Prediction of Soil Sorption Coefficient of a Diverse Set of Organic Chemicals From Molecular Structure

Jarmo Huuskonen*

Pharmaceutical Chemistry Division, Department of Pharmacy, POB 56, 00014 University of Helsinki, Finland

Received October 3, 2002

A correlation study based on simple structural descriptors for predicting the soil sorption coefficient, log K_{oc}, of a diverse set of 568 organic compounds is presented. Using a training set of 403 compounds, in which the log K_{oc} values were in the range 0−6.5, multiple linear regression (MLR) was utilized to build the models. The models were validated using a test set of 165 chemicals not included in the training set. The statistics for a linear regression model with calculated aqueous solubility, log *S*, were $r^2 = 0.80$ and $s = 0.51$ in the training set, and $r^2 = 0.76$ and $s = 0.61$ in the test set. The model parameters used allow rapid and accurate calculation of log K_{oc} values for a diverse set of organic chemicals, and propose the importance of molecular solubility, lipophilicity, size, flexibility, and ionization for a chemicals' sorption to organic soil material.

INTRODUCTION

Recent advances in solid-phase chemistry, especially combinatorial chemistry and parallel synthesis, have made possible the synthesis of a large number of new agrochemicals every day. Thus, the development of efficient and inexpensive methods for testing and estimating the biological and toxicological properties of new chemicals is of major significance. This would enable pre-screening of the potential fates of the compounds, such as biodegradation and soil sorption, and allow effective risk assessment. The experimental measurement of a chemical's soil sorption is a difficult, expensive, and time-consuming task. Hence, it would be beneficial to derive a method for estimating soil sorption of an organic chemical from known structural characteristics. This method could be used in the development of new and safe agrochemical compounds, and in the assessment of the fate of those chemicals in the environment.

The soil sorption partition coefficient normalized to organic carbon, K_{oc}, is an important parameter in the assessment of the environmental fate of organic chemicals. Several approaches have been developed for the prediction of this parameter. However, most of them use experimental molecular properties, i.e., water solubility, octanol/water partition coefficient (log *P*), or RP-HPLC capacity factor (k'), or are based on small structurally related series of chemicals. Good reviews of models for K_{oc} estimation of nonionic organic chemicals were compiled by Gawlik et al.[1] and Wauchope et al.[2]

Only a few attempts have been made in which a large and diverse set of organic pesticides was used in the modeling and prediction of K_{oc} from their chemical structure. An extensive QSAR analysis for correlation of soil sorption coefficient, log K_{oc}, and octanol/water partition coefficient, log *P*, was made by Sabljic.[3] He concluded that correlations were reasonable when structurally related chemicals were considered. It has been also indicated that the log *P* was not be a strong predictor of log K_{oc} for very lipophilic chemicals which have a log *P* value >5,[4] and adapting the more comprehensive polyparameter models rather than trying to refine existing one-parameter models in the prediction of log K_{oc}.[5] Alternative modeling schemes are based on topological indices and fragment-based polarity factors [6−10] or parameters calculated solely from molecular structure such as molar volume, molecular surface area, shape of molecules, and electronic quantum chemical descriptors.[11−14] Methods based on topological indices and atomic fragments are attractive because they are simple and give fast and accurate estimations for diverse organic compounds. However, the difficulties of these methods are how to break a molecule, especially a large pesticide molecule, into basic fragments. Quantum chemical parameters are properties of the entire solute molecule. Although these parameters give promising results for many molecules, usually for those that are structurally related, their merit is not yet sufficient for use as a general estimation method. In addition, the quantum chemical calculations are very time-consuming if their estimation for a large number of compounds is required.

In the present study, a diverse set of 568 organic chemicals was used in multiple linear regression and prediction of soil sorption coefficient values from easily calculated structural parameters.

METHODS

The soil sorption coefficients, log K_{oc} values, for a diverse set of 568 organic chemicals were compiled from the literature.[3,8,15] This data set was divided randomly into a training set of 403 chemicals and a test set of 165 chemicals. The experimental soil sorption coefficients, log K_{oc} values, ranged from 0 to 6.5 in the training set and from 0 to 6.3 in the test set.

