QSPR Treatment of the Soil Sorption Coefficients of Organic Pollutants

Iiris Kahn, Dan Fara, Mati Karelson, and Uko Maran*

Institute of Chemical Physics, Department of Chemistry, University of Tartu, 2 Jakobi Str., Tartu 51014, Estonia

Patrik L. Andersson†

Institute for Risk Assessment Sciences, Utrecht University, P.O. Box 80176, 3508 TD Utrecht, The Netherlands

Received April 8, 2004

In this study, general and class-specific QSPR models for soil sorption, logK_{OC}, of 344 organic pollutants ($0 < \log K_{OC} < 4.94$) were developed using a large variety of theoretical molecular descriptors based only on molecular structure. Two general models were obtained. The first model was derived for a structurally representative set of 68 chemicals (R^2=0.76, s=0.44), whereas the second involved a total of 344 compounds (R^2=0.76, s=0.41). The first was validated using the data for the remaining 276 pollutants (R^2=0.70, s=0.45). An additional validation of both models was performed using an independent set of 48 pollutants. Both models predict the logK_{OC} at the level of experimental precision, while the theoretical molecular descriptors appearing in the QSPR models give further insight into the mechanisms of soil sorption. The analysis of the distribution of the residuals of the logK_{OC} values calculated by both general models indicated the need and possible advantages of modeling soil sorption for smaller data sets related to individual classes of chemicals. Accordingly, QSPR models were also developed for 14 chemical classes. The descriptors appearing in these models were discussed as related to the possible interaction mechanisms in soil sorption.

INTRODUCTION

The evaluation of the soil mobility of chemicals is one of the primary tasks in estimating the ecological consequences that may follow from their distribution in environment. An important parameter that determines this distribution is the soil/sediment adsorption coefficient, commonly normalized to the content of organic carbon, K_{OC}. This parameter describes the extent to which a chemical is distributed between the solid and solution phases in soil, or between water and sediment in aquatic ecosystems, and indicates whether a chemical is likely to be transported through the soil or would be immobile.

Several experimental methods exist to measure the soil sorption coefficients.[1,2] Due to the large amount of existing and new arriving chemicals to the market, various experimental and theoretical methods have been explored that evaluate soil sorption coefficients and can therefore be used for fast screening and risk assessment of chemicals. The empirical methods usually take advantage of the correlation between K_{OC} and other experimental partition coefficients that can be obtained more easily. In the majority of cases, the experimental octanol/water partition coefficient has been used for this purpose. The alternatives have been the aqueous solubility (the partitioning of the solute between water and the pure solute phase), the chromatographic partition coefficients, bioconcentration factor (the partitioning of the solute between water and biolipids), parachor, Hildebrand solubility parameter, and solvatochromic parameters.[1] One of the most comprehensive analyses in this area has been carried out by Sabljić et al.[3] who evaluated the quality and reliability of the relationships between soil sorption coefficients and *n*-octanol/water partition coefficients (logK_{OC} vs logK_{OW}). By using the data for more than 400 different organic compounds, the authors systematically developed general, subgeneral, and class-specific QSPR (Quantitative Structure Property Relationship) models. As a result, a decision tree was built enabling to use logK_{OW} in the estimation of soil sorption coefficients. The analysis of altogether 19 models lead to the conclusion that specific interactions with soils and sediments (hydrogen bonding, dipole interactions, charge transfer, etc.), inherent for chemicals such as alkyl ureas, amines, alcohols, organic acids, amides, and dinitroanilines, cannot be adequately described by *n*-octanol/water partition coefficients alone. Therefore, the authors concluded that in order to get a better estimate of soil sorption coefficients for these chemicals, other molecular descriptors that reflect more specific interactions must be used together with the *n*-octanol/water partition coefficient.

Theoretical methods that proceed solely from the molecular structure of chemicals can be used to find correlations between logK_{OC} and theoretical molecular descriptors. Within a variety of theoretical molecular descriptors,[4] the molecular connectivity indices (MCIs) have gained high attention. A frequently used descriptor in the estimation of K_{OC} values is the first-order molecular connectivity index ($^1\chi$).[5] However, the applicability of these theoretical indices still needs further examination together with other theoretical descriptors.

A review of more than 200 QSPRs for the estimation of logK_{OC} has been presented by Gawlik et al.[6] The reviewed

* Corresponding author e-mail: uko@chem.ut.ee.
† Current address: Environmental Chemistry, Umeå University, SE-901 87 Umeå, Sweden.

10.1021/ci0498766 CCC: $30.25
Published on Web 12/22/2004

Table 1. List of Recently Published QSPR Models on K_{OC}[a]

no.	compounds	N	model descriptors	R^2	s	F	reference
1	nonpolar compounds	64	$^1\chi$	0.96	0.27	1371	Meylan et al.[7]
2	nonpolar and polar organic compounds	189	$^1\chi$, $\Sigma P_f N$ (f=26)	0.96	0.23		Meylan et al.[7]
3	heterocyclic nitrogen compounds	12	$^1\chi$	0.88	0.38	74	Liao et al.[8]
4	heterocyclic nitrogen compounds	12	LogS_w, $\Delta^1\chi^v$	0.91	0.32	54	Liao et al.[8]
5	heterocyclic nitrogen compounds	12	LogK_{OW}, $\Delta^1\chi^v$	0.87	0.38	38	Liao et al.[8]
6	phenylthio, phenylsulfinyl, phenylsulfonyl	25	Logk'_w	0.93	0.13	320	Hong et al.[9]
7	phenylthio, phenylsulfinyl, phenylsulfonyl	25	LogK_{OW}	0.83	0.21	115	Hong et al.[9]
8	phenylthio, phenylsulfinyl, phenylsulfonyl	25	$^3\chi^v_c$(Ph), $^0\chi$(R_4), $^0\chi$(Ph), $^3\chi_c$(Ph)	0.91	0.15	63	Hong et al.[9]
9	diverse organic compounds	72	LogK_{OW}	0.91			Baker et al.[10]
10	POPs (logK_{OW} > 5)	18	LogK_{OW}(calc.)	0.29	0.59		Baker et al.[11]
11	POPs (logK_{OW} > 5)	18	$^1\chi$, $^4\chi^v_C$, $^3\chi_C$	0.81	0.30	25	Baker et al.[12]
12	diverse organic compounds	66	LogK_{OW}(calc.), V^+, B_{MAX}	0.84	0.38		Müller[13]
13	phthalates	8	S_{ester}, S_{alkyl}	0.82			Thomsen et al.[14]
14	diverse set of reference substances	21	Logk'_w (cyanopropyl phase)	0.91			Szabo et al.[15]
15	diverse set of reference substances	21	Logk'_w (humic acid phase)	0.93			Szabo et al.[15]
16	PCBs	48	RRT	0.92	0.16		Hansen et al.[16]
17	PCBs	48	TSA	0.92	0.17		Hansen et al.[16]
18	PCOCs	65	LogK_{OW}	0.86	0.44	386	Dai et al.[17]
19	PCOCs	65	μ, E_{homo}, qH^+, q^-, TE	0.85	0.46	69	Dai et al.[17]
20	benzaldehydes (AM1)	14	μ, qH^+, $^3\chi_{pc}$, $^2\chi^v_p$	0.91	0.10	35	Dai et al.[18]
21	benzaldehydes (PM3)	14	μ, qH^+, $^3\chi_{pc}$, $^2\chi^v_p$	0.92	0.10	40	Dai et al.[18]
22	benzaldehydes (AM1)	14	μ, qH^+, q^-	0.86	0.13	28	Dai et al.[18]
23	benzaldehydes (PM3)	14	μ, qH^+, q^-	0.82	0.16	19	Dai et al.[18]
24	heterogeneous pesticides	141	MW, nNO, nHA, CIC, MAXDP, Ts	0.84	0.35	120	Gramatica et al.[19]
25	carbamates	29	nO, nX, nNO, ξ^C	0.95	0.17	110	Gramatica et al.[19]
26	organophosphates	28	I^E_{deg}, IC, MAXDP, η1u, Ts	0.89	0.23	35	Gramatica et al.[19]
27	phenylureas	43	MW, nCl, nCIT, λ1v, η2s	0.91	0.12	76	Gramatica et al.[19]
28	triazines	13	$^1\chi$, DELS	0.97	0.08	136	Gramatica et al.[19]
29	diverse organic compounds	592	74 fragment constants 24 structural factors	0.97	0.37		Tao et al.[20]
30	diverse organic compounds	592	$^1\chi^v$, $^2\chi$, $^4\chi_c$, $^6\chi$ ΣP_j (j=17)	0.77	0.44		Tao et al.[20]
31	diverse organic compounds	387	5 σ moments	0.71	0.62[b]	189	Klamt et al.[21]
32	substituted aromatic compounds	28	LogK_{OW}	0.61	0.22	43	Wu et al.[22]
33	substituted aromatic compounds	28	$^2\chi^v$, $\Delta^5\chi^v$	0.69	0.20	31	Wu et al.[22]
34	substituted aromatic compounds	28	MW, π*, V^-, EN	0.95	0.08	128	Wu et al.[22]
35	substituted aromatic compounds	27	LogK_{OW}	0.79	0.08	92	Wu et al.[23]
36	substituted aromatic compounds	27	LogK_{OW}, $^3\chi_c$	0.88	0.06	91	Wu et al.[23]
37	substituted aromatic compounds	27	α, π*, O	0.86	0.07	48	Wu et al.[23]
38	diverse organic pesticides	143	$^1\chi$, 11 E-state indices (S_i)	0.82	0.37	51	Huuskonen[24]
39	diverse organic compounds	403	Log$S$(calc.)	0.80	0.51	1622	Huuskonen[25]
40	diverse organic compounds	403	Log$S$(calc.), HBA, NAR, MW, I_{acid}	0.85	0.44	451	Huuskonen[25]
41	diverse organic compounds	403	LogK_{OW}(calc.)	0.79	0.52	1475	Huuskonen[25]
42	diverse organic compounds	403	LogK_{OW}(calc.), NAR, ROT, MW, I_{acid}	0.86	0.43	491	Huuskonen[25]
43	organic compounds containing C,H,N,O,S	82	N_ϕ, M_W, N_N, N_O, N_S	0.94	0.33	228	Delgado et al.[26]

