0.00
86	Molinate	1.92	2.10	−0.18
87	Monolinuron	2.10	2.01	0.09
88	Monuron	1.95	2.08	−0.13
89	N-(3,4-Dichlorophenyl)-N′-methylurea	2.46	2.32	0.14
90	N-(3,5-DiMe-4-Br-phenyl)-N′,N′-dimethylurea	2.53	2.67	−0.14
91	N-(3,5-Dimethylphenyl)-N′,N′-dimethylurea	1.73	2.05	−0.32
92	N-(3-Chloro-4-methoxyphenyl)-N′methylurea	1.84	2.08	−0.24
93	N-(3-Chloro-4-methylphenyl)-N′-methylurea	2.10	2.07	0.03
94	N-(3-Chlorophenyl)-N′,N′-dimethylurea	1.79	2.09	−0.30
95	N-(3-Chlorophenyl)-N′-methylurea	1.93	1.90	0.03
96	N-(3-Fluorophenyl)-N′,N′-dimethylurea	1.73	1.79	−0.06
97	N-(3-Methoxyphenyl)-N′,N′-dimethylurea	1.72	1.84	−0.12
98	N-(4-Fluorophenyl)-N′,N′-dimethylurea	1.43	1.79	−0.36
99	N-(4-Methoxyphenyl)-N′,N′-dimethylurea	1.40	1.83	−0.43
100	N-(4-Methylphenyl)-N′,N′-dimethylurea	1.51	1.84	−0.33
101	Nitralin	2.92	3.24	−0.32
102	N-Phenyl-N′-cycloheptylurea	2.37	2.67	−0.30

(*Continued on next page*)

Table 2. (*Continued*)

Number	Name	Obs. (Log K_{oc})	Pred. (Log K_{oc})	Residual (Log K_{oc})
103	N-Phenyl-N′-cyclohexylurea	2.07	2.51	−0.44
104	N-Phenyl-N′-cyclopentylurea	1.93	2.32	−0.39
105	N-Phenyl-N′-cyclopropylurea	1.74	2.00	−0.26
106	N-Phenyl-N-methylurea	1.29	1.51	−0.22
107	Oryzalin	3.40	3.24	0.16
108	Oxadiazon	3.51	3.66	−0.15
109	Oxamyl	1.00	1.62	−0.62
110	p,p-DDE	4.82	4.37	0.45
111	p,p-DDT	5.31	5.14	0.17
112	Parathion	3.20	2.75	0.45
113	Parathion methyl	3.00	2.25	0.75
114	Pebulate	2.80	2.23	0.57
115	Pentyl-N-phenylcarbamate	2.61	2.37	0.24
116	Phenylurea	1.50	1.27	0.23
117	Phorate	2.70	2.91	−0.21
118	Phosalone	3.71	3.72	−0.01
119	Profenofos	3.03	3.27	−0.24
120	Profluralin	4.01	3.69	0.32
121	Prometon	2.60	2.10	0.50
122	Prometryn	2.85	2.62	0.23
123	Propachlor	2.42	2.32	0.10
124	Propazine	2.40	2.26	0.14
125	Propham	1.83	2.04	−0.21
126	Propiconazole	3.39	3.77	−0.38
127	Propoxur	1.67	2.19	−0.52
128	Propyl-N-phenylcarbamate	2.06	2.03	0.03
129	Secbumeton	2.78	2.08	0.70
130	Siduron	2.31	2.72	−0.41
131	Simazine	2.10	1.86	0.24
132	Tebuthiuron	1.83	1.63	0.20
133	Terbufos	2.82	3.37	−0.55
134	Terbutryn	2.85	2.61	0.24
135	Thiabendazole	3.24	–	–
136	Thiobencarb	3.27	3.07	0.20
137	Triadimefon	2.71	2.94	−0.23
138	Triallate	3.35	3.21	0.14
139	Trichlorfon	1.90	1.69	0.21
140	Tricyclazole	3.09	2.61	0.48
141	Trietazine	2.76	2.31	0.45
142	Trifluralin	3.93	3.44	0.49
143	Vernolate	2.33	2.24	0.09

2. Calculate correlation coefficient between D and all other descriptors.
3. Exclude descriptor having the modulus of the correlation coefficient with D higher than R_{max}.
4. Let D be the next descriptor with the highest variance. Go to step (2). If there are no descriptors left, stop.