It has been shown that lipophilicity, molecular size, aqueous solubility, polarity, and flexibility of chemicals are

* Author phone: 358 09 19159170; fax: 358 09 19159566; e-mail: jarmo.huuskonen@helsinki.fi.

10.1021/ci020342j CCC: $25.00
Published on Web 07/04/2003

important factors in the soil sorption process.[1,2] Hence, the following physicochemical properties were calculated and investigated as structural parameters in the modeling. Chemical structures for each compound were searched from ChemFinder and saved as SDF format. The octanol/water partition coefficients, log *P*, were calculated by using the KOWWIN program (v. 1.62, Syracuse Research Corporation, Syracuse, NY) from SDF batch mode. SDF formats were converted to SMILES notation, and aqueous solubility values were calculated by using a nonlinear method for calculation of log *S* proposed by Huuskonen.[16] The polarity of chemical was expressed as the number of OH and NH groups (HBD) and the number of O and N atoms (HBA) in a molecule, and flexibility was expressed as the number of rotational bonds (ROT) in a molecule. The effect of possible ionization was expressed as an indicator value for carboxylic acids (I_{acid}), and was given the value of one for acids and zero for the remaining compounds. The aromaticity was expressed simply as the number of 5- and 6-member aromatic rings (NAR).

Using the set of 403 compounds, multiple linear regression models were developed based on regression algorithms in the SPSS package.[17] The quality of the model was considered as statistically satisfactory on the basis of squared correlation coefficient (r^2), standard deviation (s), and F-statistics (F) when all the parameters in the model were significant at 95% confidence level. The predictive ability inside the training set was carried out using leave-one-out cross-validation. For a reliable model, the squared predictive correlation coefficient (q^2) should be greater than 0.6 and the standard deviation in cross-validation, s_{cv}, should not be much higher than that in regression.[18,19] The actual prediction ability was evaluated using a test set of 165 compounds not included in the training set, and predictive r^2 and s are considered.

RESULTS AND DISCUSSION

The following regression equations with one or five structural parameters gave reasonable statistics

$$\log K_{oc} = -0.54 \log S + 0.81 \quad (1)$$

($n = 403$, $r^2 = 0.80$, $s = 0.51$, $F = 1622.3$, $q^2 = 0.80$, $s_{press} = 0.52$)

$$\log K_{oc} = -0.35 \log S - 0.18\mathrm{HBA} + 0.24\mathrm{NAR} + 0.004\mathrm{MW} - 0.55 I_{acid} + 0.88 \quad (2)$$

($n = 403$, $r^2 = 0.85$, $s = 0.44$, $F = 450.5$, $q^2 = 0.84$, s_{press} 0.46)

$$\log K_{oc} = 0.60 \log P + 0.84 \quad (3)$$

($n = 403$, $r^2 = 0.79$, $s = 0.52$, $F = 1474.7$, $q^2 = 0.78$, $s_{press} = 0.53$)

$$\log K_{oc} = 0.48 \log P + 0.26\mathrm{NAR} - 0.07\mathrm{ROT} + 0.002\mathrm{MW} - 0.77 I_{acid} + 0.56 \quad (4)$$

($n = 403$, $r^2 = 0.86$, $s = 0.43$, $F = 491.3$, $q^2 = 0.85$, s_{press} 0.45)

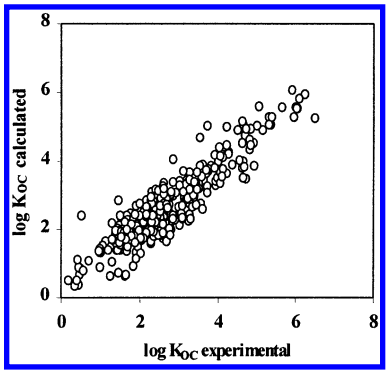

Figure 1. Correlation of experimental log K_{oc} and calculated log K_{oc} values by Equation 2 in the training set.

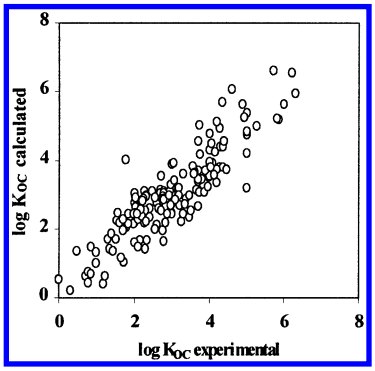

Figure 2. Correlation of experimental log K_{oc} and predicted log K_{oc} values by Equation 2 in the test set.