[a] N − number of compounds used to develop a model, R^2 − correlation coefficient, s − standard deviation, F − F-test value. [b] RMS (root-mean-square). PCOCs − polychlorinated organic compounds; POPs − persistent organic pollutants; PCBs −polychlorinated biphenyls; $P_f N$ − P_j-structural fragment contribution factors (P_f) for polar structural fragments, N - number of times the fragment occurs in the structure. $^n\chi$ − molecular connectivity indices; S_w − water solubility; $\Delta^1\chi^v$- nondisperse force factor; K_{OW} − n-octanol/water partition coefficient; V^+ − potential of the positive atomic charges; B_{MAX} − maximum charge difference between connected atoms; k′ − HPLC capacity factor; S_{ester}, S_{alkyl} − group electrotopological indices; RRT − gas chromatographic relative retention time; TSA − molecular total surface area; μ − dipole moment; E_{homo} − energy of the highest occupied molecular orbital; qH^+ − most positive net atomic charge on hydrogen atom; q^- − largest negative atomic charge on an atom; TE − total energy; MW − molecular weight; nNO − number of NO bonds or groups; nHA − number of hydrogen bond acceptor atoms; CIC − complementary information content index; MAXDP − maximum positive intrinsic state difference; Ts − global WHIM descriptor of molecular size; nO, nCl, nX − numbers of O, Cl and halogen atoms; ξ^C − eccentric connectivity index; I^E_{deg} − mean information content on vertex degree equality; IC − information content on multigraph; η1u, λ1v,η2s − directional WHIM descriptors; nCIT − number of total rings; DELS − index, mainly related to total charge transfer in the molecule; σ moments − real solvents sigma-moment descriptors; E-state indices (S_i) − electrotopological-state indices; V^- − potential of the negative atomic charges; EN − electronegativity; α − polarizability; π* − α/Connolly accessible volume; O − ovality of a molecule; HBA − number of N and O atoms; NAR − number of aromatic rings; I_{acid} − indicator variable for ionization of carboxylic acids; ROT − number of rotational bonds; N_ϕ − number of benzene rings; N_N, N_O, N_S − numbers of N, O, and S atoms.

models were grouped according to the nature of the descriptors involved, showing that logK_{OC} values were most frequently modeled with water solubility (S_w), n-octanol/water partition coefficient (K_{OW}), RP-HPLC capacity factor (k'), topological indices, or linear solvation energy parameters. In addition, some attempts were made to combine various above-mentioned descriptors. It was concluded that logK_{OW} was most commonly used to describe soil sorption. Most of the presented QSPRs were class- and soil-specific and therefore lack generality. In addition to the QSPRs given in the review,[6] several new contributions have appeared in the literature as summarized in Table 1.[7−26] Table 1 shows that research has been carried out toward finding descriptors that can complement the frequently used empirical descriptors or to be alternatives for the empirical descriptors, to be able to capture the influence of polar functional groups to the sorption. The examples of these descriptors are as follows: the constitutional (Table 1: #43),[26] topological

(Table 1: #4, 5, 8, 11, 20, 21, 30, 36, 38),[8,9,12,18,20,23,24] quantum chemical (Table 1: #12, 19−23, 34, 37),[13,17,18,22,23] and weighted holistic invariant molecular (WHIM) descriptors (Table 1: #24−27).[19] Also, the fragment contribution approach has quite successfully been used to model soil sorption for large data sets (Table 1: #2, 29).[7,20] Gramatica et al. used genetic algorithms for the selection of relevant descriptors to the models.[19] Besides linear models there were two attempts to model soil sorption with neural networks.[24,27]

A somewhat different approach was proposed by Winget et al.[28] who used the results of quantum mechanical calculations to develop a set of effective solvent descriptors using SM5 solvational parametrization to characterize the organic carbon component of the soil. These descriptors were subsequently used to develop QSPR models to be applied in partitioning of solutes between soil and air. The combination of this set of effective solvent descriptors with solute atomic surface tension parameters developed for water/air and organic solvent/air partitioning allows the prediction of the partitioning of solutes between soil and water.

The lack of complete and homogeneous soil sorption data lead in the late 1980s and early 1990s to the development of the European reference soil set (the EUROSOILS).[29,30] Gawlik et al. predicted logK_{OC} values with the capacity factor from HPLC measurements on the first and the second generation (produced several years later from the same soils) EUROSOILS soil sets.[31,32] The R^2-values for the QSPRs of logK_{OC} in the five reference soils varied from 0.794 to 0.884 for the first and from 0.817 to 0.916 for the second generation. It was concluded that "the possibility of maintaining the principal soil properties in a second generation enables the permanent use of these soils as reference materials for soil-related studies". Based on the analysis of the results and data from these two studies, it was realized that normalization by the organic carbon content may cause false interpretation in case of the soils rich in clay, which is a strong sorbent for many compounds. Therefore, the use of non-normalized $K_{f/D}$-values referring to the reference soils would be preferable. Also, Gerstl[33] has emphasized the importance of sorption to clay and other mineral surfaces in soils, especially in the soils with a small fraction of organic matter.

The present work is an extension of our earlier study in this field,[34] where partial least squares (PLS) and principal component analysis (PCA) were used to compare the applicability of five molecular descriptor sets for the modeling of soil sorption. The analysis showed that logK_{OW} describes most of the variation in all sets. However, this descriptor alone is not sufficient for obtaining the prediction at experimental quality equally for all classes of organic chemicals and inclusion of additional variables is necessary. An example of variable selection considerably improved the quality of the model.[34] In the following we aimed to model the soil sorption coefficient using a large database of only theoretical molecular descriptors. An automatic forward selection of descriptors was applied for the variable selection during the development of multilinear regression models.

DATA SET AND METHODOLOGY

The data set of experimental K_{OC} (partition coefficient of a compound between water phase and soil, normalized to organic carbon) investigated consists of 344 compounds[3,34] (Table A in Supporting Information). As compared to the data set used in our earlier work,[34] the compound #163, fluazifop butyl, was excluded from analysis because the soil sorption values given in different literature sources disagree largely with each other.[3,25,35] According to the chemical structure, the data set was divided into 14 classes:[3] acetanilides (01), alcohols (02), amides (03), anilines (04), carbamates (05), dinitroanilines (06), esters (07), nitrobenzenes (08), organic acids (09), phenols and benzonitriles (10), phenylureas (11), phosphates (12), triazines (13), and triazoles (14).

The structures of the compounds were drawn and optimized using the Merck Molecular Force Field (MMFF)[36,37] as implemented in MacroModel 7.0-program.[38] Conformational search was carried out for every compound in isolated state, using the Monte Carlo Multiple Minimum (MCMM) method.[39,40] The molecular geometries corresponding to the lowest energy conformers were refined using AM1 semiempirical parametrization[41] and eigenvector following geometry optimization algorithm[42] implemented in the quantum chemical program package MOPAC 6.0.[43] The gradient norm limit 0.01 kcal/Å was applied to the geometry optimization.

The CODESSA program[44,45] was used to calculate various constitutional (atom counts, etc.), geometrical (molecular volume, shadow indexes, etc.), topological (information content, Kier and Hall, Randic, etc), electrostatic (partial charges, charged partial surface areas, etc.), and quantum-chemical descriptors (polarizabilities, energy partitioning, reactivity indexes, etc.)[4,46] based on MOPAC results. The octanol/water partition coefficient (logK_{OW}) for every compound was calculated with the Internet version of the KowWin program[47] using an atom/fragment contribution method[48] and added to the pool of descriptors (Table A in Supporting Information). In total 743 descriptors were calculated. The best multilinear regression (BMLR) procedure[44,45,49,50] was applied to find the best correlation models. During the BMLR procedure the pool of descriptors is cleaned from insignificant descriptors ($r^2 < 0.1$) and the descriptors with missing values followed by the construction of the best two-parameter regression, the best three-parameter regression, etc. based on the statistical significance and noncollinearity criteria ($r^2 < 0.6$) of the selected descriptors to the equation. In BLMR, the descriptor scales are normalized and centered automatically, and the final result is given in natural scales. The final model has the best representation of the property in the given descriptor pool within the given number of parameters.

QSPR models were derived for the full set of 344 compounds, a representative set of 68 compounds and for the chemical classes. The representative set of 68 compounds was composed by Eriksson et al.[51] using multivariate design and factorial design on the principal properties of the compound classes based on 64 chemical descriptors. It was found that it is sufficient to allocate only 20% of the available compounds into the training set. Accordingly, the set of 68 compounds and the test set of 276 compounds were used for the development and validation of the QSPR model.