In this connection, a total of 24 molecular descriptors were taking into account for development of the QSPR models. The statistical processing for obtaining the QSPR model was carried out by using forward stepwise regression methods [15], in which the independent variables are individually added or deleted from the model at each step of the regression, depending on the Fisher ratio values selected, until the "best" model was obtained.

Thus, by examining the regression coefficient, standard deviation, and the F of Fisher, as well as the proportion between the cases and variables in the equation and the

Table 3. Observed, predicted, and residual Log K_{oc} of compounds used in external prediction series.

Number	Name	Obs. (Log K_{oc})	Pred. (Log K_{oc})	Residual (Log K_{oc})
1	Aldicarb sulfoxide	0.56	1.38	−0.82
2	Anilazine	3.00	3.03	−0.03
3	Asulam	2.48	1.96	0.52
4	Chlorbufam	2.21	2.53	−0.32
5	Cyromazine	2.30	1.30	0.26
6	Demeton-S-methyl	1.49	1.47	0.02
7	Dichlorvos	1.67	1.06	0.61
8	EPN	3.12	3.43	−0.31
9	Fenobucarb	1.71	2.31	−0.60
10	Iprobenfos	2.40	2.51	−0.11
11	Leptophos	4.50	4.52	−0.02
12	Methidathion	1.53	1.99	−0.46
13	Neburon	4.00	3.09	0.91
14	Piperophos	3.44	3.66	−0.22
15	Pirimicarb	1.90	2.19	−0.29
16	Pirimiphos methyl	3.00	2.81	0.19
17	Sulprofos	4.08	3.98	0.10
18	Terbuthylazine	2.32	2.25	0.07
19	Thiodicarb	2.54	2.78	−0.24
20	Xylicarb	1.71	1.92	−0.21

"leave-one-out" cross validation method, we were able to evaluate the quality of the model. In addition, a calculation of the regression coefficient and standard deviation of the external prediction set was taken into account for validating the obtained model. Compounds in the external prediction set were not used to develop the prediction function.

Results and discussion

Quantitative structure property relationships

The model selection was subjected to the principle of parsimony such that a function was chosen that had as high a statistical significance but as few parameters (b_k) as possible. The six-dimensional models are thus characterized by the best compromise between predictive power and model complexity. That is to say that the addition of another variable does not increase the predictive power such that the consequent increase in complexity is counterbalanced.

The best QSPR model that we were able to obtain with the TOPS-MODE descriptors is given below, together with the statistical parameters of the regression:

$$\log(K_{oc}) = 0.173 + 1.48 \cdot 10^{-7}\mu_{10}^{Dip} - 3.97 \cdot 10^{-11} \cdot \mu_{15}^{Dip} + 0.002 \cdot \mu_4^{Dist} + 0.30 \cdot \mu_1^H - 0.002 \cdot \mu_5^H + 8.88 \cdot 10^{-6} \cdot \mu_7^P \quad (1)$$

$$N = 143 \qquad S = 0.370 \qquad R^2 = 0.838$$
$$F = 117.24 \qquad p < 10^{-5} \qquad q^2 = 0.812$$

where N is the number of compounds included in the model, R^2 is the square of the correlation coefficient, S is the standard deviation of the regression, F is the Fisher ratio, q^2 is the determination coefficient of the cross-validation, and p is the significance of the variables in the model.

The variables included in the model are designated as follows: the sub-index represents the order of the spectral moment while the super-index indicates the type of bond weight used, i.e., *Dip* for dipole moment, *Dist* for standard distance, *H* for hydrophobicity, and *P* for polarizability.

The good correlation between the selected variables and the soil sorption coefficient is shown in Figure 1.

It should be noted that two outliers have been removed from the complete data set. While it is inappropriate to remove compounds from a data set simply to improve the correlation, an analysis of outliers omitted from a QSAR or QSPR can provide important information. In this context, an examination of the two omitted outliers is in order.

