ELSEVIER

Journal of Molecular Graphics and Modelling 20 (2001) 95–109

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

Molecular similarity-based estimation of properties: a comparison of three structure spaces

Brian D. Gute, Subhash C. Basak*

Center for Water and the Environment, Natural Resources, Research Institute, University of Minnesota, 5013 Miller Trunk Hwy, Duluth, MN 55811, USA

Received 11 December 2000; received in revised form 6 April 2001; accepted 2 May 2001

Abstract

Similarity, like beauty, is an intuitive concept based on personal perception and bias. In the realm of molecular similarity, each method is user defined based on the features deemed important. A method's efficacy depends on the set of descriptors used to define the intermolecular similarity of chemicals and on the mathematical function used to quantify similarity. Quantitative molecular similarity analysis (QMSA) methods, based on experimental data or computed molecular descriptors, have emerged as powerful tools for analog selection and property estimation. We have carried out a comparative study of similarity spaces derived from atom pairs and a large set of topological indices for two diverse sets of chemicals: (a) a set of 469 chemicals with vapor pressure data from the TSCA inventory, and (b) a set of 213 chemicals with lipophilicity data from the STARLIST inventory. These spaces were used for the KNN-based estimation of properties ($K = 1$–10, 15, 20, 25). The results for the QMSA models developed in this paper are also compared with model estimates derived from hierarchical QSARs. © 2001 Elsevier Science Inc. All rights reserved.

Keywords: Structure space; Hierarchical QSAR; Molecular similarity; Atom pairs; Topological indices; Property estimation

1. Introduction

A current trend in predictive toxicology is the estimation of physicochemical, biomedicinal, and toxicological properties of chemicals directly from calculated structural parameters. Two popular techniques adopted in this area are (a) quantitative structure–activity/property relationships (QSARs/QSPRs) [1] and (b) quantitative molecular similarity analysis (QMSA) [2–11].

Risk assessment, a fundamental part of chemical regulation, often must be carried out with limited or no experimental data [1]. Case in point, the Toxic Substances Control Act (TSCA) inventory has entries for nearly 80,000 distinct chemicals. Of the chemicals included in the inventory, only 50% have some physicochemical property data and only about 15% have data from any genotoxicity bioassays. Yet the premanufacture notification (PMN) process of the United States Environmental Protection Agency (USEPA) requires the agency to respond to the PMN submission within 90 days with an assessment of the human health and environmental hazard posed by the chemical. Likewise, the National Toxicology Program (NTP) has a long list of chemicals that need to be tested in the 2-year rodent bioassay. This testing carries a price tag of about 4 million dollars per chemical, a staggering cost when applied to several hundred compounds. Worldwide, millions of chemicals are known; but few of them have the physicochemical and biological test data necessary for proper risk assessment.

Predictive toxicology methods for assessing human and environmental health hazards usually attempt to predict more complex biological or toxicologically-relevant physicochemical properties from more basic experimental test data. However, the long list of candidate chemicals to be tested precludes the possibility of thoroughly testing even a small number of them. Recently, chemical manufacturers have begun an initiative to test some 2800 high-volume chemicals at a cost of about 700 million dollars. These test batteries are expected to be completed by the end of 2004. Even if this testing program goes as planned, the majority of chemicals in commerce in the United States will still have only patchy experimental test data of toxicological and ecotoxicological relevance. A whole suite of physicochemical and biological data is necessary for effective risk assessment. Regulators need to know the hazard posed by a chemical, but they also need to be able to assess the likelihood and potential route of exposure. Thus, not only do we need a suite of toxicity data, we also need a suite of physicochemical data

* Corresponding author. Tel.: +1-218-720-4230; fax: +1-218-720-4328.
E-mail address: sbasak@nrri.umn.edu (S.C. Basak).

PII: S1093-3263(01)00104-8

to predict how the chemical will behave under expected conditions.

Due to this lack of data, pragmatic approaches have been devised by chemical regulators for the hazard and risk assessment of new and existing chemicals. Various research groups have suggested the use of available structural and functional data for the development of models to predict the hazard posed by new chemicals [12–14].

One such method uses chemical analogs in property estimation. Chemical analogs (neighbors) are selected based on their degree of similarity with the chemical of interest (probe) with respect to molecular architecture or some experimental property. In other words, two chemicals, X_1 and X_2, must have some reasonable "proximity" in the descriptor space to be considered similar. This allows a regulator to examine the properties of "well-defined" analogs and then make more informed decisions based on an extrapolation of their properties. Use of experimental properties in such situations is not very practical since experimental data are not available for a majority of chemicals.

