A Comparative Study of Topological and Geometrical Parameters in Estimating Normal Boiling Point and Octanol/Water Partition Coefficient

Subhash C. Basak,* Brian D. Gute, and Gregory D. Grunwald

Natural Resources Research Institute, University of Minnesota, Duluth, Duluth, Minnesota 55811

Received February 22, 1996[⊗]

We have used topological, topochemical and geometrical parameters in predicting: (a) normal boiling point of a set of 1023 chemicals and (b) lipophilicity (log *P*, octanol/water) of 219 chemicals. The results show that topological and topochemical variables can explain most of the variance in the data. The addition of geometrical parameters to the models provide marginal improvement in the model's predictive power. Among the three classes of descriptors, the topochemical indices were the most effective in predicting properties.

1. INTRODUCTION

A contemporary trend in theoretical chemistry, biomedicinal chemistry, drug design, and toxicology is the prediction of relevant properties of chemicals using structure−activity relationships (SARs).[1−9] A large number of SARs published in recent literature use parameters which can be calculated directly from molecular structure, as opposed to experimentally derived properties or parameters.[1−6,10−30] The principal motivating factor behind this trend is our need to know many properties of a very large number of chemicals, both for practical drug design and hazard assessment of chemicals.[13,31] All these properties cannot be determined experimentally due to limited resources. The modeling of the properties of chemicals using SARs based on calculated molecular descriptors has the following three major components:[13,32]

1. Optimal representation of the chemical species by a chosen model object (structure representation).

2. Enumeration of relevant characteristics of the model object (parameterization).

3. Development of qualitative or quantitative models to predict properties using the selected structural characteristics (property prediction).

The first step in the overall process is representation (Figure 1). The term molecular structure represents a set of nonequivalent concepts. There is no reason to believe that when discussing different topics, e.g., chemical synthesis, reaction rates, spectroscopic transitions, reaction mechanisms, and ab initio calculations, that the term "molecular structure" represents the same fundamental reality.[13,33] In fact, the various models of chemicals, e.g., classical valence bond representation, different graph theoretic representations, ball and spoke model of molecules, minimum energy conformation, and symbolic representation of molecules by Hamiltonian operators, are nothing but various representations of the same chemical entity. Once the model object is chosen, subsequent processes of parameterization and property estimation can be done in more than one way. Consequently, the field of theoretical SAR is comprised of a set of diverse modeling activities.

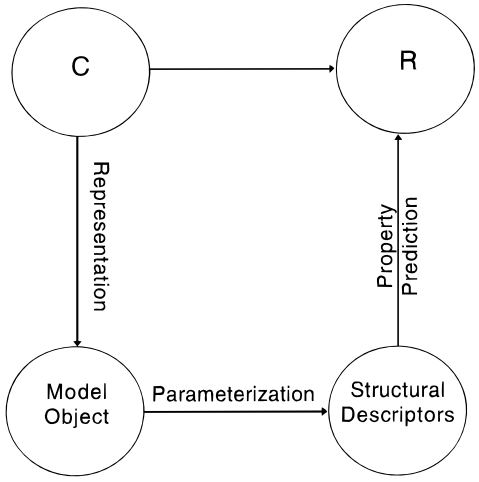

Figure 1. The processes of experimental determination vis-a-vis theoretical prediction of properties from SARs. *C* represents the set of chemicals and *R* the set of real numbers.

A convenient method of representing chemical species is by means of molecular graphs, where atoms are represented by vertices and bonds are depicted by edges.[34] Invariants derived from graphs can be used to characterize chemical structure. When a molecule is represented by a simple planar graph which does not distinguish among atoms or bond types, such invariants quantify molecular topology without being sensitive to such important chemical features like presence of heteroatoms or bond multiplicity. Such invariants may be termed "topological". On the other hand, when molecules are represented by graphs which are properly weighted to represent heterogeneity of atom types and bonding pattern, invariants derived from such graphs are chemically more realistic.[35] Such invariants have been found to be more useful as compared to the topological indices. We call such indices "topochemical" parameters, because they quantify both topology (connectivity) of atoms as well as the chemical characteristics of the specific molecular structure.

Another set of descriptors which have been used in many SARs are the geometrical or shape parameters, which encode information about the spatial characteristics of atoms in the molecule.[36−38]

In practical drug design and hazard assessment, where it is necessary to carry out very rapid estimation of a large

* All correspondence to be addressed to: Dr. Subhash C. Basak, Natural Resources Research Institute, University of Minnesota, Duluth, Duluth, MN 55811

[⊗] Abstract published in *Advance ACS Abstracts,* October 1, 1996.

S0095-2338(96)00024-8 CCC: $12.00

number of properties with no or very little empirical input, SARs based on topological, topochemical, and geometrical parameters can be of practical use. Therefore, in this paper, we have carried out a comparative study of topological, topochemical, and geometrical parameters in estimating (a) boiling point of a subset of the Toxic Substances Control Act (TSCA) Inventory comprising 1023 molecules and (b) lipophilicity of a set of 219 diverse compounds. The results are presented here with an analysis of the relative contributions of the three classes of indices in the development of SAR models.

2. MATERIALS AND METHODS

2.1. Normal Boiling Point Database. We used a subset of the Toxic Substances Control Act (TSCA) Inventory[31] for which measured normal boiling point values were available and where HB_1 was equal to zero. HB_1 is a measure of the hydrogen bonding potential of a chemical. There were 1023 chemicals in the TSCA Inventory which satisfied these two criteria. Because of the large number of chemicals in this study, we are not listing the data for these chemicals in this paper. An electronic copy of the data may be obtained by contacting the authors.

2.2. Log *P* Database. Measured values of log *P* were obtained from CLOGP,[39] namely, the STARLIST group of chemicals. For this study, we used only chemicals where HB_1 was equal to zero. Also, the range of log *P* values for the purpose of estimation was restricted to −2 to 5.5. Actual measurements for log *P* beyond this range have been shown to be problematic.[14] Table 1 provides a listing of the 219 chemicals that met these conditions.

