Prediction of partitioning properties for environmental pollutants using mathematical structural descriptors

Subhash C. Basak* and Denise Mills

Natural Resources Research Institute, University of Minnesota Duluth,
5013 Miller Trunk Highway, Duluth, MN 55811, USA
E-mail: sbasak@nrri.umn.edu

Dedicated to Dr. Nitya Anand on the occasion of his 80th birthday

(received 08 Sep 04; accepted 14 Feb 05; published on the web 02 Mar 05)

Abstract

Predictive models, based solely on molecular structure, were developed for three environmentally-related partitioning properties: Water solubility, soil/sediment partition coefficient, and octanol/water partition coefficient. Data for a diverse set of 136 chemicals were taken from the literature, and include aromatic and aliphatic compounds, as well as herbicides, pesticides, and polycyclic aromatic hydrocarbons. The hierarchical QSAR (HiQSAR) approach to model building was employed, in which increasingly more computer-resource intensive classes of structural descriptors are used only when the simpler and more easily calculable descriptors do not provide adequate models. The results indicate that the simple topostructural (TS) and topochemical (TC) descriptors provide the best models, and that, in many cases, these structure-based models are superior to those based on properties.

Keywords: Ridge regression, hierarchical QSAR, octanol/water partition coefficient, soil/ sediment partition coefficient, water solubility, environmental pollutants

Introduction

Modern lifestyle in the industrialized world is dependent upon the use of thousands of chemicals for various industrial processes as well as for special purposes as drugs, pesticides, herbicides, etc. In the United States, the Toxic Substances Control Act (TSCA) Inventory currently has over 75,000 chemicals of which over 2,800 are high production volume chemicals (HPVs).[1] These chemicals may be released to the environment during their production, transport, and intended uses. Pollutants can also be released into the environment from underground storage tanks, hazardous waste disposal sites, municipal landfills, and accidental spills. The Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) priority list contains 275 chemicals, many of which are found at facilities on the National Priority List (NPL).[2] These chemicals pose a substantial threat to human, wildlife, and ecological health.

Understanding the distribution of pollutants among different environmental phases is crucial to their hazard assessment and remediation of contaminated sites. Many contaminants end up in the soil and sediment. Remediation often involves extraction of the polluting chemicals into the aqueous phase and then treatment by physical, chemical or biological processes. The extraction methodology is critically dependent on the partitioning properties of the chemicals.[3]

Physicochemical properties such as octanol/water partition coefficients (K_{ow}) and aqueous solubility (S) have been used in the estimation of partitioning of chemicals among various environmental phases.[3,4] Reversed-phase high performance liquid chromatography has also been used in the prediction of S and K_{ow}. Whereas such property-property correlations designed to estimate properties of environmental interest from other known physicochemical properties work reasonably well, this approach is limited by the unavailability of the latter properties for the majority of chemicals of environmental concern.[5] Various studies have shown that properties such as S[6] and K_{ow}[7,8] can be predicted using mathematical molecular descriptors, which can be calculated directly from chemical structures alone without the input of any other experimental data. Such molecular descriptors quantify aspects of structure, which contribute to interactions of chemicals with hydrophobic and hydrophilic phases. Topostructural, topochemical, geometrical (3D), and quantum chemical indices comprise a set of descriptors which have proved useful in the prediction of toxicity and toxicologically relevant properties of both congeneric and structurally diverse sets of molecules.[9–11] Predictive models based on calculated descriptors can provide cost effective and rapid estimates of partitioning behavior of environmental pollutants. They can also provide insight into the environmental behavior of chemicals not yet synthesized or those that cannot be examined experimentally due to their extremely hazardous nature.

Chu and Chan used K_{ow} to predict S and soil/sediment partition coefficients (K_{oc}) of a diverse collection of pollutants, viz., aliphatics, aromatics, pesticides, herbicides, and polycyclic aromatic hydrocarbons (PAHs).[3] They also developed predictive models for K_{oc} based on solubility. As stated earlier, such property-property correlation methods are of limited applicability. Therefore, we were interested to investigate whether properties of environmental interest can be estimated from molecular structural descriptors. We have formulated a hierarchical quantitative structure-activity relationship (HiQSAR) approach where calculated descriptors are used in a graduated manner such that computationally more resource intensive parameters are used only when easily calculable indices do not provide acceptable results. We have carried out a comparative study of physicochemical properties vis-à-vis theoretically based HiQSAR approach in the estimation of partitioning of chemicals of environmental concern.

Methods and Materials

Experimental data

Chu and Chan[3] collected data for water solubility (*S*), octanol/water partition coefficient (K_{ow}), and soil/sediment partition coefficient (K_{oc}) from several sources including an EPA report on ground water remediation and the Handbook of Environmental Data on Organic Chemicals.[12,13] While Chu and Chan selected 148 compounds, we omitted 12 from their collection, resulting in a total number of 136 chemicals in our data set. The omitted compounds include: a) isomers that are indistinguishable with respect to our software (1,2-Dichloroethene, Hexachlorocyclohexane), b) compounds with fewer than three non-hydrogen atoms, for which our complete set of descriptors cannot be calculated (chloromethane, iodomethane), c) mixtures (chlorotoluene, cresol, xylene), and d) those compounds that contain atoms not represented in our software (cacodylic acid). Based on Chu and Chan's classification scheme, the 136 chemicals were partitioned into five categories: aliphatics (26), aromatics (43), herbicides (18), polycyclic aromatic hydrocarbons (19), and pesticides (30). The data are provided in Table 1.