Analysis of the residuals for Equation (1) identified Metalaxyl and Thiabendazole as significant outliers. This latter compound (number 135 in our list) is a special case that combines two different heterocyclic moieties through a conjugated bond. Hence the dual character of Thiabendazole as both a thiazole and a benzoimidazole keep this compound from being classified among the diazoles. In addition, the interaction of the fused phenyl ring with the conjugated bond of both heterocyclic identities should be taken into account.

Removal of these compounds and subsequent reanalysis of the dataset produced the following QSPR:

$$\log(K_{oc}) = 0.165 + 1.52 \cdot 10^{-7}\mu_{10}^{Dip} - 4.08 \cdot 10^{-11} \cdot \mu_{15}^{Dip} + 0.002 \cdot \mu_4^{Dist} + 0.29 \cdot \mu_1^H - 0.002 \cdot \mu_5^H + 8.52 \cdot 10^{-6} \cdot \mu_7^P \quad (2)$$

$$N = 141 \qquad S = 0.340 \qquad R^2 = 0.859 \qquad F = 135.54$$
$$p < 10^{-5} \qquad q^2 = 0.831$$

The structural significance of this model will become more evident later, when we analyze the contribution of the different structural fragments to the soil sorption coefficient. From the statistical point of view, however, it is obvious that this is a robust model.

As mentioned above, one of our objectives was to compare the reliability of the TOPS-MODE approach in describing the property under study with that of other descriptors and methods. We thus carried out a comparison of our approach with that of Gramatica et al. [12], in which the model was developed using the same data set and the same number of variables that had been included in the TOPS-MODE QSPR model. The results are given in Table 4.

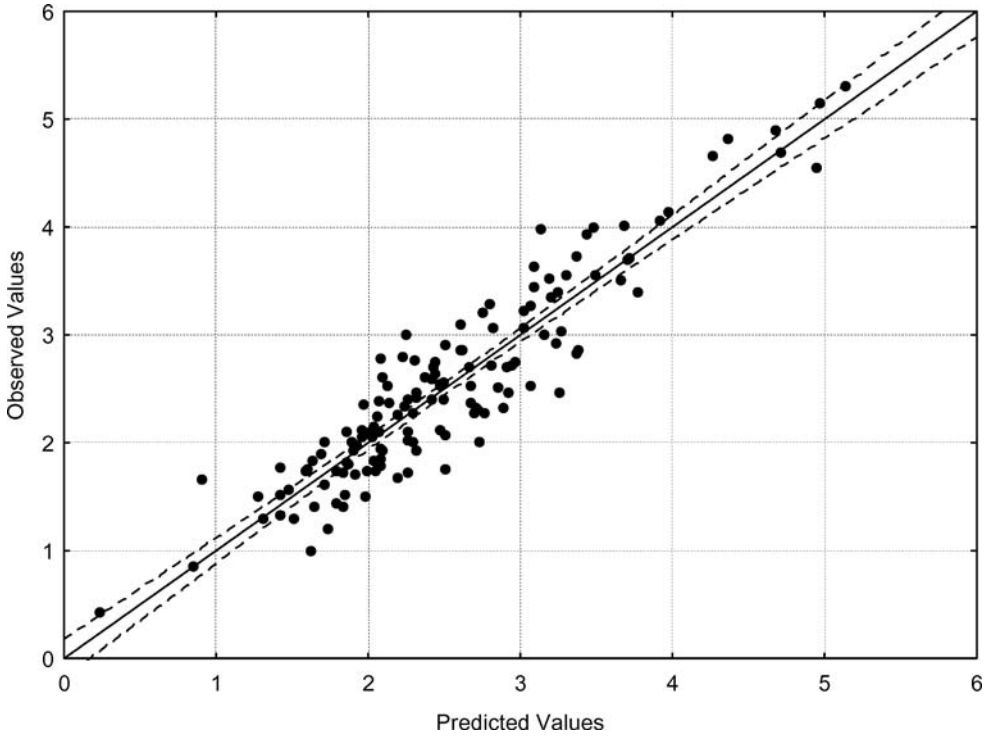

Figure 1. The linear relation between observed and predicted soil sorption coefficients for Equation (1).

As can be seen from the table, there are no remarkable differences in the explanation of the experimental variance given by the various models. While the TOPS-MODE QSPR model explains more than 85% of the soil sorption coefficient, Gramatica et al.'s model is able to explain just over 84% of such variance.