Using a subset of nearly 20,000 TSCA chemicals available to us through our collaborators at the USEPA — Environmental Research Laboratory in Duluth, we queried this database and others for property data on these compounds and found that only 76 of them had experimental data for the following seven properties — boiling point, melting point, $\log P$ (octanol–water), α (hydrogen bond donor acidity), β (hydrogen bond acceptor basicity), $V/100$ (molar volume), and π (polarizability) [15]. These are only a few of the properties necessary for effective chemical risk assessment.

As can be seen from the above, for molecular similarity methods to be useful in practical chemical risk assessment they must be based on calculated, rather than experimental, properties. Additionally, they must either select useful analogs or give sufficiently accurate property estimates that will be useful in risk assessment. However, there are many theoretical descriptors that quantify different, sometimes partially overlapping, aspects of molecular structure. Our research group has been involved in extracting useful and only weakly intercorrelated structural information from large collections of calculated molecular descriptors. Principal components (PCs) derived from descriptor sets have been used in the development of structure spaces for analog selection [4–10,15–18]. In this study, we have created three structure spaces based on: (a) atom pairs (APs); (b) PCs calculated from 102 topological indices (TIs), 98 of which are calculated by POLLY 2.3 and 4 calculated by in-house software (PC1); and (c) PCs generated from an expanded set of 202 TIs (PC2). We will present a comparative study of these three spaces in predicting $\log P$ for a set of 213 STARLIST chemicals and normal vapor pressure for a set of 469 TSCA chemicals, both properties necessary for effective risk assessment.

2. Methodology

2.1. log P database

The $\log P$ set consists of 213 diverse compounds used in an earlier study by Basak et al. [19]. Measured values of $\log P$ were obtained from CLOGP [20] as this set represents a subset of the STARLIST group of chemicals. In this study, as in the earlier study, we have used only chemicals where HB_1 was equal to 0 (HB_1 is a measure of the hydrogen bonding potential of a chemical). Therefore, none of the chemicals have available hydrogen bonding centers, creating a more homogeneous group of chemicals and a more densely packed similarity space. Also, the chemicals were selected such that their $\log P$ values fall within the range of -2 to 5.5, since actual measurements for $\log P$ beyond this range have been shown to be problematic [21]. It should be noted that six compounds included in the earlier study have been excluded from this study. All six of these compounds consisted of only two non-hydrogen atoms. This was done to eliminate compounds with a single unique atom pair (AP) and because some of the topological indices (TIs) cannot be calculated for compounds with fewer than three non-hydrogen atoms. The chemicals used in this study and their $\log P$ values from CLOGP are reported in Table 1.

2.2. Normal vapor pressure database

The vapor pressure dataset, consisting of 469 diverse compounds, represents a subset of the TSCA inventory [1] used in an earlier study by Basak et al. [22]. Measured values for normal vapor pressure were obtained from the ASTER [23] (Assessment Tools for the Evaluation of Risk) database of the USEPA. This subset contains a diverse set of chemicals for which normal vapor pressure (p_{vap}) ranges between 3 and 10 000 mmHg measured at standard temperature (25°C). As with the $\log P$ set, seven compounds, consisting of only two non-hydrogen atoms, included in the earlier study have been removed. This set of chemicals and their observed data are not reported for the sake of brevity, but the set can be characterized as follows: 253 hydrocarbons, 92 halogenated hydrocarbons, 124 other compounds including alcohols, esters, carboxylic acids, amines, ketones, nitriles and sulfides.

2.3. Calculation of topological indices

The first set of 102 TIs used in this study include the Wiener number [24], molecular connectivity indices as calculated by Randić and coworkers [25,26] and Kier and Hall [27], frequency of path lengths of varying size, information theoretic indices defined on distance matrices of graphs using the methods of Bonchev and Trinajstić [28], as well as those of Raychaudhury et al. [29], parameters defined on the neighborhood complexity of vertices defined by Basak and coworkers for hydrogen-filled molecular graphs [30–32] and Balaban's J indices [33–35]. The 98

Table 1
List of the 213 STARLIST chemicals included in this study and their experimental and estimated values for log *P*