Leave-one-out cross-validation, internal validation (leave-1/3-out), and external validation were used to validate the models for the set of 68 compounds and the full data set.

Table 2. Experimental and Calculated Soil Sorption Coefficients (logK_{OC}) and Calculated Octanol/Water Partition Coefficients (logK_{OW}) for the Independent External Validation Set: A − Calculated with Model for the Set of 344 Compounds Table 5 and B − Calculated with Model for 68 Compounds in Table 3

no.	class	CAS	name	LogK_{OC} (exp)	A (344)	Δ_A	B (68)	Δ_B	LogK_{OW}(calc)
1	01	709−98−8	propanil	2.17	2.49	0.32	2.56	0.39	2.88
2	01	2307−68−8	pentanochlor	2.76	2.88	0.12	3.11	0.35	4.18
3	02	115−32−2	dicofol	3.70	4.32	0.62	3.78	0.08	5.81
4	02	4780−79−4	1-naphthalenemethanol	2.17	2.27	0.10	2.39	0.22	2.25
5	02	103−74−2	2-pyridineethanol	1.45	1.27	−0.18	1.32	−0.13	0.38
6	03	86−86−2	1-naphthaleneacetamide	2.00	2.03	0.03	2.14	0.14	1.72
7	03	957−51−7	diphenamid	2.32	2.32	0.00	2.40	0.08	2.86
8	03	26644−46−2	triforine	2.30	2.59	0.29	2.16	−0.14	2.02
9	04	134−32−7	1-naphthalenamine	3.51	2.30	−1.21	2.44	−1.07	2.25
10	04	101−77−9	4,4-methylenedianiline	1.99	2.19	0.20	2.24	0.25	2.18
11	04	60−09−3	4-aminoazobenzene	2.79	2.66	−0.13	2.87	0.08	3.19
12	04	122−66−7	1,2-diphenylhydrazine	2.98	2.61	−0.37	2.57	−0.41	3.06
13	05	2686−99−9	trimethacarb	2.60	2.21	−0.39	2.37	−0.23	2.81
14	05	79127−80−3	fenoxycarb	3.00	3.10	0.10	3.13	0.13	4.24
15	05	22781−23−3	bendiocarb (ficam)	2.75	2.12	−0.63	2.38	−0.37	2.55
16	06	55283−68−6	ethalfluralin	3.60	3.83	0.23	3.69	0.09	5.23
17	06	62924−70−3	flumertalin	4.00	4.49	0.49	4.11	0.11	6.09
18	06	33820−53−0	isopropalin	4.00	3.91	−0.09	3.94	−0.06	5.80
19	06	40487−42−1	pendimethalin	3.70	3.48	−0.22	3.51	−0.19	4.82
20	07	32357−46−3	2,4-DB butoxyethyl ester	2.70	3.40	0.70	3.48	0.78	5.08
21	07	2122−70−5	ethyl 1-naphthyl acetate	2.48	2.79	0.31	3.02	0.54	3.75
22	07	510−15−6	chlorobenzilate	3.30	3.13	−0.17	2.96	−0.34	3.99
23	07	6422−86−2	bis(2-ethylhexyl) terephthalate	4.16	4.66	0.50	4.87	0.71	8.39
24	07	85−44−9	1,3-isobenzofurandione	1.56	2.04	0.48	2.09	0.53	2.07
25	07	103−23−1	bis(2-ethylhexyl) adipate	4.19	4.08	−0.11	4.26	0.07	8.12
26	08	42874−03−3	oxyfluorofen	5.00	3.89	−1.11	3.62	−1.38	5.21
27	08	42576−02−3	bifenox	4.00	3.20	−0.80	3.12	−0.88	4.15
28	08	99−30−9	dicloran (botran)	3.00	2.76	−0.24	2.60	−0.40	2.76
29	09	86−87−3	1-naphthaleneacetic acid	2.20	2.41	0.21	2.52	0.32	2.60
30	09	115−28−6	chlorendic acid	2.79	3.00	0.21	2.54	−0.25	3.14
31	09	81334−34−1	imazapyr acid	2.00	1.88	−0.12	1.98	−0.02	1.57
32	10	56−53−1	diethylstilbestrol	4.14	3.58	−0.56	3.76	−0.38	5.64
33	10	90−15−3	1-naphthol	2.72	2.46	−0.26	2.58	−0.14	2.69
34	10	842−07−9	1-(phenylazo)-2-naphthol	3.58	3.84	0.26	3.83	0.25	5.51
35	10	100−54−9	3-cyanopyridine	1.65	1.31	−0.34	1.32	−0.33	0.35
36	11	90982−32−4	chlorimuron-ethyl	2.04	2.48	0.44	2.33	0.29	2.29
37	11	36734−19−7	iprodione	2.85	2.58	−0.27	2.47	−0.38	2.85
38	12	30560−19−1	acephate	0.30	0.69	0.39	0.94	0.64	−0.90
39	12	732−11−6	phosmet	2.91	2.54	−0.37	2.61	−0.30	2.48
40	12	961−11−5	stirofos	3.07	3.15	0.08	2.99	−0.08	3.81
41	12	55−38−9	fenthion	3.18	3.30	0.12	3.38	0.20	4.08
42	12	35400−43−2	sulprofos	4.08	4.16	0.08	4.23	0.15	5.65
43	12	6923−22−4	monocrotophos	0.00	0.32	0.32	0.53	0.53	−1.31
44	13	101−05−3	anilazine	3.00	3.33	0.33	2.91	−0.09	3.64
45	13	66215−27−8	cyromazine	2.30	1.53	−0.77	1.55	−0.75	0.96
46	13	673−04−1	simetone	2.34	2.20	−0.14	2.32	−0.02	2.73
47	14	2593−15−9	etridiazole	3.00	2.90	−0.10	2.80	−0.20	3.60
48	14	51707−55−2	thidiazuron	2.04	2.26	0.22	2.34	0.30	2.10

For the internal validation, the data set was divided into three subsets using the same order of compounds as in Table A in Supporting Information (the 1st, 4th, 7th . . . etc. entries go into the first subset, the 2nd, 5th, and 8th . . . etc. into the second subset, and the 3rd, 6th, 9th, . . . etc. into the third subset). Two of the subsets were combined into one in three different combinations, and the correlation equations were derived for these sets with the same descriptors as of the original model. The obtained equations were used to predict the remaining subsets not used in the development of these models.

An additional validation set of 48 compounds was constructed from the publication by Huuskonen[25] in order to have an independent external validation for both QSPR models: the 68 representative compounds and the full data set. Depending on the availability of the data, 2 to 6 chemicals not present in our data set were selected from each chemical class as presented in Table 2. The standard error of the predictions was compared to the error of the experimental measurements.

RESULTS AND DISCUSSION

The Set of 68 Representative Compounds. For the training set of 68 compounds, initially, a three-descriptor model was selected according to the improvement in the statistical criteria (R^2 and R^2_{CV}) with the increase in the number of descriptors in the model. In the following search for the best predictive QSPR, the test set of 276 compounds was used to predict the property with the obtained models containing 1, 2, and 3 descriptors. The correlation coefficient of the observed versus calculated soil sorption coefficient of the test set for these models suggested the model with 2 descriptors. The parameters for this model are given in Table 3 and the respective plot in Figure 1.

The *logarithm of calculated n-octanol/water partition coefficient*, logK_{OW}, covers the range from −1.1 to 8.39. Its

Table 3. Best Two-Parameter Correlation for Training Set of 68 Chemicals ($F = 100.5$)[a]

no.	X	$\pm\Delta X$	t-test	R^2	R^2_{CV}	s	descriptor
0	2.8841E+00	3.5903E−01	8.0329				intercept
1	4.3608E−01	3.5683E−02	12.2209	0.6579	0.6323	0.5153	LogK_{OW}
2	−3.6063E−01	7.0726E−02	−5.0989	**0.7557**	**0.7288**	**0.4389**	η [AM1]

[a] R^2 correlation coefficient, R^2_{CV} cross-validated correlation coefficient, s standard deviation, F F-test value.

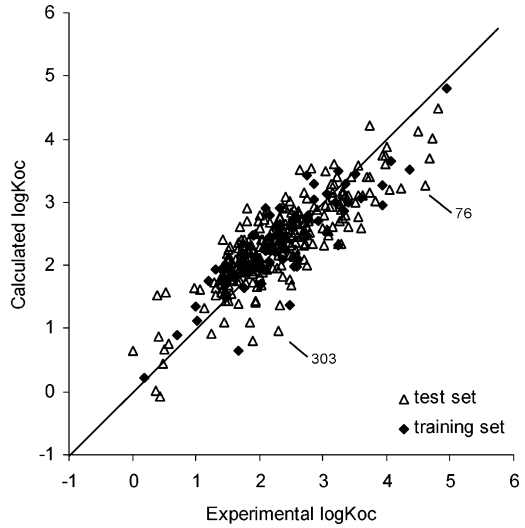

Figure 1. Calculated vs experimental soil sorption coefficient (logK_{OC}) of the two-parameter model for the training set of 68 compounds and test set of 276 compounds. Compounds #76 and 303 have the largest residuals.

dominance in the model is expected because it describes intermolecular interactions similar to those involved in soil sorption. Compounds with high logK_{OC} values also have high logK_{OW} values, being consequently more hydrophobic.