However, the TOPS-MODE model is slightly superior to the other model not only in the statistical parameters of the regression, but also, and more importantly, in its stability upon the inclusion or exclusion of compounds, as measured by the correlation coefficient and standard deviation of the cross-validation. Because of the structural variability of the compounds in the data set, the statistics from the leave-one-out cross-validation can be considered a good measurement of the predictive ability of the models. As can be seen in Table 4, the value of the determination coefficient of the leave-one-out cross-validation for the model obtained with the spectral moments ($q^2 = 0.831$) is slightly higher.

Table 4. Statistical parameters of the lineal regression models obtained for the two kinds of descriptors.

Descriptors	Number of variables	N	S	R^2	F	q^2
González MP et al.[a]	6	143	0.37	0.838	117.4	0.812
Gramatica et al.[b]	6	143	0.38	0.824	106.3	0.805
González MP et al.[a]	6	141	0.34	0.859	135.5	0.831
Gramatica et al.[b]	6	141	0.35	0.843	119.9	0.824

[a]Models according to the TOPS-MODE approach.
[b]Models reported by Gramatica et al. according to the reference [12].

Thus, although the TOPS-MODE approach presents better descriptive features than the model reported by Gramatica et al. from statistical point of view, there are no significant differences between the models in their prediction of the soil sorption coefficients with the leave-one-out methodology. This can be seen from the fact that the level of significance of an examination of this data set was $p = 0.242$; if the means were significantly different, the value would have been below 0.05.

A superficial analysis of this comparison would lead to the conclusion that both QSAR models are equivalent from a statistical standpoint. A more profound analysis, however, gives the TOPS-MODE model a definite edge over that reported by the Italian group in terms of quality. Thus, when the model was tested on an external validation set of 20 compounds belonging to the model applicability domain mainly on the basis of their leverages, the result was $R^2_{\mathrm{EXT}} = 0.777$. Although this suggests that the effective prediction power of our model is less than that obtained with internal validation, it is still higher than the result obtained with Grammatica et al.'s model, which resulted in an $R^2_{\mathrm{EXT}} = 0.607$. The statistical fit should therefore not be confused with the ability of a model to make predictions, as this result makes clear that the model with WHIM descriptors possesses a more limited prediction capability than the TOPS-MODE model. This validation process is of great importance in demonstrating the usefulness of the QSAR model for predictive purposes since such a model is of no value if it presents a good fit, but is unable to make predictions.

Furthermore, the TOPS-MODE showed how it is possible to produce a good QSPR model that permits easy structural interpretation of the results in terms of group contributions to soil sorption coefficients.

Fragment contributions

One of the most important advantages that the TOPS-MODE brings to the study of QSPR and QSAR involves the structural interpretability of the models. This interpretability comes from the fact that the spectral moments can be expressed as linear combinations of structural fragments. In this way, we can determine which fragments make a positive or a negative contribution to the property under study, as interpreted in terms of physicochemical or biological processes.

For example, both Figure 2 and Table 5 shows that the increase of the aliphatic ring size of the fragment series F_{35} to F_{38} leads to an increased contribution to the property in question.

In accordance with equation 2, higher hydrophobicity of the fragment increases its contribution to the soil sorption coefficient. This behavior has already been noted by several authors who correlated this property to the partition coefficient n-octanol/water (log K_{ow}), as shown in Table 6.

The increase of the aliphatic contribution to fragments F_{32}, F_{34}, and F_{45} by virtue of longer aliphatic chains leads to a straightforward increase in the soil sorption coefficient. Gramatica et al. thus propose that the bulk of the substituents plays a central role in the increase of the soil sorption coefficient and emphasize the strong dependence of the soil sorption of non-ionic pesticides on the size of the compounds. According to their regression coefficient, then, an increase in molecular size leads to an increase in sorption.