Table 1. Solubility (*S*), soil/sediment partition coefficient (K_{oc}), octanol/water partition coefficient (K_{ow})

	Aliphatics	*S*	K_{oc}	K_{ow}
1	1,1,1-Trichloroethane (Methyl chloroform)	1.12 exp-02	1.52 exp+02	3.16 exp+02
2	1,1,1,2-Tetrachloroethane	1.73 exp-02	1.18 exp+02	2.45 exp+02
3	1,1,2-Trichloroethane	3.37 exp-02	5.60 exp+01	2.95 exp+02
4	1,1-Dichloroethane	5.56 exp-02	3.00 exp+01	6.17 exp+01
5	1,1-Dichloroethene	2.32 exp-02	6.50 exp+01	6.92 exp+01
6	1,2-Dichloroethane	8.61 exp-02	1.40 exp+01	3.02 exp+01
7	1,3-butadiene	1.36 exp-02	1.20 exp+02	9.77 exp+01
8	Acetonitrile [methyl cyanide]	Infinity	2.20 exp+00	4.57 exp-01
9	Acrylonitrile [2-propenenitrile]	1.50 exp+00	8.50 exp-01	1.78 exp+00
10	Bromodichloromethane	2.69 exp-02	6.10 exp+01	7.59 exp+01
11	Chloroethane	8.90 exp-02	1.70 exp+01	3.50 exp+01
12	Chloroethene	4.27 exp-02	5.70 exp+01	2.40 exp+01
13	Dibromochloromethane	1.92 exp-02	8.40 exp+01	1.23 exp+02
14	Dichlorodifluorormethane [Freon12]	2.32 exp-03	5.80 exp+01	1.45 exp+02
15	Dichloromethane	2.35 exp-01	8.80 exp+00	2.00 exp+01
16	Ethylene dibromide	2.29 exp-02	4.40 exp+01	5.75 exp+01
17	Hexachlorobutadiene	5.74 exp-07	2.90 exp+04	6.02 exp+04
18	Hexachloropentadiene	7.69 exp-06	4.80 exp+03	1.10 exp+05
19	Hexachloroethane (perchloroethane)	2.11 exp-04	2.00 exp+04	3.98 exp+04
20	Pentachloroethane	1.83 exp-04	1.90 exp+03	7.76 exp+02

Table 1. Contuined

21	Tetrachloroethene	8.97 exp-04	3.64 exp+02	3.98 exp+02
22	Tetrachloromethane [Carbon tetrachloride]	4.92 exp-03	4.39 exp+02	4.37 exp+02
23	Tribromomethane [Bromoform]	1.19 exp-02	1.16 exp+02	2.51 exp+02
24	Trichloroethene	8.29 exp-03	1.26 exp+02	2.40 exp+02
25	Trichlorofluoromethane [Freon 11]	8.01 exp-03	1.59 exp+02	3.39 exp+02
26	Trichloromethane [Chloroform]	6.87 exp-02	4.70 exp+01	9.33 exp+01
	Aromatics			
27	1,2,3,4-Tetrachlorobenzene	1.62 exp-05	1.80 exp+04	2.88 exp+04
28	1,2,3,5-Tetrachlorobenzene	1.11 exp-05	1.78 exp+04	2.88 exp+04
29	1,2,3-Trichlorobenzene	6.61 exp-05	7.40 exp+03	1.29 exp+04
30	1,2,4,5-Tetrachlorobenzene	2.78 exp-05	1.60 exp+03	4.68 exp+04
31	1,2,4-Trichlorobenzene	1.65 exp-04	9.20 exp+03	2.00 exp+04
32	1,2-Dichlorobenzene [*o*-dichlorobenzene]	6.80 exp-04	1.70 exp+03	3.98 exp+03
33	1,3,5-Trichlorobenzene	3.20 exp-05	6.20 exp+03	1.41 exp+04
34	1,3-Dichlorobenzene [*m*-dichlorobenzene]	8.37 exp-04	1.70 exp+03	3.98 exp+03
35	1,3-Dinitrobenzene	2.80 exp-03	1.50 exp+02	4.17 exp+01
36	1,4-Dichlorobenzene [*p*-dichlorobenzene]	5.37 exp-04	1.70 exp+03	3.98 exp+03
37	2,3,4,6-Tetrachlorophenol	3.02 exp-05	9.80 exp+01	1.26 exp+04
38	2,3-Dinitrotoluene	1.70 exp-02	5.30 exp+01	1.95 exp+02
39	2,4,5-Trichlorophenol	6.03 exp-03	8.90 exp+01	5.25 exp+03
40	2,4,6-Trichlorophenol	4.05 exp-03	2.00 exp+03	7.41 exp+03
41	2,4-Dichlorophenol	2.82 exp-02	3.80 exp+02	7.94 exp+02
42	2,4-Dimethylphenol [*as*-*m*-xylenol]	3.44 exp-02	2.22 exp+02	2.63 exp+02
43	2,4-Dinitrophenol	3.04 exp-02	1.66 exp+01	3.16 exp+01
44	2,4-Dinitrotoluene	1.32 exp-03	4.50 exp+01	1.00 exp+02
45	2,5-Dinitrotoluene	7.25 exp-03	8.40 exp+01	1.90 exp+02
46	2,6-Dinitrotoluene	7.25 exp-03	9.20 exp+01	1.00 exp+02
47	2-Chlorophenol [*o*-chlorophenol]	2.26 exp-01	4.00 exp+02	1.45 exp+02
48	3,4-Dinitrotoluene	5.93 exp-03	9.40 exp+01	1.95 exp+02
49	4,6-Dinitro-*o*-cresol	1.46 exp-03	2.40 exp+02	5.01 exp+02
50	4-Chloro-*m*-cresol [chlorocresol]	2.70 exp-02	4.90 exp+02	9.80 exp+02
51	Benzene	2.24 exp-02	8.30 exp+01	1.32 exp+02
52	Bromobenzene [phenyl bromide]	2.84 exp-03	1.50 exp+02	9.00 exp+02
53	Chlorobenzene	4.14 exp-03	3.30 exp+02	6.92 exp+02
54	Diethylstilbestrol [DES]	3.58 exp-08	2.80 exp+01	2.88 exp+05
55	Ethylbenzene [phenylethane]	2.42 exp-04	1.10 exp+03	1.41 exp+03
56	Hexachlorobenzene [perchlorobenzene]	2.11 exp-08	3.90 exp+03	1.70 exp+05
57	Hexachlorophene [dermadex]	9.83 exp-09	9.10 exp+04	3.47 exp+07
58	*m*-Chlorotoluene	3.79 exp-04	1.20 exp+03	1.90 exp+03
59	*m*-Xylene [1,3-dimethylbenzene]	1.22 exp-03	9.82 exp+02	1.82 exp+03