2.3. Calculation of Topological and Geometric Parameters. Most of the topological indices used for property estimation were calculated by the computer program POLLY.[40] These indices include the molecular connectivity indices developed by Randić[18] and Kier and Hall,[35] Wiener number,[41] and frequency of path lengths of varying size. Information theoretic indices defined on the hydrogen-filled and hydrogen-suppressed molecular graph were calculated by POLLY using the methods of Basak *et al.*,[42,43] Roy *et al.*,[44] Raychaudhury *et al.*,[45] and Bonchev and Trinajstić.[46] The *J* indices of Balaban[47−49] were calculated using software developed by the authors. The hydrogen bonding parameter, HB_1, was calculated using a program developed by Basak[50] and is based on the ideas of Ou *et al.*[51]

van der Waal's volume (V_W) was calculated using Sybyl 6.2.[52] The 3-D Wiener numbers[37] were calculated using Sybyl with an SPL (Sybyl Programming Language) program developed by the authors. The calculation of the 3-D Wiener number consists of summing the entries in the upper triangular submatrix of the topographic Euclidean distance matrix for a molecule. The 3-D coordinates of each atom, needed for these computations, was determined using CONCORD 3.2.1.[53] For this paper, two variants of the 3-D Wiener number have been calculated, ${}^{3D}W$ and ${}^{3D}W_H$, where the hydrogen atoms have been excluded and included in the calculation, respectively.

In Table 2, the symbols for all topological and geometric parameters have been listed. A brief definition of each parameter is provided in Table 2 as well.

The parameters in Table 2 were then classified as being topological, topochemical, or geometric. Table 2 is organized to show where each parameter was classed. The topological parameters consist of those indices in which atom specific information and bonding type are ignored in calculation of the index. The topochemical indices account for atom and bond type information. The geometric parameters are based upon 3-D coordinate information of the molecule.

2.4. Statistical Analyses. Since the difference in magnitude for the topological and topochemical indices can vary greatly, they were transformed by the natural logarithm of the index plus one. One was added since many of the indices can be zero. The geometric parameters were transformed by the natural logarithm of the parameter.

Two regression procedures were used in the development of models. When the number of independent variables was high, typically greater than 25, a stepwise regression procedure to maximize improvement to R^2 was used. When the number of independent variables was small, all possible subsets regression was used. All regression models were developed using procedure REG of the SAS statistical package.[54]

For both data sets, we randomly split the chemicals into approximately equal (50%/50%) training and test sets. For the BP data, there were 512 chemicals in the training set and 511 chemicals in the test set. For log *P*, there were 114 chemicals in the training set and 105 chemicals in the test set. The training set and test set of chemicals are identified in Table 1 for the log *P* data. Models were developed using the training set of chemicals. These models were then used to predict the property values of the test chemicals. Final models were then developed using the combined training and test set of chemicals.

Initial models for the dependent property (BP or log *P*) were developed using only the topological class of indices. Once the best topological model was determined, the topological indices used in the model were added to the set of topochemical indices. Then the best model from this combined set of indices was determined. Finally, the topological and/or topochemical indices used in the best model so far were added to the set of geometric parameters, and the best model using all of these parameters was determined.

3. RESULTS

3.1. TSCA Boiling Point Estimation. Stepwise regression analyses for BP of the training set of chemicals is summarized in Table 3. As is shown in Table 3, the topological model using 11 parameters resulted in an explained variance (R^2) of 80.8% and standard error (*s*) of 40.9 °C. Addition of the topochemical parameters with the 11 topological parameters increased the effectiveness of the model significantly. The resulting model used nine parameters, two topological parameters, and seven topochemical parameters. This model had an R^2 of 96.5% and *s* of 17.4 °C. All subsets regression of the nine topological and topochemical parameters retained thus far and the three geometric parameters resulted in a ten parameter model. This model included the nine topological and topochemical parameters and the geometric parameter ${}^{3D}W_H$. This model represented a slight improvement with R^2 of 96.7% and *s* of 16.8 °C.

Application of the three models to the test set of chemicals resulted in comparable R^2 and *s* and are listed in Table 3.

Table 1. Observed and Estimated Lipophilicity (Log *P*, Octanol/Water) for 219 Chemicals with HB_1 Equal to Zero