However, as can be seen from fragments F_{33} and F_{34}, which have a slightly higher sorption, Gramatica et al.'s assertions do not necessarily hold true. Thus, while certain shifts of the ramification of these fragments affect their size, they exhibit no steady changes in their hydrophobicity. On the basis of the above results, then, we hypothesize that

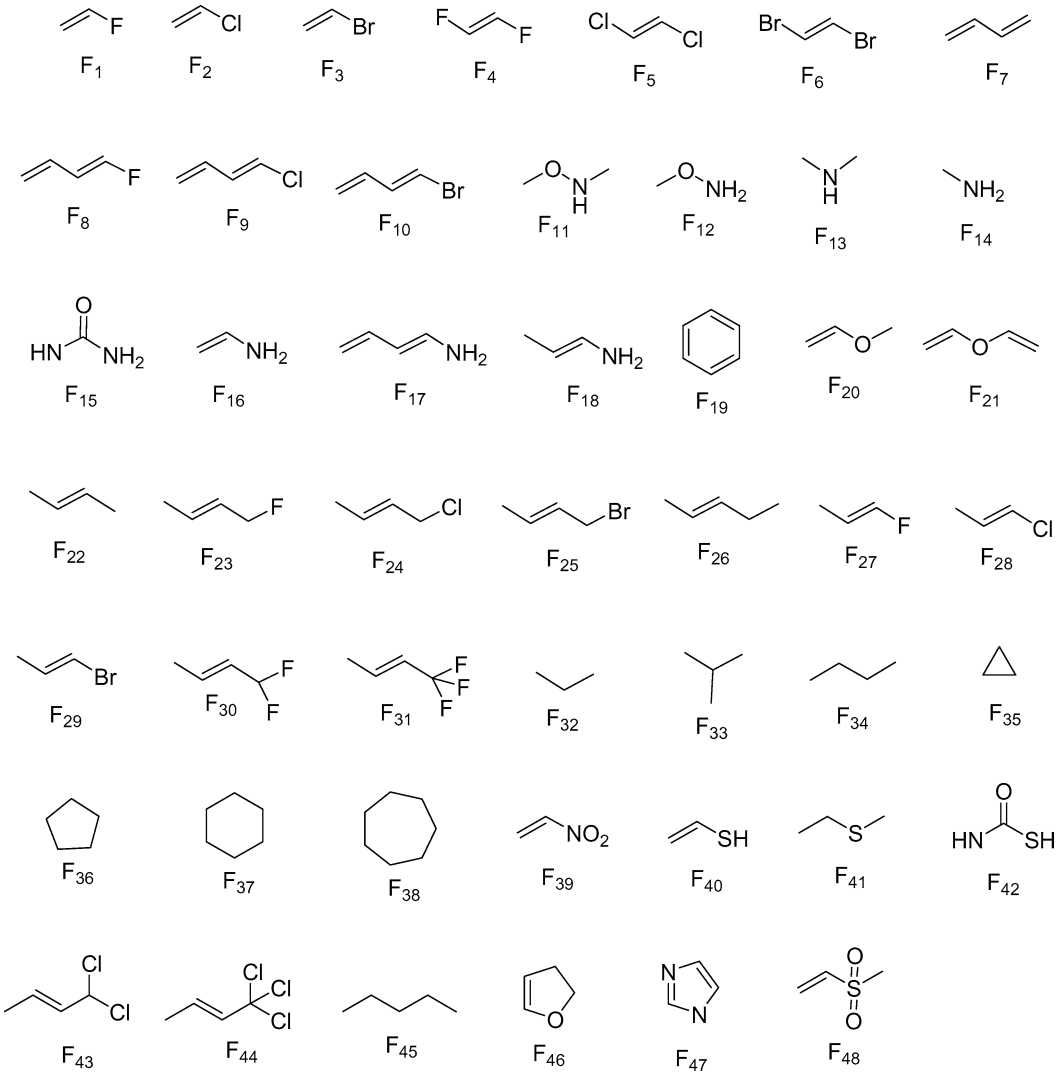

Figure 2. Structures of selected fragments whose contributions to soil sorption coefficients were calculated in this study.

Table 5. Contribution of some selected fragments to the soil sorption coefficients.