Parameters in eqs 2 and 4 are arranged in order of their relative importance. Correlation for experimental log K_{oc} and calculated log K_{oc} values by eq 2 is given in Figure 1, and Figure 2 gives the correlation of experimental log K_{oc} and predicted log K_{oc} values by eq 2 in the test set. An analysis of possible outliers showed that only for two compounds (2,6-chlorobenzamide and pentafluorophenyl methyl sulfone) were the residuals $> 3 \times s$ in the training set. The results for the prediction by eq 2 in the test set of 165 compounds not included in the training set were $r^2 = 0.79$ and $s = 0.58$. Hence, these results are in a good agreement with the results obtained for the training set. The correlation matrix for structural parameters in the equations above is given in Table 1. Table 2 gives the comparison of eqs 1−4, and experimental and predicted soil sorption coefficients in the test set are given in Table 3. All the 568 compounds along with their experimental soil sorption coefficients are available as Supporting Information. It should be noted that there are three compounds (fomesafen, thiophanate methyl, and esfenvalerate) in the test set which are largely overestimated by eqs 2 and 4. These compounds are carboxylic or sulfonic acid esters, and hydrolysis in an aqueous environment is possible, which would decrease the soil sorption coefficients of these compounds. When these three compounds were excluded, statistics for the remaining 162 compounds in the test set were $r^2 = 0.91$ and $s = 0.56$ by eq 2 and $r^2 = 0.90$ and $s = 0.55$ by eq 4.

As can be noted from the regression equations above, using only the calculated octanol/water partition coefficient and the aqueous solubility gave a reasonable prediction of log K_{oc}. Addition of simple structural parameters improved the prediction ability in both cases. This might be explained in two ways. First, estimation errors exist in the calculation of

Table 1. Correlation Matrix for Parameters in Regression Equations

	log K_{OC}	mlog P	log P	ROT	MW	I_{acid}	NAR	log S	HBA
log K_{OC}	1.00								
mlog P^a	0.90	1.00							
log P	0.89	0.98	1.00						
ROT	−0.09	0.06	0.09	1.00					
MW	0.61	0.68	0.67	0.34	1.00				
I_{acid}	−0.16	−0.05	−0.04	−0.01	−0.05	1.00			
NAR	0.54	0.39	0.39	−0.31	0.10	−0.04	1.00		
log S	−0.90	−0.89	−0.90	−0.00	−0.74	0.09	−0.49	1.00	
HBA	−0.37	−0.37	−0.36	0.51	0.23	0.02	−0.20	0.25	1.00

[a] Experimental octanol/water coefficient.

Table 2. Comparison of Prediction Ability of Regression Equations for the Estimation of Soil Sorption Coefficient

	training set					test set		
model	r^2	s	q^2	s_{cv}	n	r_{pred}^2	s	n
eq 1	0.802	0.515	0.800	0.518	403	0.761	0.610	165
eq 2	0.850	0.448	0.845	0.455	403	0.790	0.582	165
eq 3	0.786	0.528	0.782	0.532	403	0.726	0.679	165
eq 4	0.861	0.434	0.854	0.440	403	0.802	0.592	165