Table 1. Contuined

60	Nitrobenzene	1.54 exp-02	3.60 exp+01	7.08 exp+01
61	*o*-Chlorotoluene	5.69 exp-04	1.60 exp+03	2.60 exp+03
62	*o*-Xylene [1,2-dimethylbenzene]	1.65 exp-03	8.30 exp+02	8.91 exp+02
63	*p*-Chlorotoluene	3.48 exp-04	1.20 exp+03	2.00 exp+03
64	Pentachlorobenzene	5.39 exp-07	1.30 exp+04	1.55 exp+05
65	Pentachloronitrobenzene [quintozene]	2.41 exp-07	1.90 exp+04	2.82 exp+05
66	Pentachlorophenol	5.26 exp-05	5.30 exp+04	1.00 exp+05
67	Phenol	9.88 exp-01	1.42 exp+01	2.88 exp+01
68	*p*-Xylene [1,4-dimethylbenzene]	1.86 exp-03	8.70 exp+02	1.41 exp+03
69	Toluene [methylbenzene]	5.81 exp-03	3.00 exp+02	5.37 exp+02
	Herbicides			
70	2,4,5-Trichlorophenoxyacetic acid	9.32 exp-04	8.01 exp+01	4.00 exp+00
71	2,4-Dichlorophenoxyacetic acid [2,4-D]	2.80 exp-03	1.96 exp+01	6.46 exp+02
72	Alachlor	2.72 exp-03	1.90 exp+02	4.34 exp+02
73	Amitrole [aminotriazole]	3.33 exp+00	4.40 exp+00	8.32 exp-03
74	Atrazine	1.53 exp-04	1.63 exp+02	2.12 exp+02
75	Chloramben	3.40 exp-03	2.10 exp+01	1.30 exp+01
76	Diallate	5.18 exp-05	1.90 exp+03	5.37 exp+00
77	Dichlobenil [2,6-dichlorobenzonitrile]	1.05 exp-04	2.24 exp+02	7.87 exp+02
78	Diuron	1.80 exp-04	3.82 exp+02	6.50 exp+02
79	Fenuron	2.34 exp-02	4.22 exp+01	1.00 exp+01
80	Fluometuron	3.88 exp-04	1.75 exp+02	2.20 exp+01
81	Linuron	3.01 exp-04	8.63 exp+02	1.54 exp+02
82	Monuron	1.16 exp-03	1.83 exp+02	1.33 exp+02
83	Paraquat	2.45 exp+00	1.55 exp+04	1.00 exp+00
84	Picloram	1.78 exp-03	2.55 exp+01	2.00 exp+00
85	Propazine	3.74 exp-05	1.53 exp+02	7.85 exp+02
86	Simazine	1.74 exp-05	1.38 exp+02	8.80 exp+01
87	Trifluralin	1.79 exp-06	1.37 exp+04	2.20 exp+05
	Polycyclic aromatic hydrocarbons			
88	1,2:7,8-Dibenzopyrene [Dibenzo[*a*,*i*]pyrene]	3.34 exp-07	1.20 exp+03	4.17 exp+06
89	1-Naphthylamine	1.64 exp-02	6.10 exp+01	1.17 exp+02
90	2-Methylnaphthalene	1.79 exp-04	8.50 exp+03	1.30 exp+04
91	2-Napthylamine	4.09 exp-03	1.30 exp+02	1.17 exp+02
92	Acenaphthylene	2.58 exp-05	2.50 exp+03	5.01 exp+03
93	Acenapthene	2.22 exp-05	4.60 exp+03	1.00 exp+04
94	Anthracene	2.52 exp-07	1.40 exp+04	2.82 exp+04
95	Benzo[*a*]anthracene	2.50 exp-08	1.38 exp+06	3.98 exp+05
96	Benzo[*a*lpyrene	4.76 exp-09	5.50 exp+06	1.15 exp+06
97	Benzo[*b*]fluoranthene	5.03 exp-08	5.50 exp+05	1.15 exp+06