Fragment	Contribution	Fragment	Contribution	Fragment	Contribution
F_1	0.38	F_{17}	0.32	F_{33}	0.66
F_2	0.59	F_{18}	0.26	F_{34}	0.65
F_3	0.68	F_{19}	1.03	F_{35}	0.52
F_4	0.49	F_{20}	0.37	F_{36}	0.83
F_5	0.89	F_{21}	0.55	F_{37}	1.02
F_6	1.06	F_{22}	0.55	F_{38}	1.19
F_7	0.47	F_{23}	0.64	F_{39}	0.32
F_8	0.58	F_{24}	0.74	F_{40}	0.59
F_9	0.78	F_{25}	0.93	F_{41}	0.87
F_{10}	0.87	F_{26}	0.69	F_{42}	0.54
F_{11}	0.17	F_{27}	0.52	F_{43}	1.19
F_{12}	−0.25	F_{28}	0.72	F_{44}	2.05
F_{13}	0.03	F_{29}	0.80	F_{45}	0.82
F_{14}	−0.16	F_{30}	0.82	F_{46}	0.60
F_{15}	0.00	F_{31}	1.00	F_{47}	0.20
F_{16}	0.10	F_{32}	0.49	F_{48}	0.51

Table 6. Other published models with Log K_{ow} as a molecular descriptor.

Authors	Chemicals	N	Model descriptors	R^2
Sabljic et al.	Carbamates	43	Log K_{ow}	56.8
Gerstl et al.	Carbamates	39	Log K_{ow}	86.3
Sabljic et al.	Organophosph	41	Log K_{ow}	72.6
Sabljic et al.	Phenylureas	52	Log K_{ow}	61.6
Gerstl et al.	Triazines	16	Log K_{ow}	89.5
Sabljic et al.	Het. pesticides	216	Log K_{ow}	68.1

hydrophobicity, not molecular size, determines the level of sorption.

Nevertheless, although the relation log K_{oc} vs log K_{ow} has been used to explain this phenomenon, specific interactions with soils and sediments (hydrogen bonding, dipole interactions, charge transfer, etc.) which are feasible for chemicals such as alkyl ureas, amines, alcohols, organic acids, amides, and dinitroanilines cannot be adequately described by log K_{ow} alone since this global molecular descriptor accounts primarily for nonspecific interactions resulting from dispersive forces.

Thus, in order to improve the quality of estimates for these chemicals classes, other molecular descriptors which reflect more specific interactions must be used in substitutions to log K_{ow}. These descriptors include spectral moments, WHIM, and molecular connectivity indices.

As compared with fragment F_{37}, F_{19} exhibited a smaller contribution of the benzene ring to hydrophobicity; however, this fragment showed a higher sorption due to its interaction with the positive fraction of colloids and the dominant organic material of the soil.

Other examples of the predominance of electronic interactions of the compounds in this series were observed in fragments F_{13}, F_{14}, and F_{16} in which the level of the contribution was $F_{16} > F_{13} > F_{14}$. This arrangement reflects the ability of these amines to form hydrogen bonds. For example, fragment F_{16}, which corresponds to amine, has a conjugated double bond; hence it stimulates and strengthens this type of interaction. It therefore exerts a negative influence on the contribution of F_{14} in sorption. This leads to the conclusion that the ability of pesticides to form hydrogen bonds with groundwater prevents their soil sorption.

Finally, when the number of halogens is increased in one of the fragments of the families (F_{23}, F_{30}, F_{31}) or (F_{24}, F_{43}, F_{44}), a remarkable increase in the soil sorption coefficient is observed. The interest in this phenomenon has been keen for some time due to environmental pollution considerations. We recently demonstrated how increasing the halogen atoms in a chemical structure generally increase the herbicidal property of a compound [16]. But if this brings with it a higher soil sorption coefficient, with the consequent increase in difficulty of their elimination from and biodegradation in the soil, this property not desirable for new herbicides.

The issue is thus more complex than would seem at first glance. Precisely for this reason, a combination of models is necessary for solving these problems in the future.

Concluding remarks

A topological approach (TOPS-MODE) was used to predict the sorption coefficient in soils of pesticides. The theoretical pattern revealed that the dipole moment, standard distance, atom polarizability, and hydrophobicity are important factors in predicting the soil sorption coefficient of this set of compounds. This model explains more than 85% of the variance in the experimental activity with good predictive power. This last feature is significantly better than that obtained for other methodologies using the same dataset. Finally, taking into consideration the calculation speed and easy interpretation of the descriptors used, as well as the good results obtained in this particular study, it may be worth extending this prediction methodology for the sorption coefficient in soils to other families of active compounds.