no.	chemical name	obs log *P*	est log *P* (eq 4)	est log *P* (eq 5)	est log *P* (eq 6)	no.	chemical name	obs log *P*	est log *P* (eq 4)	est log *P* (eq 5)	est log *P* (eq 6)
1[a]	1,4-dimethylnaphthalene	4.37	4.25	4.37	4.41	74[a]	1,2,4-trichlorobenzene	4.02	3.65	3.84	3.79
2	cyclopropane	1.72	1.26	0.83	0.82	75[a]	2,2′,6-pcb	5.48	4.89	5.07	5.09
3	3,4-dimethylchlorobenzene	3.82	3.65	3.68	3.74	76	2-butyne	1.46	2.22	2.49	2.46
4	2,2-diphenyl-1,1,1-trichloroethane	4.87	4.90	4.93	4.99	77	azulene	3.20	3.59	3.52	3.45
5	2,6-dimethylnaphthalene	4.31	4.15	4.24	4.27	78[a]	trifluoromethylthiobenzene	3.57	3.56	2.91	2.93
6	hexafluoroethane	2.00	2.63	2.59	2.33	79[a]	2,5-pcb	5.16	4.62	4.90	4.89
7[a]	1-iodoheptane	4.70	4.04	4.27	4.24	80[a]	1,2,3-trichlorocyclohexene(34)	2.84	3.60	3.57	3.58
8[a]	allylbromide	1.79	2.22	2.04	2.06	81	biphenyl	4.09	4.18	4.33	4.32
9[a]	1,5-dimethylnaphthalene	4.38	4.23	4.38	4.41	82[a]	*p*-xylene	3.15	3.45	3.37	3.42
10	1,8-dimethylnaphthalene	4.26	4.31	4.41	4.43	83[a]	ethylene	1.13	0.70	0.93	0.97
11[a]	1,2,3-trichlorobenzene	4.05	3.60	3.64	3.63	84	thiophenol	2.52	3.08	3.01	3.04
12[a]	2-ethylthiophene	2.87	3.20	2.69	2.73	85[a]	bromotrifluoromethane	1.86	2.20	2.12	1.97
13	methylchloride	0.91	0.70	0.86	0.79	86	9-methylanthracene	5.07	5.07	4.90	4.92
14	γ-phenylpropylfluoride	2.95	3.73	3.26	3.29	87[a]	trichloroethylene	2.42	2.63	2.44	2.44
15	iodobenzene	3.25	3.08	3.68	3.68	88[a]	1,4-dimethyltetrachlorocyclohexane	4.40	4.18	4.03	4.11
16	1-methylpentachlorocyclohexane	4.04	4.18	4.20	4.24	89	propylene	1.77	1.38	1.59	1.71
17[a]	ethane	1.81	0.70	1.50	1.47	90	cyclohexene	2.86	2.55	2.72	2.74
18	2,3′-pcb	5.02	4.71	4.99	4.97	91[a]	methylthiobenzene	2.74	3.28	3.02	2.97
19	cyclopentane	3.00	2.19	2.35	2.37	92	methylfluoride	.51	0.70	0.57	0.53
20	ethylchloride	1.43	1.38	1.48	1.52	93	γ-phenylpropyliodide	3.90	3.73	4.06	4.06
21	2-phenylthiophene	3.74	3.88	4.01	3.96	94	2,3,4′-pcb	5.42	4.89	5.10	5.10
22	trichlorofluoromethane	2.53	2.20	2.34	2.29	95[a]	fluoropentachlorocyclohexane	3.19	4.18	3.87	3.89
23[a]	fluoroform	0.64	1.85	0.57	0.42	96	1,2,3,5-tetrachlorobenzene	4.92	3.91	4.08	4.04
24	dimethyldisulfide	1.77	2.22	1.57	1.39	97	2,2′-pcb	4.90	4.65	4.80	4.82
25[a]	propane	2.36	1.38	1.97	2.01	98	1-butene	2.40	2.22	1.96	2.05
26	hexamethylbenzene	5.11	4.18	4.94	4.97	99[a]	1,3-dimethylnaphthalene	4.42	4.29	4.43	4.46
27	butanethiol	2.28	2.80	2.81	2.87	100[a]	1,7-dimethylnaphthalene	4.44	4.23	4.43	4.45
28[a]	diethylsulfide	1.95	2.80	2.68	2.67	101[a]	1-methylnaphthalene	3.87	3.95	4.08	4.07
29	cyclohexane	3.44	2.55	2.83	2.87	102	2,6-pcb	4.93	4.70	4.83	4.85
30[a]	diphenyldisulfide	4.41	4.62	4.57	4.53	103[a]	α-bromotoluene	2.92	3.28	3.47	3.42
31	*m*-fluorobenzylchloride	2.77	3.55	2.95	2.99	104	2,2′,3′-trichlorobiphenyl	5.31	4.89	5.22	5.20
32	1-chloropropane	2.04	2.22	1.92	1.97	105	hexafluorobenzene	2.22	4.18	3.20	2.97
33	2,4-dichlorobenzylchloride	3.82	4.01	3.67	3.69	106[a]	3-bromothiophene	2.62	2.49	2.73	2.72
34	*m*-chlorotoluene	3.28	3.58	3.30	3.34	107[a]	1,2,3,5-tetramethylbenzene	4.17	3.91	4.27	4.32
35[a]	butane	2.89	2.22	2.39	2.43	108	halothane	2.30	3.01	2.16	2.19
36	1,2,3-trimethylbenzene	3.66	3.60	3.89	3.93	109	2,4,6-pcb	5.47	4.97	5.02	5.04
37	1,1-difluoroethylene	1.24	1.85	0.72	0.79	110	1,1-dichloroethylene	2.13	1.85	1.89	1.97
38[a]	1-chlorobutane	2.64	2.80	2.67	2.70	111	*o*-dibromobenzene	3.64	3.29	3.94	3.88
39	2,3-dibromothiophene	3.53	2.98	3.22	3.23	112	1,2,4,5-tetramethylbenzene	4.00	3.82	4.24	4.29
40[a]	pentafluorethylbenzene	3.36	3.24	3.15	3.18	113	1-hexene	3.39	3.25	3.17	3.21
41[a]	1,2,4,5-tetrabromobenzene	5.13	3.82	5.06	4.94	114[a]	neopentane	3.11	2.20	3.12	3.28
42	*o*-dichlorobenzene	3.38	3.29	3.19	3.19	115	chloroform	1.97	1.85	2.11	2.06
43[a]	1,2,3,4-tetrachlorobenzene	4.64	3.81	4.01	3.97	116[a]	1-fluorobutane	2.58	2.80	2.15	2.20
44[a]	tribromoethene	3.20	2.63	3.37	3.26	117[a]	pyrene	4.88	5.46	4.90	4.88
45	pentane	3.39	2.80	3.01	3.03	118	1,1-dichloro-2,2-diphenylethane	4.51	4.95	4.88	4.94
46[a]	isobutane	2.76	1.85	2.61	2.71	119[a]	isobutylene	2.34	1.85	2.47	2.61
47[a]	mirex	5.28	5.10	5.26	5.18	120	diphenylmethane	4.14	4.40	4.51	4.54
48[a]	1,3-dichlorobenzene	3.60	3.58	3.24	3.23	121	isopropylbenzene	3.66	3.35	3.62	3.67
49[a]	1,2-dimethylnaphthalene	4.31	4.25	4.41	4.43	122[a]	naphthalene	3.30	3.43	3.38	3.34
50[a]	2-ethylnaphthalene	4.38	4.32	4.23	4.24	123[a]	1-heptene	3.99	3.86	3.48	3.50
51[a]	cycloheptatriene	2.63	3.59	2.74	2.74	124	2,2-dimethylbutane	3.82	2.88	3.45	3.55
52[a]	3-chlorobiphenyl	4.58	4.42	4.65	4.64	125	1-fluoropentane	2.33	3.25	2.79	2.82
53[a]	3-ethylthiophene	2.82	3.20	2.72	2.75	126[a]	*o*-xylene	3.12	3.29	3.44	3.49
54	1,3,5-tribromobenzene	4.51	4.00	4.54	4.48	127[a]	ethylbenzene	3.15	3.28	3.25	3.26
55[a]	β-phenylethylchloride	2.95	3.50	3.31	3.33	128[a]	trichloromethylthiobenzene	3.78	3.56	3.59	3.61
56	acenaphthene	3.92	4.49	3.94	3.95	129[a]	thiophene	1.81	2.19	1.64	1.62
57	*m*-dibromobenzene	3.75	3.58	4.06	3.98	130	bromochloromethane	1.41	1.38	1.49	1.47
58	dichlorodifluoromethane	2.16	2.20	1.88	1.83	131[a]	1,2-dichlorotetrafluoroethane	2.82	2.63	2.71	2.65
59	toluene	2.73	3.08	3.04	3.05	132[a]	2-chlorobiphenyl	4.38	4.43	4.65	4.65
60[a]	anthracene	4.45	4.85	4.62	4.59	133	2,4′-dichlorobiphenyl	5.10	4.68	4.88	4.87
61[a]	hexachlorocyclopentadiene	5.04	4.00	4.99	4.86	134[a]	1,3,5-trichlorobenzene	4.15	4.00	3.48	3.50
62	3-phenyl-1-chloropropane	3.55	3.73	3.56	3.58	135	1-octene	4.57	4.04	3.77	3.78
63[a]	bibenzyl	4.79	4.62	4.69	4.71	136	methylbromide	1.19	0.70	1.23	1.07
64[a]	1-chloroheptane	4.15	4.04	3.72	3.71	137[a]	phenylethylsulfide	3.20	3.50	3.37	3.36
65[a]	2,4-dichlorotoluene	4.24	3.65	3.60	3.64	138	1-ethyl-2-methylbenzene	3.53	3.54	3.81	3.84
66[a]	1,1-dichloroethane	1.79	1.85	1.93	2.02	139[a]	propylbenzene	3.72	3.50	3.56	3.58
67[a]	(β)-benzothiophene	3.12	3.15	3.24	3.17	140[a]	indane	3.18	3.15	3.06	3.04
68[a]	2-bromothiophene	2.75	2.49	2.62	2.61	141	2-chloropropane	1.90	1.85	2.22	2.33
69	chlorodifluoromethane	1.08	1.85	0.75	0.75	142[a]	phenylazide	2.59	3.50	2.83	2.88
70[a]	pentachlorobenzene	5.17	4.02	4.62	4.52	143	2,4-dibromotetrachlorocyclohexane	3.98	4.18	4.25	4.29
71	9,10-dihydroanthracene	4.25	4.85	4.31	4.34	144[a]	tetrachloroethylene	3.40	3.01	3.69	3.52
72	1,3-(bis-chloromethyl)benzene	2.72	3.85	3.49	3.51	145	1-nonene	5.15	4.27	3.97	3.98
73	chlorobenzene	2.84	3.08	2.90	2.90	146	2,3-dimethylbutane	3.85	3.01	3.41	3.50