Table 1. Contuined

98	Benzo[*ghi*]perylene	2.54 exp-09	1.60 exp+06	3.24 exp+06
99	Benzo[*k*]fluoranthene	1.54 exp-08	5.50 exp+05	1.15 exp+06
100	Chrysene	7.89 exp-09	2.00 exp+05	4.07 exp+05
101	Dibenz[*a*,*h*]anthracene	1.80 exp-09	3.30 exp+06	6.31 exp+06
102	Fluoranthene	1.02 exp-06	3.80 exp+04	7.94 exp+04
103	Indeno[1,2,3-*cd*]pyrene	1.92 exp-09	1.60 exp+06	3.16 exp+06
104	Napthalene [napthene]	2.47 exp-04	1.30 exp+03	2.76 exp+03
105	Phenanthrene	5.61 exp-06	1.40 exp+04	2.88 exp+04
106	Pyrene	6.53 exp-07	3.80 exp+04	7.59 exp+04
	Pesticides			
107	1,2-Dibromo-3-chloropropane [DBCP]	4.23 exp-03	9.80 exp+01	1.95 exp+02
108	1,2-Dichloropropane	2.39 exp-02	5.10 exp+01	1.00 exp+02
109	1,3-Dichloropropene [telone]	2.52 exp-02	4.80 exp+01	1.00 exp+02
110	2,3,7,8-Tetrachlorodibenzo-*p*-dioxin	6.21 exp-10	3.30 exp+06	5.25 exp+06
111	Aldrin	4.93 exp-07	9.60 exp+04	2.00 exp+05
112	Captan	1.66 exp-06	6.40 exp+03	2.24 exp+02
113	Carbaryl [sevin]	1.99 exp-04	2.30 exp+02	2.29 exp+02
114	Carbofuran	1.88 exp-03	2.94 exp+01	2.07 exp+02
115	Chlordane	1.37 exp-06	1.40 exp+05	2.09 exp+03
116	Chlorobenzilate	6.73 exp-05	8.00 exp+02	3.24 exp+04
117	Chlorpyrifos [dursban]	8.56 exp-07	1.36 exp+04	6.60 exp+04
118	Cyclophosphamide	5.02 exp+03	4.20 exp-02	6.03 exp-04
119	DDD	3.12 exp-07	7.70 exp+05	1.58 exp+06
120	DDE	1.13 exp-07	4.40 exp+06	1.00 exp+07
121	DDT	1.96 exp-08	2.43 exp+05	1.55 exp+06
122	Diazinon	1.31 exp-04	8.50 exp+01	1.05 exp+03
123	Dieldrin	5.12 exp-07	1.70 exp+03	3.16 exp+03
124	Dinoseb	2.08 exp-04	1.24 exp+02	1.98 exp+02
125	Ethylene oxide	2.27 exp+01	2.20 exp+00	6.03 exp-01
126	Kepone	2.02 exp-08	5.50 exp+04	1.00 exp+02
127	Leptophos	5.82 exp-06	9.30 exp+03	2.02 exp+06
128	Malathion	4.39 exp-04	1.80 exp+03	7.76 exp+02
129	Methoxychlor	8.68 exp-09	8.00 exp+04	4.75 exp+04
130	Methyl parathion	2.28 exp-04	5.10 exp+03	8.13 exp+01
131	Mirex [dechlorane]	1.10 exp-06	2.40 exp+07	7.80 exp+06
132	N,N-diphenylamine	3.40 exp-04	4.70 exp+02	3.98 exp+03
133	Parathion	8.24 exp-05	1.07 exp+04	6.45 exp+03
134	*p*-Chloroaniline [4-chlorobenzenamine]	4.15 exp-02	5.61 exp+02	6.76 exp+01
135	Toxaphene	1.21 exp-06	9.64 exp+02	2.00 exp+03
136	Trichlorfon [chlorofos]	5.98 exp-01	6.10 exp+00	1.95 exp+02

Structural descriptors

Several software programs, including *POLLY v. 2.3*,[14] *Triplet*,[15,16] and *Molconn-Z v. 3.5*,[17] were used to calculate a set of descriptors based solely on molecular structure. The descriptors numerically represent various aspects of the chemical structure and can be classified into one of three categories based on level of complexity: Topostructural (TS), Topochemical (TC), and 3-dimensional/geometrical (3D). The TS are the simplest in that no chemical information is encoded, with molecular structure viewed only in terms of atom connectivity. The TC descriptors, in addition to encoding information about how the atoms are connected within the molecule, also take chemical information into account, including atom type and bond type. The most complex of the three descriptor classes is the 3D, which encodes information on the 3-dimensional aspects of molecular structure. TS, TC, and 3D descriptors were used in a hierarchical manner in order to identify any model improvement upon addition of increasingly complex descriptor classes. For comparative purposes, single-class models were also developed.

Table 2 contains a complete list of the calculated TS, TC, and 3D descriptors. From this set, the following descriptors were removed and not used in the subsequent analyses: 1) Any descriptor with a constant value for all, or most all, of the136 chemicals in the data set, 2) one descriptor of each perfectly correlated pair (i.e., $r = 1.0$), as determined by the CORR procedure of the *SAS* statistical package,[18] and any descriptors with undefined values. A total of 260 descriptors were available for modeling.

Table 2. Symbols, definitions and classification of calculated molecular descriptors

	Topostructural (TS)
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
I_D^W	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
IC	Information content of the distance matrix partitioned by frequency of occurrences of distance h
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
${}^h\chi$	Path connectivity index of order h = 0–10
${}^h\chi_C$	Cluster connectivity index of order h = 3–6
${}^h\chi_{PC}$	Path-cluster connectivity index of order h = 4–6
${}^h\chi_{Ch}$	Chain connectivity index of order h = 3–10
P_h	Number of paths of length h = 0–10