The second descriptor, the *absolute hardness*, η, has been defined by Parr and Pearson[52,53] as the second derivative of the energy with respect to the number of electrons, and it is also related to the electronic chemical potential. Within the molecular orbital theory, the absolute hardness is related to the energy gap between the LUMO and HOMO energies and thus depends on the closeness of the frontier orbitals. The negative correlation of soil sorption coefficient with this descriptor indicates that soil sorption coefficients decrease with the increase in absolute hardness. This reflects the situation where the compounds with higher chemical stability interact less with the soil constituents. The absolute hardness can also be related to the polarizability of the compounds, since the decrease of the energy gap usually leads to easier polarization of the molecule.[54]

The cross-validated correlation coefficient of the two-parameter model developed for the training set of 68 compounds (R^2_{CV}=0.7288) and the correlation coefficient obtained through the internal validation (R^2=0.7268) (Table 4) were close to the R^2 of the model itself. The validation of the data with the test set of 276 compounds resulted in the squared correlation coefficient, $R^2 = 0.7035$ and the standard deviation of $s = 0.4512$ (see the respective plot in Figure 1). The logK_{OC} for compounds with the highest residuals, pentachloroaniline (#76) and mevinphos (#303), were underestimated by more than 1.2 log units. The removal of these outliers improved the validation statistics: R^2 =0.7168 and $s = 0.4381$. An important criterion of the quality of the model is the comparison of its error with the error of the respective experimental measurements. Lohninger[35] has calculated the standard deviation of the experimental logK_{OC} values from different experiments for 70 pesticides with the average of about 0.44 log units (from the range of 0.09−1.22). The standard deviation, s, of the prediction for the validation set was close to the experimental error. An additional external validation with the independent set of 48 compounds in Table 2 resulted in the correlation coefficient, $R^2 = 0.7988$, and standard deviation, $s = 0.4474$, that is also comparable with the limits of the experimental error.

Table 4. Internal Validation of the Two-Parameter QSPR Model for 68 Chemicals[a]

set to fit	R^2 (fit)	s (fit)	set to predict	R^2 (pred.)	s (pred.)
#2 and #3	0.7685	0.4803	#1	0.7752	0.4035
#1 and #3	0.8043	0.3828	#2	0.6976	0.5770
#1 and #2	0.7038	0.4475	#3	0.8374	0.4656
			#1, #2, #3	**0.7268**	**0.4647**

[a] As in Table 3.

Comparison with the results from the PLS analysis[34] shows that the forward selection of variables considerably improved the statistical and validation parameters of the model. This is an indication that too many molecular descriptors in the PLS model add noise that lowers the model performance. A variety of techniques of variable selection exist that could be used in connection with QSPR analysis.[55]

Full Set of Compounds. A QSPR model was developed for the full set of 344 compounds in order to see how the extended structural diversity will be reflected by the descriptors. Considering the squared correlation and cross-validated correlation coefficients of the generated models, 4 descriptors, logK_{OW}, the *partial negative surface area*, the *absolute hardness*, and the *maximum* π-π *bond order*, were significant for the description of logK_{OC} (Table 5 and the respective plot in Figure 2). According to this model, the largest part of variation (66%) in the soil sorption coefficients is described by logK_{OW}. The additional three descriptors improved the squared correlation coefficient from 0.66 to 0.76. Three outliers to this model were detected. The largest negative residual (−1.51 log units) belongs to mevinphos (#303−12), followed by dalapon (#177−09), and dicrotophos (#315−12), with positive and negative residuals of 1.23 log units, respectively.

Two descriptors in the model for this data set, logK_{OW} and the *absolute hardness*, η, were also present in the model for the representative set of 68 compounds. Their physical meaning as related to soil sorption was explained above.

The second descriptor in the model (Table 5), the *partial negative surface area*, *PNSA-1*, belongs to the group of charged partial surface area (CPSA) descriptors,[56] which describe the projection of the charge distribution on the surface area of the molecules. The charge distribution for this descriptor was calculated using Zefirov's scheme based

Table 5. Best Four-Parameter Correlation for the Full Set of 344 Chemicals ($F = 266.39$)[a]

no.	X	$\pm\Delta X$	t-test	R^2	R^2_{CV}	s	descriptor
0	2.1560E+00	2.3519E−01	9.1671				intercept
1	4.2399E−01	1.7834E−02	23.7747	0.6606	0.6561	0.4826	LogK_{OW}
2	2.7206E−03	4.2455E−04	6.4081	0.7139	0.7080	0.4437	*PNSA-1* [Zefirov]
3	−2.4130E−01	4.0664E−02	−5.9341	0.7366	0.7292	0.4264	η [AM1]
4	−4.0438E−01	7.2723E−02	−5.5606	**0.7586**	**0.7506**	**0.4088**	$P^{max}_{\pi-\pi}$ [AM1]

[a] As in Table 3.

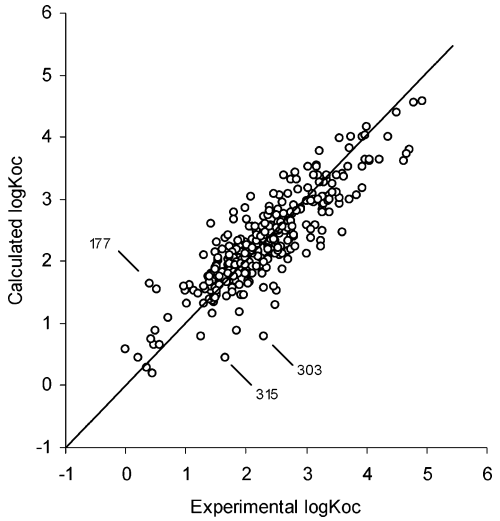

Figure 2. Calculated vs experimental soil sorption coefficient (logK_{OC}) of the four-parameter model for 344 compounds. Compounds #177, 303, and 315 have the largest residuals.

Table 6. Internal Validation of the Four-Parameter QSPR Model for 344 Chemicals[a]

set to fit	R^2 (fit)	s (fit)	set to predict	R^2 (pred.)	s (pred.)
#2 and #3	0.7411	0.4225	#1	0.7925	0.3923
#1 and #3	0.7681	0.3941	#2	0.7410	0.4486
#1 and #2	0.7684	0.4117	#3	0.7356	0.4126
			#1, #2, #3	**0.7546**	**0.4122**

[a] As in Table 3.

on Sanderson's electronegativity equalization principle.[57] The positive contribution of *PNSA-1* to the model suggests that the compounds with larger negatively charged surface area interact stronger with the organic components of the soil and result in higher sorption.

The last term in the four-parameter equation (Table 5), the *maximum* π-π *bond order*, $P^{\max}_{\pi-\pi}$, belongs to the group of valence-related descriptors.[58] It is calculated directly from the Mulliken charge partition scheme and measures the extent of sharing π-electrons between two atoms, pointing to the most stable bond in the molecule. The negative contribution of this descriptor improves the correlation coefficient from 0.74 to 0.76.

The statistical characteristics of the developed model were comparable to those from cross-validated correlation analysis (Table 5) and internal validation (Table 6). The standard deviation of 0.4122 log units of the internal validation of the four-parameter model (Table 6) lies close to the experimental error of the K_{OC} measurements.[35] Finally, the predictive ability of the model (Table 5) was tested with the independent set of 48 compounds (Table 2) to give a standard deviation of $s = 0.4413$, which also corresponds to the limits of the experimental error. The relatively high correlation coefficient between experimental and predicted logK_{OC} (R^2=0.8163) compared to the developed model (Table 5) can be explained by the higher correlation of the test compounds with logK_{OW}.

The small differences between the cross-validated, internal validation, and the correlation coefficients of the model demonstrate higher stability of this model compared to the model based on the representative set of 68 compounds. Therefore, this model would be more reliable to use for prediction. The applicability of this QSPR for prediction is determined by the classes of compounds (with the molecular weight of 32 to 420) used for its development. The range of logK_{OW}, the most influential parameter of the model, should also be considered. According to the values used for the development of the model we suggest the range of −1 to 7 log units for the best results. The first reason for lowering the limit from 8.39 to 7 is the small number of compounds with logK_{OW} over 7 log units in our data set. The second reason is the possibility of larger uncertainty in the logK_{OW} measurements for compounds with logK_{OW} higher than this value.[10]

Estimation of logK_{OC} for the Independent Chemical Classes by the General Models. Further, the predictive and descriptive capabilities of the developed two- and four-parameter QSPR models (Tables 3 and 5, respectively) were studied for the 14 individual chemical classes. The respective linear correlation coefficients of the calculated vs experimental logK_{OC} are given in Table 7.

As seen from the R^2 values in Table 7, the differences in the precision of predictions between the different classes appear to be quite large, which limits the application of these models for certain chemical classes. For a number of the classes, the estimated R^2 lies below 0.7, that cannot be considered satisfactory. The poorest results according to the R^2 value were obtained for the series of amides (03). The analysis of the residuals (by the mean residuals in Table 7) has revealed trends in systematic over- or underestimation of soil sorption by the two models for many classes of compounds. A systematic overestimation of logK_{OC} is particularly expressed in the case of organic acids (09). The logK_{OC} values are systematically underestimated for the group of dinitroanilines (06) by the two-parameter model. The underestimation is also evident in the group of esters (07) and phenols and benzonitriles (10). In these cases, the inclusion of an additional correction term to the model could improve the result.