Table 1 (Continued)

no.	chemical name	obs log *P*	est log *P* (eq 4)	est log *P* (eq 5)	est log *P* (eq 6)	no.	chemical name	obs log *P*	est log *P* (eq 4)	est log *P* (eq 5)	est log *P* (eq 6)
147[a]	dichlorofluoromethane	1.55	1.85	1.25	1.30	184[a]	2,3,6-trimethylnaphthalene	4.73	4.46	4.61	4.64
148[a]	1,1,2,2-tetrachloroethane	2.39	3.01	2.91	2.90	185[a]	difluoromethane	.20	1.38	0.20	0.11
149[a]	1,2,4-trimethylbenzene	3.78	3.65	3.95	3.98	186	1,2,4-trifluorobenzene	2.52	3.65	2.63	2.55
150[a]	fluorobenzene	2.27	3.08	2.39	2.40	187	bromobenzene	2.99	3.08	3.38	3.33
151	butylbenzene	4.26	3.73	3.81	3.83	188	hexachloro-1,3-butadiene	4.78	4.26	5.00	4.86
152[a]	ethylbromide	1.61	1.38	1.98	1.95	189	vinylbromide	1.57	1.38	1.76	1.78
153[a]	tetrafluoromethane	1.18	2.20	1.61	1.29	190[a]	*o*-chlorotoluene	3.42	3.29	3.33	3.36
154[a]	*p*-cymene	4.10	3.93	3.88	3.92	191[a]	α-chlorotoluene	2.30	3.28	3.09	3.10
155[a]	*p*-chlorotoluene	3.33	3.45	3.17	3.22	192	1,4-cyclohexadiene	2.30	2.55	2.42	2.46
156[a]	1-bromopropane	2.10	2.22	2.35	2.34	193	1-bromoheptane	4.36	4.04	4.06	4.00
157[a]	bromocyclohexane	3.20	3.08	3.47	3.46	194	styrene	2.95	3.28	3.15	3.17
158[a]	2-methylthiophene	2.33	2.49	2.39	2.41	195	chlorotrifluoromethane	1.65	2.20	1.56	1.45
159	diphenylsulfide	4.45	4.40	4.48	4.47	196[a]	(dimethyl)phenylphosphine	2.57	3.35	2.99	2.92
160[a]	1,2,4,5-tetrachlorobenzene	4.82	3.82	3.93	3.91	197	cycloocta-1,5-diene	3.16	3.23	2.88	2.93
161	1,1,1-trichloroethane	2.49	2.20	2.43	2.52	198	tetrachlorocyclohexane	2.82	3.82	3.50	3.55
162[a]	p-dichlorobenzene	3.52	3.45	3.08	3.10	199	1-bromooctane	4.89	4.27	4.23	4.17
163	1-bromobutane	2.75	2.80	3.15	3.14	200	2-methylnaphthalene	3.86	3.90	4.03	4.01
164[a]	*p*-chlorobiphenyl	4.61	4.50	4.57	4.56	201	3-methylthiophene	2.34	2.49	2.44	2.46
165[a]	cyclopropylbenzene	3.27	2.98	3.01	3.02	202[a]	methylenechloride	1.25	1.38	1.38	1.35
166[a]	2,6-dichlorotoluene	4.29	3.60	3.48	3.54	203[a]	hexachlorobenzene	5.31	4.18	5.09	4.93
167[a]	allene	1.45	1.38	1.42	1.48	204	indene	2.92	3.15	3.01	3.01
168[a]	b-phenylethylbromide	3.09	3.50	3.66	3.64	205	*tert*-butylbenzene	4.11	3.26	3.92	3.99
169[a]	1,3-butadiene	1.99	2.22	1.88	1.97	206	1,2-dichloroethane	1.48	2.22	1.92	1.91
170	2-chlorothiophene	2.54	2.49	2.14	2.16	207[a]	1,3,5-trimethylbenzene	3.42	4.00	3.81	3.87
171	1-bromopentane	3.37	3.25	3.62	3.58	208[a]	phenanthrene	4.46	4.88	4.69	4.68
172[a]	γ-phenylpropylbromide	3.72	3.73	3.87	3.84	209[a]	benzene	2.13	2.55	2.40	2.39
173	1,3-cyclohexadiene	2.47	2.55	2.47	2.50	210	3,3,3-trifluoropropylbenzene	3.31	3.80	3.19	3.24
174[a]	pentamethylbenzene	4.56	4.02	4.65	4.69	211[a]	α-(2,2,2-trichloroethyl)styrene	4.56	3.93	4.04	4.13
175[a]	*p*-dibromobenzene	3.79	3.45	3.81	3.76	212[a]	2,3-dimethylnaphthalene	4.40	4.20	4.25	4.28
176	1,4-pentadiene	2.48	2.80	2.32	2.42	213[a]	1,3-dichloropropane	2.00	2.80	2.47	2.47
177[a]	methyliodide	1.51	0.70	1.48	1.42	214	1,2,3,4-tetramethylbenzene	4.11	3.81	4.26	4.31
178[a]	1,1-difluoroethane	.75	1.85	1.04	1.11	215[a]	stilbene-t	4.81	4.62	4.79	4.78
179[a]	1-bromohexane	3.80	3.86	3.80	3.75	216	fluorene	4.18	4.65	4.22	4.21
180[a]	*m*-xylene	3.20	3.58	3.44	3.49	217[a]	2-fluoro-3-bromotetrachlorocyclohexane	3.28	4.18	4.06	4.09
181	dibenzothiophene	4.38	4.65	4.44	4.40	218[a]	allylbenzene	3.23	3.50	3.37	3.41
182	ethyliodide	2.00	1.38	2.28	2.34	219[a]	carbontetrachloride	2.83	2.20	3.27	3.10
183	trifluoromethylbenzene	3.01	3.26	2.77	2.80						