Table 3. Experimental and Predicted Soil Sorption Coefficients in the Test Set

no.	CAS no.	compound	log K_{oc}	eq 1	eq 2	eq 3	eq 4
1	86-86-2	1-naphthaleneacetamide	2.00	2.53	2.75	1.88	2.23
2	32357-46-3	2,4-DB butoxyethyl ester	2.70	3.20	3.14	3.90	3.37
3	33089-61-1	amitraz	3.00	4.03	3.89	4.18	4.21
4	12771-68-5	ancymidol	2.08	2.62	2.69	2.04	2.39
5	1861-40-1	benfluralin	4.26	4.02	3.79	4.04	3.86
6	741-58-2	bensulide	3.00	3.18	3.11	3.32	3.23
7	1689-99-2	bromoxynil octyl ester	4.00	3.85	3.92	4.37	4.09
8	510-15-6	chlorobenzilate	3.30	3.40	3.61	3.24	3.46
9	76-06-2	chloropicrin	1.79	2.06	2.07	1.63	1.53
10	52315-07-8	cypermethrin	5.00	5.00	4.75	4.68	4.70
11	96-12-8	DBCP	1.85	1.94	2.43	2.46	2.30
12	1861-32-1	chlorthal-dimethyl	3.70	3.69	3.39	3.39	3.41
13	120-36-5	2,4-DP butoxyethyl ester	3.00	3.29	3.20	3.39	3.03
14	115-32-2	dicofol	3.70	3.99	4.56	4.30	4.59
15	55283-68-6	ethalfluralin	3.60	3.94	3.73	3.99	3.89
16	2593-15-9	etridiazole	3.00	2.39	2.45	3.01	3.03
17	66441-23-4	fenoxaprop-ethyl	3.98	3.89	3.51	3.55	3.66
18	55-38-9	fenthion	3.18	3.39	3.19	3.30	3.13
19	62924-70-3	flumetralin	4.00	4.72	4.78	4.51	4.78
20	72178-02-0	fomesafen	1.78	4.26	4.00	2.89	3.39
21	78587-05-0	hexythiazox	3.79	3.66	3.44	4.20	4.10
22	81334-34-1	imazapyr acid	2.00	2.75	2.18	1.78	2.08
23	33820-53-0	isopropalin	4.00	3.49	3.55	4.33	4.04
24	58-89-9	lindane	3.04	3.99	3.92	3.40	3.34
25	7085-19-0	mecoprop	1.30	2.26	1.70	2.61	1.78
26	26087-47-8	iprofenfos	2.40	2.29	2.35	2.99	2.97
27	20354-26-1	methazole	3.48	2.71	2.34	2.78	2.95
28	74223-64-6	metsulfuron methyl	1.54	2.96	2.25	2.05	2.42
29	2122-70-5	ethyl 1-naphthyl acetate	2.48	2.93	3.10	3.10	3.14
30	300-76-5	naled	2.26	1.92	2.19	1.81	1.92
31	5259-88-1	oxycarboxin	1.98	2.39	2.16	1.69	1.95
32	42874-03-3	oxyfluorfen	5.00	4.17	4.22	3.98	4.21
33	732-11-6	phosmet	2.91	3.34	2.93	2.34	2.45
34	3478-94-2	piperalin	3.70	3.11	3.44	3.90	4.00
35	67747-09-5	prochloraz	2.70	3.07	3.04	3.33	3.51
36	709-98-8	propanil	2.17	2.84	2.82	2.57	2.54
37	74051-80-2	sethoxydim	2.00	3.54	3.03	3.24	2.67
38	35400-43-2	sulprofos	4.08	4.00	3.91	4.24	3.79
39	3383-96-8	temephos	5.00	3.58	3.18	4.55	4.13
40	23564-05-8	thiophanate methyl	3.25	2.91	2.20	1.50	1.63
41	66841-25-6	tralomethrin	5.00	4.61	5.39	5.39	5.88
42	1582-09-8	trifluralin	4.37	4.03	3.80	4.04	3.86
43	2686-99-9	trimethacarb	2.60	2.14	2.10	2.53	2.52
44	1918-13-4	chlorthiamide	2.26	3.09	3.11	2.62	2.69
45	93-72-1	fenoprop	1.75	2.73	2.19	3.06	2.28
46	957-51-7	diphenamid	2.32	2.57	2.96	2.56	2.77

Table 3 (Continued)