Correlation of logK_{OW} within the Chemical Classes. The previous QSPRs (Tables 5 and 3) had logK_{OW} as the dominant term. The results presented in Table 8 demonstrate that only alcohols (#2 in Table 8), dinitroanilines (#6), and esters (#7) have a high correlation between logK_{OC} and

Table 7. Linear Correlation Coefficients of Relationships between the Predicted vs Experimental Data According to the Models in Tables 3 and 5 for the Individual Chemical Classes[a]

no.	chemical class	R^2 (training set)	estimation (mean residual)	outlier	R^2 (full set)	estimation (mean residual)	outlier
1	acetanilides	0.6107 (0.8257)	0.20	norfluorazon (#17)	0.7585 (0.8730)	0.08	norfluorazon (#17)
2	alcohols	0.8795	−0.06		0.8970	0.12	
3	amides	0.4160 (0.4988)	0.13	2,6-dichlorobenzamide (#50) 3,5-dinitrobenzamide (#51)	0.5615 (0.6780)	−0.03	2,6-dichlorobenzamide (#50) diethylacetamide (#54)
4	anilines	0.7844	−0.16		0.8274 (0.8405)	−0.16	pentachloroaniline (#76) 2,6-dichloroaniline (#70)
5	carbamates	0.7111 (0.8197)	0.18	asulam (#104)	0.7649 (0.8601)	0.01	asulam (#104)
6	dinitroanilines	0.8093	−0.29		0.7668	−0.06	
7	esters	0.8471	−0.15		0.8629 (0.8915)	−0.28	di-*n*-hexylphthalate (#145) ethyl-octanoate (#153)
8	nitrobenzenes	0.7632	−0.14		0.8290 (0.9357)	0.05	dinoseb (#167)
9	organic acids	0.7665	0.54		0.7195	0.55	
10	phenols and benzonitriles	0.7423 (0.7694)	−0.22	2,3,5-trimethylphenol (#214)	0.6783 (0.7327)	−0.20	2,3,5-trimethylphenol (#214)
11	phenylureas	0.5695	0.12		0.6307 (0.7647)	0.03	VEL-3510 (#265) isouron (#266)
12	phosphates	0.5867 (0.7095)	0.08	mevinphos (#303) trichlorfon (#312) dicrotophos (#315)	0.6424 (0.7380)	0.02	mevinphos (#303) dicrotophos (#315)
13	triazines	0.6587 (0.8822)	0.14	NIA-23486 (#331)	0.5987 0.7542)	0.01	NIA-23486 (#331)
14	triazoles	0.7485 (0.8189)	−0.11	thiabendazol (#333)	0.7196 (0.7912)	−0.19	thiabendazol (#333)
	all compounds	0.7142 (0.7245)	0.05	pentachloroaniline (#76) mevinphos (#303)	0.7586 (0.7778)	0.00	mevinphos (#303) dalapon (#177) dicrotophos (#315)

[a] The R^2 in parentheses have been calculated after removal of the indicated outliers.

Table 8. Single-Parameter Correlations with logK_{OW} for the 14 Individual Chemical Classes[a]

no.	class	R^2	R^2_{CV}	s	F	N	range of logK_{OC}	σ^b logK_{OC}	range of logK_{OW}
	full data set	0.6606	0.6561	0.4826	665.75	344	0.00−4.94	0.83	−1.10−8.39
1	acetanilides	0.5246	0.4537	0.3585	19.86	20	1.40−3.28	0.49	1.10−4.84
2	alcohols	0.8123	0.7405	0.3565	47.60	13	0.20−2.59	0.76	−0.78−3.79
3	amides	0.3359	0.2417	0.5022	12.65	27	0.53−3.32	0.59	−0.18−3.98
4	anilines	0.7836	0.7072	0.3726	65.17	20	1.41−4.62	0.76	1.08−4.30
5	carbamates	0.6479	0.5922	0.3697	73.61	42	0.42−3.35	0.61	−0.78−4.57
6	dinitroanilines	0.8099	0.7750	0.2502	72.45	19	2.37−4.01	0.54	0.95−5.62
7	esters	0.8459	0.8170	0.3709	120.75	24	1.60−4.94	0.90	1.66−8.39
8	nitrobenzenes	0.7774	0.6884	0.5047	27.93	10	1.73−4.36	0.96	1.47−5.03
9	organic acids	0.7155	0.6427	0.3571	52.82	23	0.00−3.28	0.64	0.09−4.23
10	phenols and benzonitriles	0.7552	0.7022	0.3678	67.86	24	0.98−3.73	0.71	1.03−4.74
11	phenylureas	0.5205	0.4635	0.3800	54.27	52	1.29−3.83	0.54	0.71−4.15
12	phosphates	0.6161	0.5680	0.5148	60.99	40	1.20−4.67	0.81	−1.10−6.34
13	triazines	0.6384	0.5188	0.2575	24.71	16	1.70−3.07	0.40	1.44−4.22
14	triazoles	0.6246	0.5365	0.5053	19.96	14	1.25−3.73	0.76	−0.47−4.81

[a] As in Table 3. [b] Standard deviation.

logK_{OW}. Four sets of data, anilines (#4), nitrobenzenes (#8), organic acids (#9), and phenols and benzonitriles (#10), have R^2 between 0.7 and 0.8. The remaining seven subsets have single-parameter correlations with logK_{OW} with R^2 below 0.7. Consequently, additional descriptors are needed for many chemical classes to complement or to be alternatives for logK_{OW} in the QSPR models for better description of logK_{OC}.

For 297 compounds in our full set, experimental logK_{OW} data are available in the database of KowWin.[47] The correlation between the experimental and calculated logK_{OW} for these compounds gave $R^2 = 0.94$ and $s = 0.33$. The class with the poorest correlation between experimental and calculated logK_{OW} was phenylureas (#11), with $R^2 = 0.84$, the rest of the classes had R^2 above 0.90.

Next, the correlation of logK_{OC} with experimental logK_{OW} values was attempted. For the 50 phenylureas (#11), the experimental logK_{OW} performed considerably better than the calculated data, giving an improvement in the R^2 from 0.61 to 0.78. The correlation with the experimental logK_{OW} was considerably poorer for only the 14 triazines (#13), that had the squared correlation coefficient of 0.71 for experimental logK_{OW} compared to the 0.83 obtained by using the calculated logK_{OW} from KowWin.

This means that in addition to the experimental error of the soil sorption measurements, in some cases there is also a considerable error originating from the calculation of the logK_{OW}. In the current set of data such error is most significant in the case of phenylureas.

Table 9. Statistical Characteristics of Class-Specific QSPR Models[a]

no.	class	N	N_{desc}	R^2	R^2_{cv}	s
01	acetanilides	20	2	0.9328	0.9005	0.1386
02	alcohols	13	1	0.9371	0.9069	0.2064
03	amides	27	3	0.8733	0.8239	0.2287
04	anilines	20	2	0.8969	0.8627	0.2646
05	carbamates	42	3	0.8955	0.8711	0.2066
06	dinitroanilines	19	2	0.9393	0.9212	0.1456
07	esters	24	2	0.8913	0.8663	0.3189
08	nitrobenzenes	10	1	0.9653	0.9475	0.1992
09	organic acids	23	3	0.8620	0.8056	0.2613
10	phenols and benzonitriles	24	3	0.8805	0.8239	0.2696
11	phenylureas	52	3	0.8483	0.8270	0.2182
12	phosphates	40	4	0.8500	0.8014	0.3353
13	triazines	16	2	0.9479	0.9122	0.1015
14	triazoles	14	2	0.9206	0.8913	0.2427

[a] As in Table 3.

Models for Individual Chemical Classes. The statistics and model parameters of the QSPRs of logK_{OC} for the 14 independent chemical classes are presented in Tables 9 and 10, respectively. The notations for the descriptors are provided in the Supporting Information (Table B). Only the models with the squared correlation coefficients above 0.8 were selected for the further analysis. In the case of many alternative models, the models with the highest cross-validated correlation coefficients and relative significance of the descriptors according to the *t*-test were selected.

Notably, for the classes with more than 20 compounds, the best single-descriptor correlations with logK_{OC} were obtained with logK_{OW} (8 classes). The single-parameter correlations for other chemical classes involved topological, geometrical, or charged partial surface area (CPSA) descriptors. The best multiparameter models for most classes contained at least one descriptor belonging to one of these groups. The additional descriptors in the multiparameter models were mostly obtained from quantum-chemical calculations (Table 10).

For the group of amides (03) that had the lowest correlation with logK_{OW} (Table 8), the logK_{OC} was also very poorly estimated by both general models (Table 7). This group was modeled with the combination of a CPSA descriptor, a shadow index and a quantum chemical site-specific energy descriptor (Tables 9 and 10: #03). Such a combination suggests a complex sorption mechanism for the members of this group. From the descriptors appearing in this equation, the $^{HA}HDSA\text{-}1_{TMSA}$ describes the relative abundance of hydrogen-bonding donor centers in the molecule and relates thus to the solubility in hydrogen-bonding media. The bulkier shape described by the shadow index, $ZX_{shadow}/ZX_{rectangle}$, causes the increase of the sorptive affinity of the molecule. Finally, some contribution to the sorption may be due to the chemical stability or reactivity of the molecule as reflected by the maximum resonance energy of a C−O bond, $E_R^{\max}(C-O)$. This may also be the reason logK_{OW} fails to model logK_{OC} for the set of amides.