Acknowledgments

The authors would like to express their gratitude to the Ministry of Science, Technology, and Environment of Villa Clara City (project 0624) for financial support. In addition, Maykel Pérez González wishes to thank the owners of the software Modeslab 1.0 for donating this valuable tool in order to perform this research.

References

1. Cheng, H., *Pesticides in the soil environment: Processes, Impact, and modeling*, Editor, Soil Science Society of America, Inc., Madison, Wisconsin. 1990.
2. Briggs, G., *Adsorption of pesticides by some Australian soils*, Aust. J. Soil Res., 19 (1981) 61–68.

3. Baker, J.R., Mihelcic, J.R. and Sabljic, A., *Reliable QSAR for estimating K_{oc} for persistent organic pollutants: Correlation with molecular connectivity indices*. Chemosphere, 2 (2001) 213–221.
4. Gousheng, L. and Jianguo, Y. *QSAR analysis of soil sorption coefficients for polar organic chemicals: Substituted anilines and phenols*, Water Research, 10 (2005) 2048–2055.
5. Gawlik, B.M., Sotiriou, N., Feicht, E.A., Schulte-Hostede, S. and Kettrup, A., *Alternatives for the determination of the soil adsorption coefficient K_{oc} of non-ionicorganic compounds – A review*, Chemosphere, 34 (1997) 2525–2551.
6. Estrada, E., *Spectral moments of the edge adjacency matrix in molecular graphs. 1. Definition and applications to the prediction of physical properties of alkanes,* J. Chem. Inf. Comput. Sci., 36 (1996) 844–849.
7. Estrada, E., *Spectral moments of the edge adjacency matrix in molecular graphs. 2. Molecules containing heteroatoms and QSAR applications,* J. Chem. Inf. Comput. Sci., 37 (1997) 320–328.
8. González, M.P., González, H.D., Cabrera, M.A. and Molina, R.R., *A novel approach to predict a toxicological property of aromatic compounds in the Tetrahymena pyriformis*, Bioorg. Med. Chem., 12 (2004) 735–744.
9. Morales, A.H., González, M.P. and Rieumont, J.B., TOPS-MODE *approach to predict mutagenicity in dental monomers*, Polymer, 45 (2004) 2045–2050.
10. González, M.P., Morales, A.H. and Cabrera, M.A., *Quantitative structure activity relationship to predict toxicological properties of benzene derivative compounds*, Bioorg. Med. Chem., 13 (2005) 1775–1781.
11. Morales, A.H., Cabrera, M.A., González, M.P., Molina, R.R., and González, H.D., *A topological substructural approach applied to the computational prediction of rodent carcinogenicity*, Bioorg. Med. Chem., 13 (2005) 2477–2488.
12. Gramatica, P., Corradi, M. and Consonni, V., *Modelling of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors*, Chemosphere., 41 (2000) 763–772.
13. Gutierrez, Y., Estrada, E., TOPS-MODE (1997) (Topological Sub-Structural Molecular Design) for Windows Version 4.0, Universidad de Santiago de Compostela, Spain.
14. Gutierrez, Y. & Estrada, E., Modes Lab®, 2002, version 1.0 b.
15. van der Waterbeemd, H., *Discriminant analysis for activity prediction*, In: R. Manhnhold, Krogsgaard-Larsen & H. Timmerman (Eds.) Method and Principles in Medicinal Chemistry, vol 2, Chemometric methods in molecular design Ed: H. Van Waterbeemd, VCH, Weinhiem. 1995, pp 265–282.
16. González, M.P., González, H., Molina, R., Cabrera, M.A. and Ramos, R., TOPS MODE *Based QSARs derived from heterogeneous series of compounds. Application to the Design of New Herbicides*, J. Chem. Inf. Comput. Sci., 43 (2003) 1192–1199.
17. Ivanciuc, O., Ivanciuc, T. & Balaban, T., *Vertex and edge-weighted molecular graphs and derived structural descriptors*. In: Devillers, J. & Balaban, A.T. (Eds.) Topological Indices and Related Descriptors. Gordon and Breach Sci. Pub., The Netherlands, pp. 169–220.