no.	CAS no.	compound	log K_{oc}	eq 1	eq 2	eq 3	eq 4
47	70124-77-5	fluthrinate	5.00	4.76	4.74	4.74	5.05
48	6923-22-4	monocrotophos	0.00	0.75	0.54	0.05	0.06
49	13684-63-4	phenmedipharm	3.38	3.24	2.87	2.81	2.91
50	76578-14-8	quizalofop ethyl	2.71	3.91	3.55	3.46	3.61
51	78-48-8	tribufos	3.70	3.83	3.71	4.30	3.42
52	2675-77-6	chloroneb	3.22	2.71	2.69	2.91	2.86
53	79127-80-3	fenoxycarb	3.00	3.42	2.93	3.39	3.05
54	13171-21-6	phosphamidon	0.85	1.53	1.48	1.07	0.92
55	542-75-6	1,3-dichloropropene	1.51	1.52	1.72	2.22	1.87
56	93-76-5	2,3,5-T acid	1.72	2.51	2.00	2.80	2.04
57	30560-19-1	acephtate	0.30	0.78	0.23	0.30	−0.48
58	101-05-3	anilazine	3.00	3.08	3.05	3.03	3.38
59	22781-23-3	bendiocarb	2.75	2.35	1.97	2.38	2.39
60	42576-02-3	bifenox	4.00	4.16	3.96	3.34	3.65
61	5234-68-4	carboxin	2.41	2.68	2.59	1.74	1.91
62	90982-32-4	chlorimuron-ethyl	2.04	3.10	2.64	2.22	2.64
63	68359-37-5	cyfluthrin	5.00	5.01	4.83	4.80	4.84
64	66215-27-8	cyromazine	2.30	2.07	1.42	1.42	1.55
65	99-30-9	dicloran	3.00	2.76	2.73	2.50	2.66
66	13684-56-5	desmedipham	3.18	3.39	2.96	2.78	2.81
67	1194-65-6	diclobenil	2.60	2.73	2.76	2.54	2.54
68	51338-27-3	diclofop-methyl	4.20	3.80	3.73	3.57	3.69
69	66230-04-4	esfenvalerate	3.72	5.04	5.04	4.91	5.15
70	13194-48-4	ethoprop	1.85	2.03	2.14	2.73	2.12
71	82-68-8	quintozene	4.30	3.84	4.39	3.87	4.50
72	60168-88-9	fenarimol	2.78	2.97	3.11	3.02	3.17
73	69409-94-5	fluvalinate	6.00	5.81	5.63	4.94	5.31
74	944-22-9	fonofos	2.94	3.22	3.09	3.26	2.76
75	51235-04-2	hexazinone	1.73	1.39	1.04	2.14	2.08
76	67485-29-4	hydramethylnon	5.86	5.26	5.20	5.97	6.06
77	36734-19-7	iprodione	2.85	3.48	2.89	2.56	2.73
78	25311-71-1	isofenphos	2.78	3.35	3.04	3.69	3.33
79	91465-08-6	cyhalothrin	5.26	5.21	5.01	4.96	5.01
80	10265-92-6	methamidophos	0.70	0.50	0.64	0.28	0.32
81	74-83-9	methyl bromide	1.34	1.12	1.42	1.55	1.37
82	301-12-2	oxydemeton-methyl	1.00	1.29	1.33	0.22	0.18
83	2439-01-2.	oxythioquinox	3.36	3.29	2.72	2.87	2.77
84	40487-42-1	pendimethalin	3.70	3.30	3.14	3.74	3.57