[a] Training chemicals.

The largest difference in variance explained was for the topological parameter model. For this model, R^2 decreased from 80.8% to 79.5% or 1.3% less variance explained. However, the standard error for the test chemicals was 0.1 °C lower. For the other two models, the R^2 of the test chemicals was within 0.6% of that seen for the training chemicals. Standard errors for the test chemicals were within 1 °C of the standard error for the training set of chemicals.

Regression analysis of the set of training and test chemicals combined showed similar results as analysis of the training set of chemicals. Using only the topological class of indices, stepwise regression resulted in an eight parameter model to estimate boiling point:

$$\mathrm{BP} = -21.9 + 30.6(\mathrm{W}) - 21.5(\mathrm{O}) + 69.9({}^{3}\chi) + 35.8({}^{6}\chi) - 106.5({}^{6}\chi_{\mathrm{C}}) - 96.1({}^{5}\chi_{\mathrm{Ch}}) - 17.7\,({}^{5}\chi_{\mathrm{PC}}) + 19.5(\mathrm{P}_{10}) \quad (1)$$

$$n = 1023, \quad R^2 = 81.2\%, \quad s = 39.7\ ^{\circ}\mathrm{C}, \quad F = 547$$

These eight parameters were added to the set of topochemical parameters. Again, stepwise regression was used to develop a model using the eight topological and all topochemical indices. The best model to estimate boiling point consisted of eight parameters again:

$$\mathrm{BP} = -332.9 + 134.6({}^{6}\chi) + 10.9(\mathrm{P}_{10}) + 110.0(\mathrm{IC}_{0}) - 133.8({}^{6}\chi^{\mathrm{b}}) - 80.2({}^{3}\chi^{\mathrm{b}}_{\mathrm{C}}) + 176.5({}^{0}\chi^{\mathrm{v}}) + 44.8({}^{2}\chi^{\mathrm{v}}) + 16.8({}^{5}\chi^{\mathrm{v}}_{\mathrm{PC}}) \quad (2)$$

$$n = 1023, \quad R^2 = 96.1\%, \quad s = 18.0\ ^{\circ}\mathrm{C}, \quad F = 3151$$

Only two of the topological indices used in eq 1 were retained by the regression procedure in eq 2: ${}^{6}\chi$ and P_{10}. The improvement in R^2 was very significant, going from 81.2% for eq 1 to 96.1% for eq 2. Also, the model error decreased by over half, dropping from 39.7 °C to 18.0 °C.

Using all subsets regression on the eight parameters of eq 2 and the three geometric parameters resulted in a ten parameter model as follows:

$$\mathrm{BP} = -285.7 + 125.3({}^{6}\chi) + 10.6(\mathrm{P}_{10}) + 74.5(\mathrm{IC}_{0}) - 125.0({}^{6}\chi^{\mathrm{b}}) - 86.3({}^{3}\chi^{\mathrm{b}}_{\mathrm{C}}) + 175.3({}^{0}\chi^{\mathrm{v}}) + 49.1({}^{2}\chi^{\mathrm{v}}) + 18.7({}^{5}\chi^{\mathrm{v}}_{\mathrm{PC}}) - 9.1({}^{3\mathrm{D}}W_{\mathrm{H}}) + 8.1({}^{3\mathrm{D}}W) \quad (3)$$

$$n = 1023, \quad R^2 = 96.3\%, \quad s = 17.6\ ^{\circ}\mathrm{C}, \quad F = 2650$$

Equation 3 contains all of the parameters from eq 2 plus the two variants of the 3-D Wiener number. The addition

Table 2. Symbols, Definitions, and Classifications of Topological and Geometrical Parameters

	Topological
I_D^W	information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\overline{I_D^W}$	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
$\overline{IC}$	information content of the distance matrix partitioned by frequency of occurrences of distance h
O	order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
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 $h = 0-6$
${}^h\chi$	cluster connectivity index of order $h = 3-5$
${}^h\chi_{PC}$	path-cluster connectivity index of order $h = 4-6$
${}^h\chi_{Ch}$	chain connectivity index of order $h = 5-6$
P_h	number of paths of length $h = 0-10$
J	Balaban's J index based on distance
	Topochemical
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 rth ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	structural information content for rth ($r = 0-6$) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	complementary information content for rth ($r = 0-6$) 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 $h = 3-5$
${}^h\chi^b_{Ch}$	bond chain connectivity index of order $h = 5-6$
${}^h\chi^b_{PC}$	bond path-cluster connectivity index of order $h = 4-6$
${}^h\chi^v$	valence path connectivity index of order $h = 0-6$
${}^h\chi^v_C$	valence cluster connectivity index of order $h = 3-5$
${}^h\chi^v_{Ch}$	valence chain connectivity index of order $h = 5-6$
${}^h\chi^v_{PC}$	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
	Geometric
V_W	van der Waal's volume
${}^{3D}W$	3-D Wiener number for the hydrogen-suppressed geometric distance matrix
${}^{3D}W_H$	3-D Wiener number for the hydrogen-filled geometric distance matrix