Table 2. Contuined

J	Balaban’s J index based on topological distance
nrings	Number of rings in a graph
ncirc	Number of circuits in a graph
DN^2S_y	Triplet index from distance matrix, square of graph order (# of non-H atoms), and distance sum; operation y = 1–5
DN^21_y	Triplet index from distance matrix, square of graph order, and number 1; operation y = 1–5
$AS1_y$	Triplet index from adjacency matrix, distance sum, and number 1; operation y = 1–5
$DS1_y$	Triplet index from distance matrix, distance sum, and number 1; operation y = 1–5
ASN_y	Triplet index from adjacency matrix, distance sum, and graph order; operation y = 1–5
DSN_y	Triplet index from distance matrix, distance sum, and graph order; operation y = 1–5
DN^2N_y	Triplet index from distance matrix, square of graph order, and graph order; operation y = 1–5
ANS_y	Triplet index from adjacency matrix, graph order, and distance sum; operation y = 1–5
$AN1_y$	Triplet index from adjacency matrix, graph order, and number 1; operation y = 1–5
ANN_y	Triplet index from adjacency matrix, graph order, and graph order again; operation y = 1–5
ASV_y	Triplet index from adjacency matrix, distance sum, and vertex degree; operation y = 1–5
DSV_y	Triplet index from distance matrix, distance sum, and vertex degree; operation y = 1–5
ANV_y	Triplet index from adjacency matrix, graph order, and vertex degree; operation y = 1–5
	Topochemical (TC)
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
O_{orb}	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-suppressed graph
I_{orb}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
IC_r	Mean information content or complexity of a graph based on the r^th (r = 0–6) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r^th (r = 0–6) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r^th (r = 0–6) order neighborhood of vertices in a hydrogen-filled graph
${}^{h}\chi^{b}$	Bond path connectivity index of order h = 0–6
${}^{h}\chi^{b}_{C}$	Bond cluster connectivity index of order h = 3–6
${}^{h}\chi^{b}_{Ch}$	Bond chain connectivity index of order h = 3–6
${}^{h}\chi^{b}_{PC}$	Bond path-cluster connectivity index of order h = 4–6
${}^{h}\chi^{v}$	Valence path connectivity index of order h = 0–10
${}^{h}\chi^{v}_{C}$	Valence cluster connectivity index of order h = 3–6
${}^{h}\chi^{v}_{Ch}$	Valence chain connectivity index of order h = 3–10

Table 2. Contuined

${}^{h}\chi^{v}_{PC}$	Valence path-cluster connectivity index of order h = 4–6
J^{B}	Balaban's J index based on bond types
J^{X}	Balaban's J index based on relative electronegativities
J^{Y}	Balaban's J index based on relative covalent radii
AZV_{y}	Triplet index from adjacency matrix, atomic number, and vertex degree; operation y = 1−5
AZS_{y}	Triplet index from adjacency matrix, atomic number, and distance sum; operation y = 1–5
ASZ_{y}	Triplet index from adjacency matrix, distance sum, and atomic number; operation y = 1–5
AZN_{y}	Triplet index from adjacency matrix, atomic number, and graph order; operation y = 1–5
ANZ_{y}	Triplet index from adjacency matrix, graph order, and atomic number; operation y = 1–5
DSZ_{y}	Triplet index from distance matrix, distance sum, and atomic number; operation y = 1–5
$DN^{2}Z_{y}$	Triplet index from distance matrix, square of graph order, and atomic number; operation y = 1–5
nvx	Number of non-hydrogen atoms in a molecule
nelem	Number of elements in a molecule
fw	Molecular weight
si	Shannon information index
totop	Total Topological Index t
sumI	Sum of the intrinsic state values I
sumdelI	Sum of delta-I values
tets2	Total topological state index based on electrotopological state indices
phia	Flexibility index (kp1* kp2/nvx)
IdCbar	Bonchev-Trinajstić information index
IdC	Bonchev-Trinajstić information index
Wp	Wienerp
Pf	Plattf
Wt	Total Wiener number
knotp	Difference of chi-cluster-3 and path/cluster-4
knotpv	Valence difference of chi-cluster-3 and path/cluster-4
nclass	Number of classes of topologically (symmetry) equivalent graph vertices
numHBd	Number of hydrogen bond donors
numwHBd	Number of weak hydrogen bond donors
numHBa	Number of hydrogen bond acceptors
SHCsats	E-State of C sp^{3} bonded to other saturated C atoms
SHCsatu	E-State of C sp^{3} bonded to unsaturated C atoms
SHvin	E-State of C atoms in the vinyl group, =CH–
SHtvin	E-State of C atoms in the terminal vinyl group, $=CH_{2}$
SHavin	E-State of C atoms in the vinyl group, =CH–, bonded to an aromatic C
SHarom	E-State of C sp^{2} which are part of an aromatic system

Table 2. Contuined

SHHBd	Hydrogen bond donor index, sum of Hydrogen E-State values for –OH, =NH,–NH_2, –NH–, –SH, and #CH
SHwHBd	Weak hydrogen bond donor index, sum of C–H Hydrogen E-State values for hydrogen atoms on a C to which a F and/or Cl are also bonded
SHHBa	Hydrogen bond acceptor index, sum of the E-State values for –OH, =NH,–NH_2, –NH–, >N–, –O–, –S–, along with –F and –Cl
Qv	General Polarity descriptor
$NHBint_y$	Count of potential internal hydrogen bonders (y = 2–10)
$SHBint_y$	E-State descriptors of potential internal hydrogen bond strength (y =2–10)
	Electrotopological State index values for atoms types: SHsOH, SHdNH, SHsSH, SHsNH2, SHssNH, SHtCH, SHother, SHCHnX, Hmax Gmax, Hmin, Gmin, Hmaxpos, Hminneg, SsLi, SssBe, Sssss,Bem, SssBH, SsssB, SssssBm, SsCH3, SdCH2, SssCH2, StCH, SdsCH, SaaCH, SsssCH, SddC,StsC, SdssC, SaasC, SaaaC, SssssC, SsNH3p, SsNH2, SssNH2p, SdNH, SssNH, SaaNH, StN, SsssNHp, SdsN, SaaN, SsssN, SddsN, SaasN, SssssNp, SsOH, SdO, SssO, SaaO, SsF, SsSiH3, SssSiH2, SsssSiH, SssssSi, SsPH2, SssPH, SsssP, SdsssP, SsssssP, SsSH, SdS, SssS, SaaS, SdssS, SddssS, SssssssS, SsCl, SsGeH3, SssGeH2, SsssGeH, SssssGe, SsAsH2, SssAsH, SsssAs, SdsssAs, SsssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SddssSe, SsBr, SsSnH3, SssSnH2, SsssSnH, SssssSn, SsI, SsPbH3, SssPbH2, SsssPbH, SssssPb
	Geometrical / Shape (3D)
kp0	Kappa zero
kp1–kp3	Kappa simple indices
ka1–ka3	Kappa alpha indices

Statistical methodology

Each of the descriptors was transformed by the natural logarithm prior to model development, as their scales differed by several orders of magnitude. In order to avoid possible arithmetic error, a constant was added to the descriptor before log transforming. For descriptors with minimum values less than –1, the constant added was the smallest natural number that would provide a positive sum. For descriptors with minimum values greater than –1, the constant '1' was used. The dependent variables, i.e., S, K_{oc}, and K_{ow}, were also scaled by the natural logarithm. (The log scaled descriptors are available as supplemental material.)