85	29232-93-7	pirimiphos-methyl	3.00	2.92	2.44	2.91	2.75
86	2312-35-8	propargite	3.60	4.06	3.68	4.19	3.96
87	51707-55-2	thidiazuron	2.04	2.74	2.45	2.10	2.35
88	59669-26-0	thiodicarb	2.54	2.88	1.98	1.99	1.66
89	137-26-8	thiram	2.83	2.86	2.66	1.87	1.63
90	58138-08-2	tridiphane	3.75	4.11	4.16	3.96	3.90
91	26644-46-2	triforine	2.30	2.73	2.53	2.05	2.18
92	55290-64-7	dimethipin	0.48	1.56	1.37	0.87	1.10
93	128-66-5	dibenzo[*b*,*d*]chrysene-7,12-dione	4.28	4.01	4.94	4.62	5.71
94	366-18-7	2,2-dipyridine	1.60	1.81	2.19	1.67	2.06
95	119-91-5	2,2-biquinoline	4.02	3.81	4.30	3.28	4.12
96	101-77-9	4,4′-methylenedianiline	1.99	1.98	2.45	2.15	2.48
97	6422-86-2	bis(2-ethylhexyl)terephthalate	4.16	4.14	4.24	5.89	4.86
98	78-59-1	isophorone	1.40	1.91	1.88	2.42	2.17
99	314-42-1	isocil	2.11	1.92	2.27	1.56	1.40
100	91-17-8	decahydronaphthalene	3.67	3.61	3.13	3.37	2.94
101	117-84-0	di-*n*-octylphthalete	4.38	4.08	4.39	5.98	4.66
102	84-65-1	9,10-anthracenedione	3.57	3.44	3.83	2.85	3.46
103	1610-17-9	atratone	2.64	2.24	2.60	2.74	2.51
104	225-11-6	benzo[*c*]acridine	4.39	4.09	4.57	3.54	4.33
105	1468-95-7	9-anthracenemethanol	3.61	3.12	3.63	2.91	3.44
106	86-87-3	1-naphthaneneacetic acid	2.20	2.50	2.36	2.41	1.95
107	4780-79-4	1-naphthalenemethanol	2.17	2.09	2.56	2.20	2.49
108	134-32-7	1-naphthylamine	3.51	2.13	2.53	2.20	2.52
109	90-15-3	1-naphthol	2.72	2.10	2.52	2.46	2.73
110	82-28-0	1-amino-2-methyl-9,10-anthracenedione	3.90	3.63	3.69	3.29	3.89
111	193-39-5	indeno(1,2,3-CD)pyrene	6.20	5.92	6.56	4.87	6.03
112	189-55-9	dibenzo[*a*,*i*]pyrene	5.71	5.85	6.61	5.22	6.38
113	53-70-3	dibenz[*a*,*h*]anthracene	6.31	5.31	5.93	4.87	5.78
114	85-44-9	1,3-isobenzofurandione	1.56	2.08	2.45	2.09	2.19
115	191-24-2	benzo[*ghi*]perylene	4.61	5.94	6.08	4.87	5.52
116	207-08-9	benzo[*k*]fluoranthene	4.34	5.44	5.68	4.52	5.17
117	35554-44-0	Imazalil	3.73	2.68	3.06	3.31	3.37
118	3766-81-2	fenobucarb	1.71	2.35	2.29	2.56	2.36
119	2425-10-7	xylicarb	1.71	2.00	1.96	2.21	2.15