Table 3. Summary of Regression Results for the Training Set of Chemicals and Predictions of Test Set of Chemicals for Dependent Variable BP (°C) for Three Parameter Classes

	training set ($N = 512$)				test set ($N = 511$)	
parameter class	variables included	F	R^2	s	R^2	s
topological	$\overline{IC}$, O, M_2, ${}^3\chi$, ${}^5\chi$, ${}^6\chi$, ${}^3\chi_C$, ${}^6\chi_C$, ${}^5\chi_{Ch}$, P_{10}, J	191	80.8	40.9	79.5	40.8
topological + topochemical	${}^6\chi$, P_{10}, IC_0, ${}^6\chi^b$, ${}^3\chi^b_C$, ${}^0\chi^v$, ${}^4\chi^v$, ${}^4\chi^v_C$, ${}^4\chi^v_{PC}$	1547	96.5	17.4	96.0	18.0
topological + topochemical + geometric	${}^6\chi$, P_{10}, IC_0, ${}^6\chi^b$, ${}^3\chi^b_C$, ${}^0\chi^v$, ${}^4\chi^v$, ${}^4\chi^v_C$, ${}^4\chi^v_{PC}$, ${}^{3D}W_H$	1486	96.7	16.8	96.1	17.7

Table 4. Summary of Regression Results for the Training Set of Chemicals and Predictions of Test Set of Chemicals for Dependent Variable Log P for Three Parameter Classes

	training set ($N = 114$)				test set ($N = 105$)	
parameter class	variables included	F	R^2	s	R^2	s
topological	IDW, ${}^2\chi$, ${}^5\chi$, ${}^6\chi_C$, ${}^5\chi_{Ch}$, ${}^6\chi_{Ch}$, ${}^4\chi_{PC}$, P_7, P_9	40.7	77.9	0.57	73.8	0.60
topological + topochemical	${}^5\chi_{Ch}$, ${}^4\chi_{PC}$, SIC_1, ${}^3\chi^b$, ${}^4\chi^b_{PC}$, ${}^0\chi^v$, J^Y	122.6	89.0	0.40	85.6	0.45
topological + topochemical + geometric	${}^5\chi_{Ch}$, ${}^4\chi_{PC}$, SIC_1, ${}^4\chi^b_{PC}$, ${}^0\chi^v$, J^Y, ^{3D}W	123.0	89.0	0.39	85.5	0.45

of the two 3D-Wiener numbers resulted in only a very slight increase in the predictive power of the model. The standard error (s) decreased by only 0.4 °C with the addition of the geometric parameters and R^2 increased from 96.1% to 96.3%, an increase of only 0.2% of the variance explained by eq 2 over eq 3. A scatterplot of observed boiling point *vs* estimated boiling point using eq 3 is shown in Figure 2.

3.2. Log *P* Estimation. Stepwise regression analyses for log P of the training set of chemicals is summarized in Table 4. The topological parameter model included nine variables. These nine variables explained 77.9% of the variance with a standard error of 0.57. Regression analysis of these nine topological parameters and the topochemical parameters resulted in a better model with only seven parameters. This model included two topological parameters and five topochemical. The R^2 increased from 77.9% to 89.0% and s decreased from 0.57 to 0.40. Adding the geometric parameters provided a very minor increase. For this model, ^{3D}W replaced ${}^3\chi^b$, the R^2 remained the same, and s decreased from 0.40 to 0.39.

Application of these models to the test set of chemicals resulted in slightly decreased variance explained and slightly

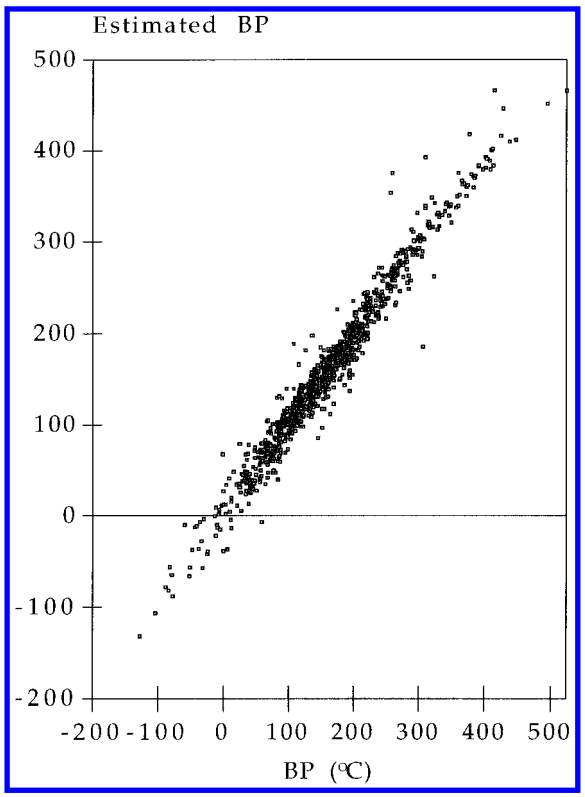

Figure 2. Scatterplot of observed boiling point *vs* estimated boiling point using eq 3 for 1023 diverse chemicals.

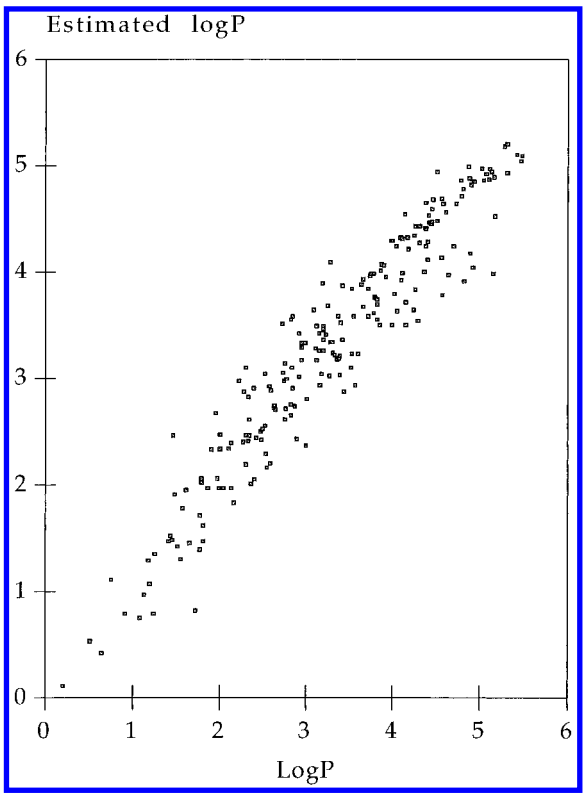

Figure 3. Scatterplot of observed log *P* *vs* estimated log *P* using eq 6 for 219 diverse chemicals.

increased standard error. All R^2 for the test set differed by no more than 4.1% of the R^2 seen for the training set. The standard error of the test set of chemicals was within 0.06 of the standard error of the training set. These results can be seen in Table 4.