For comparative purposes, results are reported based on two regression methodologies for the development of predictive models for each endpoint, namely ridge regression (RR)[19] and partial least squares (PLS).[20] Both methodologies make use of all available descriptors, as opposed to subset regression, and are useful when the number of descriptors exceeds the number of compounds in the data set (i.e., rank deficient data) and when the descriptors are highly intercorrelated. Formal comparisons have consistently shown that using a subset of available descriptors is less effective than using alternative regression methods that retain all available descriptors, such as RR and PLS, and deal with rank deficiency in another way.[21,22] With ridge regression, the descriptors are first transformed to their principal components (PCs). All PCs are retained but are "shrunk" differentially according to their eignevalues.[19] For each model developed, the cross-validated R^2 was obtained using the leave-one-out approach and can be calculated as follows (eq. 1):

$$R_{cv}^2 = 1 - \frac{PRESS}{SSTotal} \quad (1)$$

where *PRESS* is the prediction sum of squares and *SSTotal* is the total sum of squares.

It should be strongly stated that ordinary least squares (OLS) regression is inappropriate for use with rank deficient data, and that the conventional R^2 metric is without value in this situation. Unlike R^2, which tends to increase upon the addition of any descriptor, the cross-validated R^2 tends to decrease upon the addition of irrelevant descriptors and is a reliable measure of model predictability.[23]

RR and PLS models based on structural descriptors were developed for each of the five chemical subsets as well as for the combined set of 136 compounds. For comparative purposes, we also developed property-based models for the prediction of: a) S, based on K_{ow}, b) K_{oc}, based on K_{ow}, and c) K_{oc}, based on S. The *SAS* statistical package[18] was used to develop these ordinary least squares models, a methodology appropriate for the number of independent variables with respect to the number of observations.

Results and Discussion

The major objective of the study reported in this paper is to compare the relative effectiveness of physicochemical vis-à-vis calculated structural descriptors in the estimation of partitioning properties of chemicals of environmental concern.

For the sake of brevity, the many highly-parameterized models are not reported. However, Tables 3–7 provide the associated cross-validated R^2 values for the five chemical subsets, while the cross-validated R^2 values for the combined data are found in Table 8. In all cases, there is no significant improvement in model quality when the more complex 3D descriptors are added to the topological (i.e., TS and TC) descriptors.

When examining the regression results, it's important to keep in mind that while R^2 is necessarily a nonnegative number, this is not true of the cross-validated R^2, which can take on negative values if the model is extremely poor (see eq. 1).

With respect to water solubility, there is improvement in model quality upon the addition of TC descriptors to the TS indices (especially pronounced with the aromatics, herbicides, and pesticides) except in the case of the aliphatic subset, for which the TS descriptors, alone, provide the best water solubility model. The best solubility model for the total set of 136 compounds is the TS+TC model, with a cross-validated R^2 value of 0.739 (Table 8). High-quality models obtained for the various subsets include the TC model for the aromatics with a cross-validated R^2 value of 0.905 (Table 4), and the TS+TC model for the polyaromatic hydrocarbons with a cross-validated R^2 value of 0.808 (Table 6).

Table 3. Regression results for the aliphatic subset (N = 26)

	S ($R^2_{c.v.}$)		K_{oc} ($R^2_{c.v.}$)		K_{ow} ($R^2_{c.v.}$)	
Model Type	*RR*	*PLS*	*RR*	*PLS*	*RR*	*PLS*
TS	0.704	0.569	0.658	0.622	0.748	0.697
TS+TC	0.649	0.641	0.544	0.374	0.761	0.520
TS+TC+3D	0.562	0.644	0.540	0.350	0.759	0.523
TS	0.704	0.569	0.658	0.622	0.748	0.697
TC	0.582	0.595	0.615	0.491	0.765	0.627
3D	0.637	0.549	0.682	0.665	0.739	0.690

Table 4. Regression results for the aromatic subset (N = 43)

	S ($R^2_{c.v.}$)		K_{oc} ($R^2_{c.v.}$)		K_{ow} ($R^2_{c.v.}$)	
Model Type	*RR*	*PLS*	*RR*	*PLS*	*RR*	*PLS*
TS	0.459	0.314	–0.019	–0.025	0.668	0.512
TS+TC	0.901	0.880	–0.018	0.148	0.927	0.886
TS+TC+3D	0.905	0.881	–0.037	0.138	0.929	0.891
TS	0.459	0.314	–0.019	-0.025	0.668	0.512
TC	0.905	0.884	0.069	0.177	0.923	0.870
3D	0.692	0.637	0.408	0.279	0.756	0.893

Table 5. Regression results for the herbicide subset (N=18)

	S ($R^2_{c.v.}$)		K_{oc} ($R^2_{c.v.}$)		K_{ow} ($R^2_{c.v.}$)	
Model Type	*RR*	*PLS*	*RR*	*PLS*	*RR*	*PLS*
TS	–0.445	–6.44	–1.67	0.282	0.128	–0.300
TS+TC	0.502	–0.440	0.358	0.143	0.079	–0.334
TS+TC+3D	0.496	–0.436	0.353	0.136	0.084	–0.331
TS	–0.445	–6.44	–1.67	0.282	0.128	–0.300
TC	0.596	0.227	0.385	0.258	0.012	–0.387
3D	0.285	0.340	0.278	–1.07	0.198	–0.321