Table 3 (Continued)

no.	CAS no.	compound	log K_{oc}	eq 1	eq 2	eq 3	eq 4
120	5915-41-3	terbuthylazine	2.32	2.68	2.20	2.81	2.62
121	613-13-8	2-aminoanthracene	4.45	3.37	3.73	2.91	3.47
122	115-28-6	chlorendic acid	2.79	1.85	1.63	2.73	2.19
123	118-75-2	chloranil	2.32	3.28	2.94	2.18	2.24
124	91-80-5	methapyrilene	2.87	1.91	2.44	2.38	2.74
125	1698-60-8	chloridazon	2.08	2.49	2.43	1.30	1.66
126	12427-38-2	maneb	2.74	3.37	2.88	1.21	1.10
127	149-30-4	2(3*H*)benzothiazoledione	2.25	2.56	2.89	2.56	2.80
128	92-24-0	tetracene	5.81	4.85	5.22	4.16	4.83
129	143-50-0	chlordecone	4.20	4.79	5.12	3.80	4.14
130	56-04-2	metacil	2.14	1.77	1.69	1.71	1.80
131	60-11-7	4-(dimethylamino)azobenzene	3.87	3.11	3.07	3.42	3.57
132	1615-80-1	*N*,*N*-diethylhydrazine	1.18	0.42	0.40	1.11	0.86
133	15299-99-7	*N*,*N*-diethyl-2-(1-naphthalenyloxy)propanamide	2.76	2.75	2.94	2.85	2.75
134	103-23-1	di(2-ethylhexyl)adipate	4.19	3.82	3.36	5.73	4.14
135	56-53-1	diethylstibestrol	4.14	3.39	3.58	4.24	4.19
136	86-30-6	*N*-nitrosodiphenylamine	3.08	2.92	2.86	2.74	2.95
137	81-81-2	warfarin	2.96	3.26	3.28	2.18	2.63
138	2307-68-8	pentanochlor	2.76	2.81	2.87	3.36	3.01
139	961-11-5	stirofos	3.07	3.27	3.42	3.13	3.21
140	673-04-1	simetone	2.34	2.04	1.69	2.48	2.28
141	54-11-5	nicotine	2.01	1.23	1.60	1.44	1.63
142	151-56-4	aziridine	0.78	0.18	0.45	0.67	0.53
143	1861-40-1	benefin	4.03	4.02	3.79	4.04	3.86
144	123-91-1	1,4-dioxane	1.23	0.51	0.64	0.65	0.62
145	94-59-7	safrole	2.83	2.16	2.19	2.92	2.68
146	100-54-9	3-cyanopyridine	1.65	1.18	1.18	1.05	1.18
147	492-80-8	auramine	3.31	1.84	2.42	2.64	3.04
148	90-94-8	4,4′-bis(dimethylamino)benzophenone	2.21	2.47	2.82	3.17	3.47
149	60-09-3	*p*-phenylazoaniline	2.79	2.89	2.84	2.76	3.04
150	842-07-9	1-(phenylazo)-2-naphthalenol	3.58	3.46	3.62	4.16	4.47
151	7159-34-4	pyroxychlor	3.48	2.88	2.99	3.18	3.27
152	103-74-2	2-pyridineethanol	1.45	1.08	1.37	1.07	1.17
153	85-70-1	2-butoxy-2-oxoethyl butyl phthalate	3.70	3.07	2.65	3.34	2.82
154	36355-01-8	hexabromobiphenyl	4.87	4.92	5.63	6.32	6.88
155	145-73-3	endothal	2.09	1.41	1.50	1.98	2.24
156	106-89-8	epichlorohydrin	1.00	0.80	1.02	1.22	0.95
157	3597-91-9	4-biphenylmethanol	2.54	2.50	2.91	2.55	2.70
158	101-61-1	4,4′-methylenebis(*N*,*N*-dimethylaniline)	3.96	2.89	3.22	3.47	3.68
159	85-22-3	pentabromoethylbenzene	4.92	4.52	5.24	5.34	5.60
160		4-nonylphenyl diphenyl phosphate	4.06	3.99	4.49	6.37	5.77
161	62-55-5	thioacetamide	0.78	0.49	0.76	0.34	0.21
162	139-65-1	4,4′-thiodianiline	2.04	2.60	2.90	2.32	2.66
163	62-56-6	thiourea	0.85	0.68	0.71	0.05	−0.02
164	122-66-7	1,2-diphenylhydrazine	2.98	2.44	2.69	2.68	2.80
165	93-08-3	1-(naphthalenyl)ethanone	2.93	2.69	2.98	2.56	2.81

log *P* and log *S*. Experimental octanol/water partition coefficients were available for 308 compounds in the training set and when experimental log *P* was used in eq 4 the statistics were $r^2 = 0.88$ and $s = 0.41$. Hence this gave only a marginal improvement in the estimation.

Experimental errors exist also in soil sorption coefficients as pointed out by Lohninger,[8] who suggested that the uncertainty in experimental soil sorption values is ≈0.4. Although octanol/water partition coefficient and aqueous solubility gave a reasonable prediction of soil sorption coefficient, parameters for molecular size, flexibility, ionization, and aromaticity of analyzed compounds were also considered. As the results of the present study show, these easily calculated parameters could be used to improve the prediction ability of the models in both cases. Increasing molecular size and an increase in the number of aromatic rings decrease the solubility of a compound in water and increase sorption to organic soil content. Possible ionization of carboxylic acids increases their solubility and hence favors the water phase. Molecular flexibility decreases sorption to organic soil content. This might be due to association of high entropy of fusion, which makes sorption of flexible compounds difficult.

The soil organic matrix can be seen as an organized and rigid macromolecule structure and its ability to strongly interact with organic chemicals depends primarly on the chemicals' ability to adopt conformation which can interact with the organic soil content. The chemicals with high log *P* values, i.e., highly planar polyaromatic hydrocarbons (underestimated), and large and highly flexible chemicals (overestimated), had large differences between the log *P* and log K_{oc} correlations. This supports the fact that planarity, expressed as the number of aromatic rings (NAR), and flexibility, expressed as the number of rotable bonds (ROT), of a chemical are important factors in determining its ability to bond with soil organic content.