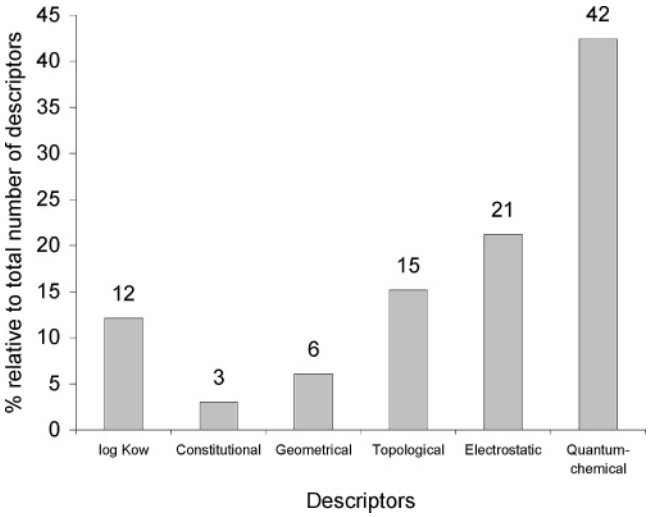

Figure 3. Relative distribution of different types of descriptors used in the models of chemical classes (total number of descriptors is 33).

In the model development for the chemical classes the pool of descriptors can be enlarged and target-specified. This is achieved by the inclusion of class-specific descriptors that for instance express properties of certain bonds or atom types related to a given class of compounds. These descriptors are normally excluded for the diverse data sets. Figure 3 provides a summary on the relative distribution of the different types of descriptors in the models of the chemical classes.

In the class-specific models, logK_{OW} was involved only in four chemical classes (Figure 3 and Table 10: #5, 7, 9, 12). This indicates that for the selected sets of chemically similar compounds, alternative descriptors can give better

Table 10. Parameters of QSPR Models (logK_{OC} =)[a]

no.	c_0	c_1	d_1	c_2	d_2	c_3	d_3	c_4	d_4
01	−9.065	0.0286	WNSA-1	0.152	$E_{ee}^{min}(C)$				
02	0.143	1.001	$^3\chi^v$						
03	−13.333	−11.193	$^{HA}HDSA-1_{TMSA}$	4.905	$ZX_{shadow}/ZX_{rectangle}$	0.490	$E_R^{max}(C-O)$		
04	485.93	0.863	$^2\chi^v$	−4.703	$E_{ee+en}^{max}(C)$				
05	−4.713	0.475	logK_{OW}	−33.165	q_N^{min}	1.900	$\#e_n^{occ}$		
06	8.000	−19.553	$\#O_{rel}$	−0.301	$^3\chi$				
07	60.172	0.484	logK_{OW}	−59.496	V_H^{max}				
08	2.006	−0.00748	DPSA-1						
09	−5.679	0.439	logK_{OW}	2.810	B_{MO}^{max}	0.00388	$DPSA-2$		
10	7.291	5.391	q_{max}-q_{min}/r^2	−4.394	RNCG	−0.199	$E_{tot}^{min}(C-C)$		
11	−60.678	0.571	$^1\chi^v$	7.064	$E_R^{min}(H-N)$	−36.687	V_H^{avg}		
12	93.356	0.439	logK_{OW}	−94.649	$P_{\sigma-\sigma}^{max}$	0.216	$^2\kappa$	−0.494	$E_C^{min}(C-H)$
13	9.625	1.176	ϵ_{HOMO-1}	−0.00107	E_{ne}(tot)				
14	−10.551	0.0470	XY_{shadow}	1.654	$E_C^{min}(C-N)$				

[a] c_0 intercept, c_n coefficients to the descriptors d_n in the QSPRs.

Table 11. Descriptors Classified According to Their Origin[a]

	electron distribution			
size	hybrid		hybrid	reactivity
${}^3\chi^v$(2)	*WNSA*−1 [Zefirov] (1)	q_N^{min} [Zefirov] (5)	$E_{ee}^{min}(C)$ [AM1] (1)	V_H^{max} [AM1] (7)
$ZX_{shadow}/ZX_{rectangle}$ (3)	${}^{HA}HDSA-1_{TMSA}$ [Zefirov] (3)	$\#e_n^{occ}$ [AM1] (5)	$E_R^{max}(C-O)$ [AM1] (3)	V_H^{avg} [AM1] (11)
${}^2\chi^v$ (4)	LogK_{OW} (5)	*RNCG* [Zefirov] (10)	$E_{ee+en}^{max}(C)$ [AM1] (4)	$P_{\sigma-\sigma}^{max}$ [AM1] (12)
$\#O_{rel}$ (6)	LogK_{OW} (7)		B_{MO}^{max} [AM1] (9)	ϵ_{HOMO-1} [AM1] (13)
${}^3\chi$ (6)	*DPSA*−1 [Zefirov] (8)		$E_{tot}^{min}(C-C)$ [AM1] (10)	
${}^1\chi^v$(11)	LogK_{OW} (9)		$E_R^{min}(H-N)$ [AM1] (11)	
${}^2\kappa$ (12)	*DPSA*−2 [Zefirov] (9)		$E_C^{min}(C-H)$ [AM1] (12)	
XY_{shadow} (14)	q_{max}-q_{min}/r^2 [Zefirov] (10)		$E_{ne}(tot)$ [AM1] (13)	
	LogK_{OW} (12)		E_C^{min} $(C-N)$ [AM1] (14)	

[a] (n) − number of subset.

QSPR models for logK_{OC}, the alcohols (02), anilines (04), and dinitroanilines (06) being vivid examples.

The constitutional descriptors appeared only once in the class-specific models (Table 10 and Figure 3). The *relative number of oxygen atoms, $\#O_{rel}$,* in the model for dinitroanilines (06), is an important structural characteristic being the second best single descriptor for this set after logK_{OW}. The negative correlation of soil sorption with this descriptor indicates that the compounds relatively richer in oxygen have smaller affinity for sorption.

The geometrical descriptors were represented by two shadow indices[59] (Table 10: #03, 14) that encode the 3D-geometry of compounds through the projections on geometrical planes. The *XY shadow index, XY_{shadow},* in the model of triazoles (14) reflects the geometrical size in the direction of the longest and shortest axes of the molecule, whereas the normalized shadow index, $ZX_{shadow}/ZX_{rectangle}$, in the model for amides (03) reflects the distortion of the geometrical shape of the molecule. Both descriptors gave positive contribution to the soil sorption coefficient.

Topological descriptors have gained much attention in the estimation of soil sorption coefficients.[5] In this study they evolved in five models (Table 10: #02, 04, 06, 11, 12) that makes 15% of the descriptors of the models (Figure 3). These descriptors have been designed to account for atomic connectivity, size, and branching as well as for the presence of heteroatoms and the hybridization of atoms.[60] Our descriptor set included molecular connectivity indices from zeroth to the third order. The mainly congeneric set of 13 alcohols (Table 10: #02) exhibited a well-defined relationship with the *Kier and Hall valence connectivity index of third order*, ${}^3\chi^v$. In the homologous series of alcohols (from methanol to decanol), soil sorption is increasing with the increase in the size of the molecule, ranging from 0 to 2.5 log units. In four of the models with topological descriptors, the positive correlation shows an increase in sorption with the increase in the value of the descriptor.

The electrostatic descriptors contributed 21% relative to the total number of descriptors in models for the chemical classes (Figure 3) indicating the importance of charge distribution induced processes (polarization, H-bonding, reactivity, etc.) in soil sorption. For the set of electrostatic descriptors, the charge distribution was calculated by Zefirov's scheme. From among these descriptors, the involvement of 4 charged partial surface area (CPSA) descriptors[56,61] (Table 10: #01, 03, 08, 09) that encode both information about the surface area and the respective charge distribution could be emphasized.

The quantum chemical descriptors gave the highest contribution to the QSPRs (Figure 3). They became involved for all subsets with relatively higher structural diversity and number of compounds, and most often they appear in combinations with logK_{OW} or topological connectivity indices. The most frequently represented quantum-chemical descriptors have site-specific character (Table 10: #01, 03, 04, 10−14), encoding intramolecular energy distribution, bond strength, and polarity/polarizability. These descriptors can be related to the various interactions taking place at sorption or chemical reactivity with the active sites on humic substances for permanent binding. The latter is exemplified by the involvement of quantum-chemical energy and valence (reactivity) descriptors (Table 10: #07, 11−13) that contain information about the stability or reactivity of the molecules or the location of the specific electrophilic or nucleophilic centers.[46]

The analysis of the distribution of different types of descriptors in QSPRs for different classes of compounds can be useful for elucidating the types of physical and chemical interactions that govern in soil sorption. The structural factors of intermolecular interaction mechanisms in condensed media can be generally addressed as related to size, shape, and electron distribution. The reactivity related structural properties can also be considered. In Table 11, we have classified the descriptors involved in the class models for chemical classes according to these concepts. Notably, the descriptors that are designed to combine two structural factors (the hybrid descriptors) are particularly successful.