As with the BP data set, regression analyses of the combined training and test sets was similar to the analyses of the training sets. Starting with topological parameters only, the following seven parameter model was developed to estimate log *P*:

$$\log P = -1.42 + 1.08(\mathrm{W}) - 1.58({}^{2}\chi) + 1.51({}^{6}\chi) - 0.92({}^{6}\chi_{\mathrm{C}}) - 0.32(\mathrm{P}_{7}) + 0.20(\mathrm{P}_{10}) + 1.97(\mathrm{J}) \quad (4)$$

$$n = 219, \quad R^2 = 78.9\%, \quad s = 0.54, \quad F = 112$$

The seven parameters of eq 4 were added to the set of topochemical indices, and a new model was developed using stepwise regression. This new model consisted of ten parameters:

$$\log P = -2.13 - 0.20({}^{2}\chi) + 0.18(\mathrm{P}_{10}) - 1.86(\mathrm{IC}_{0}) + 1.33(\mathrm{CIC}_{2}) - 0.92(\mathrm{CIC}_{3}) - 1.36({}^{6}\chi^{\mathrm{b}}) + 5.76({}^{0}\chi^{\mathrm{v}}) - 2.98({}^{1}\chi^{\mathrm{v}}) + 0.54({}^{4}\chi^{\mathrm{v}}) - 0.39({}^{3}\chi^{\mathrm{v}}_{\mathrm{C}}) \quad (5)$$

$$n = 219, \quad R^2 = 90.8\%, \quad s = 0.36, \quad F = 206$$

As with the boiling point models, only two of the topological parameters were retained in eq 5, ${}^{2}\chi$ and P_{10}. Also, just like the boiling point models, the addition of the topochemical parameters resulted in a significant increase in the quality of log *P* estimation.

All subsets regression using the ten parameters of eq 5 and the geometric parameters resulted in the following 11 parameter model:

$$\log P = -5.60 + 0.19(\mathrm{P}_{10}) - 1.46(\mathrm{IC}_{0}) + 1.09(\mathrm{CIC}_{2}) - 0.77(\mathrm{CIC}_{3}) - 1.36({}^{6}\chi^{\mathrm{b}}) + 5.34({}^{0}\chi^{\mathrm{v}}) - 3.41({}^{1}\chi^{\mathrm{v}}) + 0.55({}^{4}\chi^{\mathrm{v}}) - 0.41({}^{3}\chi^{\mathrm{v}}_{\mathrm{C}}) + 1.10(V_{\mathrm{W}}) - 0.17({}^{3\mathrm{D}}W) \quad (6)$$

$$n = 219, \quad R^2 = 91.2\%, \quad s = 0.35, \quad F = 194$$

Equation 6 differs from eq 5 with the removal of ${}^{2}\chi$ and the addition of V_{W} and ${}^{3\mathrm{D}}\mathrm{W}$. The addition of the geometric parameters resulted in only slight improvement in the ability to estimate log *P*.

Estimated log *P* values using eqs 4−6 may be found in Table 1. Figure 3 shows a scatterplot of observed log *P* *vs* estimated log *P* using eq 6.

4. DISCUSSION

The objective of this paper was to carry out a comparative study of the effectiveness of topological, topochemical, and geometrical parameters in SAR. To this end, we used these three classes of parameters in predicting normal boiling point of a diverse set of 1023 chemicals and log *P* of a set of 219 chemicals. To further assess the utility of these models for predictive purposes, the data sets were split into training and test sets by randomly assigning chemicals to one or the other. Models developed using the training sets of chemicals were used to predict the relevant property of the test chemicals.

As can be seen in Tables 3 and Table 4, the models developed using the training sets of chemicals could predict BP and log *P* of the test chemicals as accurately as they could estimate these properties for the training chemicals. Therefore, it seemed reasonable to combine the training and test sets to develop the regression models.

Both for boiling point and log *P*, topological variables gave a reasonable predictive model. The addition of topochemical parameters to the set of independent variables resulted in substantial improvement in model performance. Further addition of geometrical variables gave slight improvement in explained variance in these data.

Our modeling approach in this paper was a hierarchical one, beginning with parameters derived from the simplest (topological) representation of molecules. Such indices are derived from simple graphs which are unweighted and, consequently, do not represent the reality of chemicals very well. The next tier of variables, topochemical indices, quantify information both about topology as well as atom types and bonding pattern. Finally, geometrical or 3-D parameters were used for modeling. The results show that the addition of chemical information makes a substantial contribution to the predictive power of the models for both boiling point and log *P*. It would be interesting to see whether this trend is valid for other properties.

ACKNOWLEDGMENT

This is contribution number 183 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported in part by grant F49620-94-1-0401 from the United States Air Force, a grant from Exxon Corporation, and the Structure-Activity Relationship Consortium (SARCON) of the Natural Resources Research Institute of the University of Minnesota.