In the present study two simple models for estimation of soil sorption coefficient of a large and diverse set of chemicals is proposed. The results show that calculated aqueous solubility and octanol/water partition coefficient gave comparable predictions. However, uncertainty in estimation of aqueous solubility is at present higher than the

uncertainty of available methods for estimation of octanol/water partition coefficient. Hence, there is a need to develop more accurate methods for estimation of aqueous solubility from molecular structure, which would also improve estimation of soil sorption coefficients.

ACKNOWLEDGMENT

I thank William Meylan from SRC for giving us the KOWWIN program to use.

Supporting Information Available: Table giving a complete listing of experimental soil sorption coefficients for all 586 chemicals (18 pages; PDF). This material is available free of charge via the Internet at http://pubs.asc.org.

REFERENCES AND NOTES

(1) Gawlik, B. M.; Sotiriou, N.; Feicht, E. A.; Schulte-Hostede, S.; Kettrup, A. Alternatives for the Determination of the Soil Absorption Coefficient K_{oc} of Non-ionic Organic Compounds − Review. *Chemosphere* **1997**, *34*, 2525−2551.

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

(3) Sabljic, A.; Güsten, H.; Verhaar, H.; Hermens, J. QSAR Modelling of Soil Sorption. Improvements and Systematics of log K_{oc} vs log K_{ow} Correlations. *Chemosphere* **1995**, *31*, 4489−4514.

(4) Baker, J. B.; Mihelcic, J. R.; Shea, E. Estimating K_{oc} for Persistent Organic Pollutants: Limitations of Correlations with K_{ow}. *Chemosphere* **2000**, *41*, 813−817.

(5) Goss, K.-U.; Schwarzenbach, R. P. Linear Free Energy Relationships Used to Evaluate Equilibrium Partitioning of Organic Compounds. *Environ. Sci. Technol.* **2001**, *35*, 1−9.

(6) Sabljic, A. On the Prediction of Soil Sorption Coefficients of Organic Pollutants from Molecular Structure: Application of Molecular Topology Model. *Environ. Sci. Technol.* **1987**, *21*, 358−366.

(7) Meylan, W. M.; Howard, P. H. Molecular Topology/Fragment Contribution Method for Predicting Soil Sorption Coefficients. *Environ. Sci. Technol.* **1992**, *26*, 1560−1567.

(8) Lohninger, H. Estimation of Soil Partition Coefficients of Pesticides from Their Structure. *Chemosphere* **1994**, *29*, 1611−1626.

(9) Tao, S.; Lu, X. Estimation of Organic Carbon Normalized Sorption Coefficient (K_{oc}) for Soils by Topological Indices and Polarity Factors. Chemosphere **1999**, *39*, 2019−2034.

(10) Tao, S.; Piao, H.; Dawson, R.; Lu, X.; Hu, H. Estimation of Organic Carbon Normalized Sorption Coefficient (K_{oc}) for Soils Using the Fragment Constant Methodology. *Environ. Sci. Technol.* **1999**, *33*, 2719−2725.

(11) Müller, M. Quantum Chemical Modelling of Soil Sorption Coefficients: Multiple Linear Regression Models. *Chemosphere* **1997**, *35*, 365−377.

(12) Gramatica, P.; Corradi, M.; Consonni, V. Modelling and Prediction of Soil Coefficients of Nonionic Organic Pesticides by Molecular Descriptors. *Chemosphere* **2000**, *41*, 763−777.

(13) Winget, P.; Cramer, C. J.; Truhlar, D. G. Prediction of Soil Sorption Coefficients Using a Universal Solvation Model. *Environ. Sci. Technol.* **2000**, *34*, 4733−4740.

(14) 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 Physicochemical Descriptors. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1450−1459.

(15) Syracuse Research Corporation. *Chemfate Research Results*. SCR Environmental Science Center: Syracuse, NY, 2002 (http://esc.syr-res.com/efdb/Chemfate.htm).

(16) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.

(17) SPSS. *SPSS for Windows Release*, v. 10.0; Chicago, IL.

(18) Wold, S. Validation of QSARs. *Quantum Struct.-Act. Relat.* **1991**, *10*, 191−193.

(19) Kubinyi, H. Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. *Quantum Struct.-Act. Relat.* **1994**, *13*, 285−294.

CI020342J