The interaction mechanisms between chemicals and soil have been widely discussed in the literature. Organic compounds may be sorbed by soil organic matter through physical or chemical binding: ionic, hydrogen, and covalent bonding, charge-transfer or electron donor−acceptor mechanisms, dipole−dipole or van der Waals forces, ligand exchange, and cation and water bridging. An organic molecule may be sorbed initially by sites that provide the strongest binding, followed then by progressively weaker sites as the stronger adsorption sites become filled. Once adsorbed, the chemical may be subject to other processes that can affect retention. Some compounds may further react to become covalently or irreversibly bound, while others may become physically trapped into the humic matrix. The latter depends on the shape and structural conformation of the

chemical.[62,63] In addition it must be considered that the solid phase (the organic matter of soil) and solvent phase (water) have different hydrogen bonding properties. Most probably the formation of hydrogen bonds between the solute and solvent are stronger than between the solute and the solid phase. The solvent−solute interactions may also support the formation of a rather tightly bound solvent layer on the solute molecule that enables the solute to float faster through the soil. In addition, water, as an important part of the partitioning system, is responsible for degradation of some compounds, such as esters and amides, by hydrolysis.

Baker et al.[12] have proposed a hypothetical mechanism by which nonionic chemicals interact with the soil's organic matter. The interactions depend on the organic matter's 3D-structure and functional groups that lead to three possible types of interactions with the organic fraction of soil: nonspecific hydrophobic, nonspecific polar, and specific interactions of functional groups. Notably, the descriptors involved in the present QSPR models are in accordance with this hypothetical mechanism. The dominance of the size and shape related descriptors to account for the nonspecific diffusion into the porous matrix of the humic substances is accompanied by a strong influence of polar or charge related descriptors. To a lesser extent, the descriptors pointing to the reactivity of the molecules were also involved.

CONCLUSIONS

The first model derived for a representative set of 68 compounds contained two descriptors, logK_{OW} and the absolute hardness. This model was validated with the cross-validated correlation analysis (leave-one-out method), internal validation scheme (leave-1/3-out), and two validation sets: the test set of 276 and an independent set of 48 compounds. All validation schemes showed that the training set had satisfactory constitution and the developed QSPR model predicts soil sorption coefficients with comparable quality to the experimental error of 0.44 log units. The inclusion of the absolute hardness in the model to complement logK_{OW} resulted in the improvement in description and estimation of the studied property compared to the results obtained from the single-parameter correlations with logK_{OW}. Compared to our previous models[34] with PLS analysis the forward selection of variables used to choose the descriptors into the model improved considerably the results of fitting and predicting.

The second model was derived for the total set of 344 mainly polar, structurally diverse compounds. The best model included four parameters and described the soil sorption with experimental quality. The model was validated with cross-validated correlation analysis and the internal validation scheme. The predictive ability of the model was tested by internal validation (leave-1/3-out) and the independent set of 48 contaminants for which the obtained results were close to the experimental error. This model contains more diverse structural information, and its performance was improved by the two additional descriptors compared to the first model. We recommend this model as a general model for the prediction of the soil sorption values. The applicability of this QSPR for prediction is determined by the classes of compounds used for its development with the molecular weight of up to 420 and logK_{OW} in the range of −1 to 7 log units. The validation showed that amides, organic acids, and phosphates are not accurately modeled by this QSPR.

The QSPR models were developed to describe soil sorption for fourteen individual classes of compounds formed according to their chemical functionality. These models revealed wide applicability of various descriptors that explicitly describe bulk (size) and electron distribution of the compounds. Approximately half of the theoretical descriptors in the QSPR models developed were of quantum-chemical origin, the conventional logK_{OW} being present for only four chemical classes.

Based on the descriptor content of all the models the following can be concluded about the structural properties of the compounds that influence soil sorption. For the structurally diverse data sets hydrophobicity is the major characteristic describing a complex of physical interactions. The two other dominating structural properties are the size and shape of the compounds and the charge distribution, that often appear as alternatives to the hydrophobicity. A larger size and bulkier shape favor nonspecific interactions with the soil constituents and the humic matrix. The charge distribution describes nonspecific polar and specific interactions either with water while floating through the soil or with the soil keeping them from leaching into the groundwater. The presence of reactivity indices in the QSPR models for some chemical classes indicates that chemical reactivity affects soil sorption.

ACKNOWLEDGMENT

Financial support is greatly acknowledged from the 5th framework EU research training network entitled "Intelligent Modeling Algorithms for the General Evaluations of TOXicities" (IMAGETOX) (HPRN-CT-1999-00015) and Estonian Science Foundation (Grants #5805 and #4548). We thank Joop Hermens for the discussion and valuable comments during the preparation of the manuscript. Also Martin Müller, Fraunhofer Institute for Environmental Chemistry and Ecotoxicology, Schmallenberg, Germany and Aleksandar Sabljić, Institute Ruđjer Bošković, Zagreb, Croatia are acknowledged for providing the K_{OC} database.

Supporting Information Available: Experimental and calculated soil sorption coefficients (logK_{OC}) and calculated octanol/water partition coefficients (logK_{OW}) (Table A) and abbreviations (Table B). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Lyman, W. J. In *Handbook of Chemical Property Estimation Methods*; Lyman, W. J., Reehl, W. F., Rosenblatt, D. H., Eds.; McGraw-Hill Book Company: New York, 1982; Chapter 4.

(2) Wauchope, R. D.; Yeh, S.; Linders, J. B. H. J.; Kloskowski, R.; Tanaka, K.; Rubin, B.; Katayama, A.; Kördel, W.; Gerstl, Z.; Lana, M.; Unsworth, J. B. Review. Pesticide Soil Sorption Parameters: Theory, Measurement, Uses, Limitations and Reliability. *Pest Management Sci.* **2002**, *58*, 419−445.

(3) Sabljić, A.; Güsten, H.; Verhaar, H.; Hermens, J. QSAR Modelling of Soil Sorption. Improvements and Systematics of logK_{OC} vs. logK_{OW} Correlations. *Chemosphere* **1995**, *31*, 4489−4514.

(4) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; John Wiley & Sons: New York, U.S.A., 2000.

(5) Sabljić, A. QSAR Models for Estimating Properties of Persistent Organic Pollutants Required in Evaluation of their Environmental Fate and Risk. *Chemosphere* **2001**, *43*, 363−375.