Table 6. Regression results for the polycyclic aromatic hydrocarbon subset (*N*=19)

	S (*$R^2_{c.v.}$*)		K_{oc} (*$R^2_{c.v.}$*)		K_{ow} (*$R^2_{c.v.}$*)	
Model Type	*RR*	*PLS*	*RR*	*PLS*	*RR*	*PLS*
TS	0.746	0.698	0.437	0.416	0.842	0.772
TS+TC	0.808	0.789	0.454	0.286	0.942	0.921
TS+TC+3D	0.806	0.753	0.451	0.299	0.942	0.922
TS	0.746	0.698	0.437	0.416	0.842	0.772
TC	0.807	0.782	0.510	0.374	0.944	0.936
3D	0.693	0.695	0.360	0.289	0.793	0.690

Table 7. Regression results for the pesticides subset (*N*=30)

	S (*$R^2_{c.v.}$*)		K_{oc} (*$R^2_{c.v.}$*)		K_{ow} (*$R^2_{c.v.}$*)	
Model Type	*RR*	*PLS*	*RR*	*PLS*	*RR*	*PLS*
TS	0.453	0.032	0.305	0.219	0.189	–0.697
TS+TC	0.759	0.582	0.705	0.690	0.450	0.196
TS+TC+3D	0.757	0.564	0.703	0.698	0.451	0.212
TS	0.453	0.032	0.305	0.219	0.189	–0.697
TC	0.735	0.593	0.699	0.644	0.440	0.156
3D	0.289	0.311	0.187	0.052	0.073	–0.033

Table 8. Regression results for the combined data sets (*N*=136)

	S (*$R^2_{c.v.}$*)		K_{oc} (*$R^2_{c.v.}$*)		K_{ow} (*$R^2_{c.v.}$*)	
Model Type	*RR*	*PLS*	*RR*	*PLS*	*RR*	*PLS*
TS	0.577	0.480	0.483	0.400	0.482	0.456
TS+TC	0.739	0.688	0.717	0.649	0.544	0.488
TS+TC+3D	0.738	0.717	0.717	0.641	0.547	0.503
TS	0.577	0.480	0.483	0.400	0.482	0.456
TC	0.735	0.725	0.720	0.664	0.570	0.451
3D	0.412	0.479	0.309	0.381	0.235	0.318

With respect to K_{oc}, again we see significant improvement in model quality upon the addition of TC descriptors to the TS indices, in most cases, with a notable exception in the aliphatic subset. Overall, K_{oc} models are inferior to the water solubility models, with the cross-validated R^2 values ranging from 0.385 to 0.705 for the various chemical subsets, and the best model (TC) for the combined set of chemicals being 0.720 (Table 8).

Similar trends are seen with the K_{ow} models, except that it is the herbicide data, rather than the aliphatic data, that is better modeled with the TS descriptors than with the TC. Very good TC models were found for the aromatics and the polycyclic aromatic hydrocarbons, with cross-validated R^2 values of 0.927 and 0.944, respectively. The best K_{ow} model for the total set of 136 compounds was the TC model, with a cross-validated R^2 value of 0.570 (Table 8).

The statistical analyses of the complete set of 136 compounds revealed a number of compounds with high influence upon the models. These were considered, independently, for each of the three endpoints, and additional models for the combined set of chemicals were developed omitting these compounds as outliers (Table 9). For the *S* model, paraquat, cylclophosphamide, and dechlorane were omitted. These same compounds were omitted from the K_{ow} model, in addition to trifluralin, kepone, and trichlorofon. From the K_{oc} model, diethylstilbestrol, trifluralin, 1,2:7,8-dibenzopyrene, cyclophosphamide, kepone, dechlorane, and trichloron were omitted. With the removal of these outliers, model improvement was observed. E.g., with respect to the TC models, the cross-validated R^2 values improved from 0.735 to 0.846 for the *S* model, from 0.720 to 0.790 for the K_{oc} model, and from 0.570 to 0.865 for the K_{ow} model.

Table 9. Regression results for combined data sets, with outliers removed with respect to each endpoint

	S ($R^2_{c.v.}$)		K_{oc} ($R^2_{c.v.}$)		K_{ow} ($R^2_{c.v.}$)	
	N = 133		N = 131		N = 130	
Model Type	*RR*	*PLS*	*RR*	*PLS*	*RR*	*PLS*
TS	0.649	0.598	0.579	0.534	0.588	0.496
TS+TC	0.848	0.848	0.790	0.764	0.862	0.848
TS+TC+3D	0.849	0.846	0.766	0.749	0.864	0.849
TS	0.649	0.598	0.579	0.534	0.588	0.496
TC	0.846	0.848	0.790	0.738	0.865	0.858
3D	0.495	0.582	0.334	0.464	0.316	0.447

The results of the comparative property-based models are summarized in Table 10. The structure-based models were superior to the property-based models for the herbicides, pesticides, aromatics, and the combined set of compounds. In addition, it is the TS and TC descriptors that provide the best structural models for these chemical subsets. It should be noted that the TS descriptors alone provide the best solubility model for the aliphatic subset. The property-based models are superior to the structure-based models for the aliphatics and the polyaromatic hydrocarbons. With respect to the K_{oc} models, the herbicides, aromatics, and the combined set of chemicals are better modeled with the structure-based descriptors, while the pesticides, aliphatics, and polycyclic aromatic hydrocarbons are better modeled with K_{ow} and solubility. It is of interest to note that the 3D descriptors provide the best structure-based models for the aliphatics and the aromatics, with respect to the prediction of K_{oc}.