REFERENCES AND NOTES

(1) Wilkins, C. L.; Randić, M. *Theoret. Chim. Acta (Berl.)* **1980**, *58*, 45.
(2) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64.
(3) Johnson, M. A.; Basak, S. C.; Maggiora, G. *Mathl. Comput. Modeling* **1988**, *11*, 630.
(4) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. *Discrete Appl. Math.* **1988**, *19*, 17.
(5) Basak, S. C.; Bertelsen, S.; Grunwald, G. D. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 270.
(6) Basak, S. C.; Grunwald, G. D. *Math. Modeling Sci. Comput.* **1994**, in press.
(7) Willet, P.; Winterman, V. *Quant. Struct-Act. Relat.* **1986**, *5*, 18.
(8) Fisanick, W.; Cross, K. P.; Rusinko, III, A. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 664.
(9) Lajiness, M. S. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Publishers: New York, 1990; p 299.
(10) Niemi, G. J.; Basak, S. C.; Veith, G. D. In *Proceedings of the First Conference of the International Society of Environmental Protection, Vol. 2, No. 1;* Zirm, K., Mayer, J., Eds.; WB. Druck Bmgh. and Co.: Reiden, Austria, 1989; p 57.
(11) Basak, S. C.; Niemi, G. J.; Veith, G. D. *J. Math. Chem.* **1990**, *4*, 185.
(12) Basak, S. C.; Niemi, G. J.; Veith, G. D. *Mathematical and Computer Modeling* **1990**, *14*, 511.
(13) Basak, S. C.; Niemi, G. J.; Veith, G. D. *J. Math. Chem.*, **1991**, *7*, 243.
(14) Niemi, G. J.; Basak, S. C.; Veith, G. D.; Grunwald, G. D. *Environ. Toxicol. Chem.* **1992**, *11*, 893.
(15) Basak, S. C.; Grunwald, G. D. *Math. Modeling Sci. Comput.* **1993**, *2*, 735.
(16) Randić, M.; Basak, S. C. *Mathematical and Computer Modeling* (in preparation).
(17) Randić, M. *Int. J. Quantum. Chem.* **1984**, *11*, 137.
(18) Randić, M. *J. Am. Chem. Soc.* **1975**, *97*, 6609.
(19) Randić, M. In *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New York, 1990; p 77.
(20) Balasubramanian, K. *SAR QSAR Environ. Res.* **1994**, *2*, 59.
(21) Rouvray, D. H. *J. Mol. Struct.(Theochem)* **1995**, *336*, 101.
(22) Balaban, A. T.; Basak, S. C.; Colburn, T.; Grunwald, G. D *J. Chem. Inf. Comput. Sci*. **1994**, *34*, 1118.
(23) Basak, S. C.; Grunwald, G. D. *SAR QSAR Environ. Res.* **1994**, *2*, 289.
(24) Basak, S. C.; Grunwald, G. D. *New J. Chem.* **1995**, *19*, 231.
(25) Basak, S. C.; Grunwald, G. D. In *Proceeding of the XVI International Cancer Congress*; Rao, R. S., Deo, M. G., Sanghvi, L. D., Eds.; Monduzzi: Bologna, Italy, 1995; p 413.
(26) Basak, S. C.; Grunwald, G. D. *J. Chem. Inf. Comput. Sci*. **1995**, *35*, 366.
(27) Basak, S. C.; Grunwald, G. D. *SAR QSAR Environ. Res.* **1995**, *3*, 265.
(28) Basak, S. C.; Grunwald, G. D. *Chemosphere* **1995**, *31*, 2529.
(29) Basak, S. C.; Bertelsen, S.; Grunwald, G. D. *Toxicology Lett.* **1995**, *79*, 239.
(30) Basak, S. C.; Gute, B. D.; Drewes, L. R. *Pharm. Res.* **1996**, *13*, 775.
(31) Auer, C. M.; Nabholz, J. V.; Baetcke, K. P. *Environ. Health Perspect.* **1990**, *87*, 183.
(32) Bunge, M. *Method, Model and Matter*; D. Reidel Publishing Company: Dordrecht-Holland/Boston, 1973.
(33) Weininger, S. J. *J. Chem. Educ.* **1984**, *61*, 939.
(34) Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1983.
(35) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Research Studies Press: Chichester, England, 1986.
(36) Moriguchi, I.; Kanada, Y. *Chem. Pharm. Bull.* **1977**, *25*, 926.
(37) Bogdanov, B.; Nikolić, S.; Trinajstić, N. *J. Math. Chem.* **1989**, *3*, 299.
(38) Kamlet, M. J.; Doherty, R. M.; Abraham, M. H.; Marcus, Y.; Taft, R. W. *J. Phys. Chem.* **1988**, *92*, 5244.
(39) Leo, A.; Weininger, D. *CLOGP Version 3.2 User Reference Manual*; Medicinal Chemistry Project, Pomona College, Claremont, CA. 1984.
(40) Basak, S. C.; Harriss, D. K.; Magnuson, V. R. *POLLY Version 2.3*; Copyright of the University of Minnesota, 1988.
(41) Wiener, H. *J. Am. Chem. Soc.* **1947**, *69*, 17.
(42) Basak, S. C.; Roy, A. B.; Ghosh, J. J. In *Proceedings of the IInd International Conference on Mathematical Modeling*; Avula, X. J. R., Bellman, R., Luke, Y. L., Rigler, A. K., Eds.; University of Missouri-Rolla: Rolla, MO, 1979; p 851.
(43) Basak, S. C.; Magnuson, V. R. *Arzneim.-Forsch./Drug Res.* **1983**, *33*, 501.
(44) Roy, A. B.; Basak, S. C.; Harriss, D. K.; Magnuson, V. R. In *Mathematical Modeling in Science and Technology*; Avula, X. J. R., Kalman, R. E., Lipais, A. I., Rodin, E. Y., Eds.; Pergamon: New York, 1984; p 745.
(45) Raychaudhury, C.; Ray, S. K.; Ghosh, J. J.; Roy, A. B.; Basak, S. C. *J. Comput. Chem.* **1984**, *5*, 581.
(46) Bonchev, D.; Trinajstić, N. *J. Chem. Phys.* **1977**, *67*, 4517.
(47) Balaban, A. T. *Chem. Phys. Lett.* **1982**, *89*, 399.
(48) Balaban, A. T. *Pure Appl. Chem.* **1982**, *55*, 199.
(49) Balaban, A. T. *MATCH* **1986**, *21*, 115.
(50) Basak, S. C. *H-Bond*; Copyright of the University of Minnesota, 1988.
(51) Ou, Y. C.; Ouyang, Y.; Lien, E. J. *J. Mol. Sci.* **1986**, *4*, 89.
(52) Tripos Associates, Inc. *Sybyl Version 6.2*; Tripos Associates, Inc.: St. Louis, MO, 1995.
(53) Tripos Associates, Inc. *CONCORD Version 3.2.1*; Tripos Associates, Inc.: St. Louis, MO, 1995.
(54) SAS Institute, Inc. In *SAS/STAT User Guide, Release 6.03 Edition*; SAS Institute, Inc.: Cary, NC, 1988.

CI960024I