(6) Gawlik, B. M.; Sotiriou, N.; Feicht, E. A.; Schulte-Hostede, S.; Kettrup, A. Alternatives for the Determination of the Soil Adsorption Coefficient, K_{OC}, of Non-Ionicorganic Compounds − A Review. *Chemosphere* **1997**, *34*, 2525−2551.
(7) Meylan, W.; Howard, P. H.; Boethling, R. S. Molecular Topology Fragment Contribution Method for Predicting Soil Sorption Coefficients. *Environ. Sci. Technol.* **1992**, *26*, 1560−1567.
(8) Liao, Y.-Y.; Wang, Z.-T.; Chen, J.-W.; Han, S.-K.; Wang, L.-S.; Lu, G.-Y.; Zhao, T.-N. The Prediction of Soil Sorption Coefficients of Heterocyclic Nitrogen Compounds by Octanol/Water Partition Coefficient, Water Solubility, and by Molecular Connectivity Indices. *Bull. Environ. Contam. Toxicol.* **1996**, *56*, 711−716.
(9) Hong, H.; Wang, L.; Han, S.; Zhang, Z.; Zou, G. Prediction of Soil Adsorption Coefficient K_{OC} for Phenylthio, Phenylsulfinyl and Phenylsulfonyl Acetates. *Chemosphere* **1997**, *34*, 827−834.
(10) Baker, J. R.; Mihelcic, J. R.; Luehrs, D. C.; Hickey, J. P. Evaluation of Estimation Methods for Organic Carbon Normalized Sorption Coefficients. *Water Environ. Res.* **1997**, *69*, 136−145.
(11) Baker, J. R.; Mihelcic, J. R.; Shea, E. Estimating K_{OC} for Persistent Organic Pollutants: Limitations of Correlations with K_{OW}. *Chemosphere* **2000**, *41*, 813−817.
(12) Baker, J. R.; Mihelcic, J. R.; Sabljic, A. Reliable QSAR for Estimating K_{OC} for Persistent Organic Pollutants: Correlation with Molecular Connectivity Indices. *Chemosphere* **2001**, *41*, 213−221.
(13) Müller, M. Quantum Chemical Modelling of Soil Sorption Coefficients: Multiple Linear Regression Models. *Chemosphere* **1997**, *35*, 365−377.
(14) Thomsen, M.; Rasmussen, A. G.; Carlsen, L. SAR/QSAR Approaches to Solubility, Partitioning and Sorption of Phthalates. *Chemosphere* **1999**, *38*, 2613−2624.
(15) Szabo, G.; Guczi, J.; Kördel, W.; Zsolnay, A.; Major, V.; Keresztes, P. Comparison of Different HPLC Stationary Phases for Determination of Soil−Water Distribution Coefficient, K_{OC}, Values of Organic Chemicals on RP-HPLC System. *Chemosphere* **1999**, *39*, 431−442.
(16) Hansen, B. G.; Paya-Perez, A. B.; Rahman, M.; Larsen, B. R. QSARs for K_{OW} and K_{OC} of PCB Congeners: A Critical Examination of Data, Assumptions and Statistical Approaches. *Chemosphere* **1999**, *39*, 2209−2228.
(17) Dai, J.; Sun, C.; Han, S.; Wang, L. QSAR for Polychlorinated Organic Compounds (PCOCs). I. Prediction of Partition Properties for PCOCs Using Quantum Chemical Parameters. *Bull. Environ. Contam. Toxicol.* **1999**, *62*, 530−538.
(18) Dai, J.; Xu, M.; Wang, L. Prediction of Octanol/Water Partitioning Coefficient and Sediment Sorption Coefficient for Benzaldehydes by Various Molecular Descriptors. *Bull. Environ. Contam. Toxicol.* **2000**, *65*, 190−199.
(19) Gramatica, P.; Corradi, M.; Consonni, V. Modelling and Prediction of Soil Sorption Coefficients of Non-Ionic Organic Pesticides by Molecular Descriptors. *Chemosphere* **2000**, *41*, 763−777.
(20) Tao, S.; Lu, X. X.; Cao, J.; Dawson, R. A. Comparison of the Fragment Constant and Molecular Connectivity Indices Models for Normalized Sorption Coefficient Estimation. *Water Environ. Res.* **2001,** *73*, 307−313.
(21) Klamt, A.; Eckert, F.; Diedenhofen, M. Prediction of Soil Sorption Coefficients with a Conductor-Like Screening Model for Real Solvents. *Environ. Toxicol. Chem.* **2002,** *21*, 2562−2566.
(22) Wu, C.-D.; Wei, D.-B.; Liu, X.-H.; Wang, L.-S. Estimation of the Sorption of Substituted Aromatic Compounds on the Sediment of the Yangtse River. *Bull. Environ. Contam. Toxicol.* **2001**, *66*, 777−783.
(23) Wu, C. D.; Wei, D. B.; Hu, G. P.; Wang, L. S. Estimation of the Sorption of Substituted Aromatic Compounds onto Modified Clay. *Bull. Environ. Contam. Toxicol*. **2003**, *70*, 513−519.
(24) Huuskonen, J. Prediction of Soil Sorption Coefficient of Organic Pesticides from the Atom-Type Electrotopological State Indices. *Environ. Toxicol. Chem.* **2003**, *22*, 816−820.
(25) Huuskonen, J. Prediction of Soil Sorption Coefficient of a Diverse Set of Organic Chemicals from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1457−1462.
(26) Delgado, E. J.; Alderete, J. B.; Jana, G. A. A Simple QSPR Model for Predicting Soil Sorption Coefficients of Polar and Nonpolar Organic Compounds from Molecular Formula. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1928−1932.
(27) Gao, C.; Govind, R.; Tabak, H. H. Predicting Soil Sorption Coefficients of Chemicals Using a Neural Network Model. *Environ. Toxicol. Chem.* **1996**, *15*, 1089−1096.
(28) Winget, P.; Cramer, C. J.; Thrular, D. G. Prediction of Soil Sorption Coefficients Using Universal Solvation Model. *Environ. Sci. Technol.* **2000**, *34*, 4733−4740.
(29) Kuhnt, G.; Muntau, H. European Reference Soils for Sorption Testing. *Fresenius Environ. Bull.* **1992**, *1*, 589−594.
(30) Kuhnt, G.; Muntau, H. Eds.; EUROSOILS − Identification, Collection, Treatment, Characterization, European Commission, Special Publication No. 1.94.60, Ispra, 1994; p 154.
(31) Gawlik, B. M.; Feicht, E. A.; Kracher, W.; Kettrup, A.; Muntau, H. Application of the European Reference Soil Set (EUROSOILS) to a HPLC−Screening Method for the Estimation of Soil Adsorption Coefficients of Organic Compounds. *Chemosphere* **1998**, *36*, 2903−2919.
(32) Gawlik, B. M.; Kettrup, A.; Muntau, H. Estimation of Soil Adsorption Coefficients of Organic Compounds by HPLC Screening Using the Second Generation of the European Reference Soil Set. *Chemosphere* **2000**, *41*, 1337−1347.
(33) Gerstl, Z. Quantitative Structure−Activity Relationships (QSARs) as a Tool for Predicting the Sorption of Organic Chemicals in Soils. *Isr. J Chem.* **2002**, *42*, 55−65.
(34) Andersson, P. L.; Maran, U.; Fara, D.; Karelson, M.; Hermens, J. L. M. General and Class Specific Models for Prediction of Soil Sorption Using Various Physico-Chemical Descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1450−1459.
(35) Lohninger, H. Estimation of Soil Partition Coefficients of Pesticides from their Chemical Structure. *Chemosphere* **1994**, *29*, 1611−1626.
(36) Halgren, T. A. Merck Molecular Force Field. I.-V. *J. Comput. Chem.* **1996**, *17*, 490−519, 520−552, 553−586, 587−615, 616−641.
(37) Halgren, T. A. MMFF VII. Characterization of MMFF94, MMFF94s, and Other Widely Available Force Fields for Conformational Energies and for Intermolecular-Interaction Energies and Geometries. *J. Comput. Chem.* **1999**, *20*, 730−748.
(38) MacroModel version 7.0, Schrödinger, Inc., Portland 2000.
(39) Chang, G.; Guida, W. C.; Still, W. C. An Internal-Coordinate Monte Carlo Method for Searching Conformational Space. *J. Am. Chem. Soc.* **1989**, *111*, 4379−4386.
(40) Saunders: M.; Houk, K. N.; Wu, Y. D.; Still, W. C.; Lipton, M.; Chang, G.; Guida, W. C. Conformations of Cycloheptadecane. A Comparison of Methods for Conformational Searching. *J. Am. Chem. Soc*. **1990**, *112*, 1419−1427.
(41) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(42) Baker, J. An Algorithm for the Location of Transition States. *J. Comput. Chem.* **1986**, *7*, 385−395.
(43) Stewart, J. J. P. *MOPAC 6.0 Program Package*; QCPE, No 455, 1989
(44) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure. *Chem. Soc. Rev.* **1995**, *24*, 279−287.
(45) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *CODESSA: Reference manual (version 2.0)*; Gainesville, FL, 1994.
(46) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chem. Rev*. **1996**, *96*, 1027−1043.
(47) KowWin On-Line Version: http://esc.syrres.com/interkow/kowdemo.htm
(48) Meylan, W. M.; Howard, P. H. Atom/fragment Contribution Method for Estimating Octanol−Water Partition Coefficients. *J. Pharm. Sci*. **1995**, *84*, 83−92.
(49) Draper, N. R.; Smith, H. *Applied Regression Analysis;* Wiley: New York, 1981.
(50) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and Test Set of 9 Simple Inorganics. *J. Phys. Chem.* **1996**, *100*, 10400−10407.
(51) Eriksson, L.; Johansson, E.; Müller, M.; Wold, S. On the Selection of the Training Set in Environmental QSAR Analysis when Compounds are Clustered. *J. Chemometrics* **2000**, 14, 599−616.
(52) Parr, R. G.; Pearson R. G. Absolute Hardness: Companion Parameter to Absolute Electronegativity. *J. Am. Chem. Soc*. **1983**, *105*, 7512−7516.
(53) Pearson, R. G. The Principle of Maximum Hardness. *Acc. Chem. Res.* **1993**, *26*, 250−255.
(54) Pearson, R. G. Absolute Electronegativity and Hardness: Applications to Organic Chemistry. *J. Org. Chem.* **1989**, *54*, 4(6), 1423−1430.
(55) Maran, U.; Sild, S. In *Artificial Intelligence Methods and Tools for Systems Biology*; Dubitzky, W., Azuaje, F., Eds.; Kluwer Academic Publishers: Boston-Dordrecht-London, 2004; Chapter 2, pp 19−36.
(56) Stanton, D. T.; Jurs, P. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem*. **1990**, *62*, 2323−2329.
(57) Zefirov, N. S.; Kirpichenok, M. A.; Izmailov, F. F.; Trofimov, M. I. Scheme for the Calculation of the Electronegativities of Atoms in a Molecule in the Framework of Sanderson's Principle. *Dokl. Akad. Nauk* (*Engl. Transl.*) **1987**, *296*, 883−887.
(58) Sannigrahi, A. B. Ab Initio Molecular Orbital Calculations of Bond Index and Valency. *Adv. Quantum Chem.* **1992**, *23*, 301−351.

(59) Rohrbaugh, R. H.; Jurs, P. C. Description of Molecular Shape Applied in Studies of Structure/Activity and Structure/Property Relationships. *Anal. Chim. Acta* **1987**, *199*, 99−109.
(60) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis;* John Wiley and Sons: New York, 1986.
(61) Stanton, D. T.; Jurs, P. C. Computer-Assisted Study of the Relationship Between Molecular Structure and Surface Tension of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 109−115.
(62) Senesi, N. In: *Migration and Fate of Pollutants in Soils and Subsoils*; Petruzzelli, D., Helfferich, F. G., Eds.; Springer-Verlag: Berlin, Heidelberg, 1993; pp 47−74.
(63) Roy, W. R. In *Migration and Fate of Pollutants in Soils and Subsoils*; Petruzzelli, D., Helfferich, F. G., Eds.; Springer-Verlag: Berlin, Heidelberg, 1993; pp 169−188.

CI0498766