Table 10. Comparative property-property correlations

	$S = f(K_{ow})$	$K_{oc} = f(K_{ow})$	$K_{oc} = f(S)$
Model Type	$R^2_{c.v.}$	$R_{c.v.}$	$R^2_{c.v.}$
Herbicides (*N* = 18)	0.500	–0.128	–0.650
Pesticides (*N* = 30)	0.614	0.718	0.751
Aliphatic (*N* = 26)	0.806	0.873	0.827
PAHs (*N* = 19)	0.862	0.633	0.832
Aromatics (*N* = 43)	0.788	0.539	0.282
All chemicals (*N* = 136)	0.721	0.693	0.706

Conclusions

We used computed molecular descriptors, viz., TS, TC, and geometrical or 3-D indices, in the prediction of water solubility, octanol/water partition coefficient, and soil/ sediment partition coefficient for a diverse set of chemicals. Results show that a combination of TS plus TC indices gives the best models, in most cases, for the entire set of molecules as well as the various subsets, with the addition of the 3-D descriptors to the independent variables resulting in either marginal or no improvement in model quality. It must be noted that the breakdown of the entire set into various subsets by Chu and Chan based on structure, e.g., aliphatic, aromatics, polycyclic aromatic, and biochemical action, e.g., herbicide, pesticide, is quite arbitrary. The diversity of the set of calculated molecular descriptors allowed for the development of good quality QSARs both for the structurally diverse set and for the various subsets. This is in line with our earlier mutagen/non-mutagen classification study with a diverse set of 508 chemicals where we obtained good results both for structurally defined smaller subsets and the entire diverse set.[9]

We reported results based on two methods of QSAR model development: a) ridge regression (RR), and b) partial least squares (PLS). In earlier studies with various types of property and toxicity data,[9,24] we reported RR, PLS, and principal components regression (PCR) results and found that RR is the most effective of the three, whereas PCR was the least useful based on cross-validated R^2. PCR results were not reported in this paper for brevity. A perusal of the QSAR results in Tables 1–10 shows that here, also, RR usually outperformed PLS.

The overall high quality of the models obtained for all three properties, viz., water solubility, and octanol/water partition coefficient, and soil/sediment partition coefficient, both for the diverse set and the different subsets using purely calculated structural descriptors indicate that QSARs developed in this paper will find application in the estimation of partitioning properties of chemicals of environmental concern.

Supplementary material available

See website: http://www.arkat-usa.org/ark/journal/2005/I02_Anand/1214/1214supplinfo.htm

Acknowledgements

This paper represents contribution number 377 from the Center for Water and the Environment of the Natural Resources Research Institute. Research was supported by the United States Air Force (Grant F49620-02-1-0138) and the Agency for Toxic Substances and Disease Registry (Cooperative Agreement Number 572112).

References

1. USEPA; http://www.epa.gov/opptintr/newchems/invntory.htm: 2003.
2. ATSDR; http://www.atsdr.cdcgov/clist.html: 2003, Vol. 2002.
3. Chu, W.; Chan, K.-H. *Sci. Total Environ.* **2000**, *248*, 1.
4. Gao, C. *Environ. Toxicol. Chem.* **1996**, *15*, 1089.
5. Auer, C. M.; Zeeman, M.; Nabholz, J. V.; Clements, R. G. *SAR QSAR Environ. Res.* **1994**, *2*, 29.
6. Huuskonen, J. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773.
7. Basak, S. C.; Gute, B. D.; Grunwald, G. D. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1054.
8. Niemi, G. J.; Basak, S. C.; Veith, G. D.; Grunwald, G. *Environ. Toxicol. Chem.* **1992**, *11*, 891.
9. Basak, S. C.; Mills, D.; Gute, B. D.; Hawkins, D. M. In *Quantitative Structure-Activity Relationships (QSAR) Models of Mutagens and Carcinogens*; Benigni, R., Ed.; CRC Press: Boca Raton, FL, 2003, p 207.
10. Basak, S. C.; Hawkins, D. M.; Mills, D. In *Advances in Molecular Similarity*;Carbo-Dorca, R., Mezey, P. G., Eds.; Kluwer: Amsterdam, 2002; Vol. 5, in press.
11. Basak, S. C.; Mills, D.; Hawkins, D. M.; El-Masri, H. A. *SAR QSAR Environ. Res.* **2002**, *13*, 649.
12. EPA Report, EPA-600/8-90/-003, 1990.
13. Verschueren, K. *Handbook of Environmental Data on Organic Chemicals*; 2nd ed.; Van Norstrand Reinhold Co., Inc., 1983.
14. POLLY v. 2.3, Basak, S. C.; Harriss, D. K.; Magnuson, V. R. Copyright of the University of Minnesota, 1988.
15. Filip, P. A.; Balaban, T. S.; Balaban, A. T. *J. Math. Chem.* **1987**, *1*, 61.
16. Basak, S. C.; Balaban, A. T.; Grunwald, G. D.; Gute, B. D. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 891.
17. Molconn-Z v 3.50, Hall Associates Consulting: Quincy, MA, 2000.

18. SAS/STAT User Guide, Release 6.03 Edition; SAS Institute Inc.: Cary, NC, 1988
19. Hoerl, A. E.; Kennard, R. W. *Technometrics* **1970**, *8*, 27.
20. Wold, H. In *Perspectives in Probability and Statistics, Papers in Honor of M. S. Bartlett*; Gani, J., Ed.; Academic Press: London, 1975.
21. Rencher, A. C.; Pun, F. C. *Technometrics* **1980**, *22*, 49.
22. Frank, I. E.; Friedman, J. H. *Technometrics* **1993**, *35*, 109.
23. Hawkins, D. M.; Basak, S. C.; Mills, D. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 579.
24. Basak, S. C.; Mills, D.; Mumtaz, M. M.; Balasubramanian, K. *Indian J. Chem.* **2003**, *42A*, 1385.