ID no.	Compound name	log *P* (experimental)	Estimated log *P*		
			AP	TI 102	TI 202
1	1,4-Dimethylnaphthalene	4.37	4.35	4.35	4.32
2	Cyclopropane	1.72	3.22	1.75	2.67
3	3,4-Dimethylchlorobenzene	3.82	4.01	3.66	3.48
4	2,2-Diphenyl-1,1,1-trichloroethane	4.87	4.33	4.28	4.54
5	2,6-Dimethylnaphthalene	4.31	4.57	4.38	4.39
6	Hexafluoroethane	2.00	2.00	2.83	3.09
7	1-Iodoheptane	4.70	4.26	4.47	4.63
8	Allylbromide	1.79	2.44	1.63	2.43
9	1,5-Dimethylnaphthalene	4.38	4.34	4.34	4.32
10	1,8-Dimethylnaphthalene	4.26	4.38	4.38	4.38
11	1,2,3-Trichlorobenzene	4.05	4.01	4.62	4.61
12	2-Ethylthiophene	2.87	2.99	2.50	2.67
14	γ-Phenylpropylfluoride	2.95	3.73	3.41	3.91
15	Iodobenzene	3.25	2.79	2.95	3.37
16	1-Methylpentachlorocyclohexane	4.04	3.80	3.62	3.24
18	2,3′-pcb	5.02	5.03	5.19	5.13
19	Cyclopentane	3.00	3.32	2.91	2.79
20	Ethylchloride	1.43	2.20	1.73	1.81
21	2-Phenylthiophene	3.74	3.21	2.91	3.02
22	Trichlorofluoromethane	2.53	2.33	1.89	2.01
23	Fluoroform	0.64	0.92	1.78	1.16
24	Dimethyldisulfide	1.77	2.42	1.81	1.72
25	Propane	2.36	2.16	1.98	1.60
26	Hexamethylbenzene	5.11	4.37	4.47	4.66
27	Butanethiol	2.28	3.02	2.66	2.61
28	Diethylsulfide	1.95	2.88	2.69	2.70
29	Cyclohexane	3.44	3.85	2.64	3.08
30	Diphenyldisulfide	4.41	4.62	4.68	4.62
31	*m*-Fluorobenzylchloride	2.77	2.51	3.34	4.03
32	1-Chloropropane	2.04	2.77	2.16	2.46
33	2,4-Dichlorobenzylchloride	3.82	4.13	3.68	4.13
34	*m*-Chlorotoluene	3.28	3.40	3.89	3.38
35	Butane	2.89	2.72	2.68	2.72
36	1,2,3-Trimethylbenzene	3.66	3.62	4.02	3.98
37	1,1-Difluoroethylene	1.24	0.42	1.17	0.70
38	1-Chlorobutane	2.64	2.16	2.54	2.43
39	2,3-Dibromothiophene	3.53	3.20	2.64	2.69
40	Pentafluorethylbenzene	3.36	3.29	2.71	2.92
41	1,2,4,5-Tetrabromobenzene	5.13	3.77	4.05	4.15
42	*o*-Dichlorobenzene	3.38	3.74	3.66	3.83
43	1,2,3,4-Tetrachlorobenzene	4.64	4.87	4.42	4.49
44	Tribromoethene	3.20	3.59	3.26	2.52
45	Pentane	3.39	2.59	2.70	2.61
46	Isobutane	2.76	2.88	2.38	2.12
47	Mirex	5.28	4.91	4.11	4.39
48	1,3-Dichlorobenzene	3.60	3.65	3.49	3.36
49	1,2-Dimethylnaphthalene	4.31	4.38	4.24	4.43
50	2-Ethylnaphthalene	4.38	4.13	4.24	4.43
51	Cycloheptatriene	2.63	2.30	3.03	2.60
52	3-Chlorobiphenyl	4.58	4.50	4.70	4.50
53	3-Ethylthiophene	2.82	3.01	2.51	2.70
54	1,3,5-Tribromobenzene	4.51	4.44	4.25	4.46
55	β-Phenylethylchloride	2.95	2.93	3.08	3.46
56	Acenaphthene	3.92	4.22	4.34	4.28
57	*m*-Dibromobenzene	3.75	3.32	3.54	3.28
58	Dichlorodifluoromethane	2.16	2.09	2.01	2.09
59	Toluene	2.73	2.95	3.10	2.71
60	Anthracene	4.45	4.77	4.30	4.49
61	Hexachlorocyclopentadiene	5.04	5.05	4.61	4.87
62	3-Phenyl-1-chloropropane	3.55	2.95	3.52	3.61

Table 1 (*Continued*)

ID no.	Compound name	$\log P$ (experimental)	Estimated $\log P$		
			AP	TI 102	TI 202
63	Bibenzyl	4.79	4.28	4.47	4.63
64	1-Chloroheptane	4.15	4.53	4.29	4.86
65	2,4-Dichlorotoluene	4.24	3.92	3.71	3.92
66	1,1-Dichloroethane	1.79	1.94	1.59	2.02
67	β-Benzothiophene	3.12	3.33	2.91	2.63
68	2-Bromothiophene	2.75	2.93	2.68	2.58
69	Chlorodifluoromethane	1.08	1.10	1.39	1.10
70	Pentachlorobenzene	5.17	4.71	4.56	4.78
71	9,10-Dihydroanthracene	4.25	4.32	4.62	4.48
72	1,3-(Bis-chloromethyl)benzene	2.72	2.54	3.08	3.08
73	Chlorobenzene	2.84	2.78	2.59	2.40
74	1,2,4-Trichlorobenzene	4.02	4.17	3.78	4.03
75	2,2′,6-pcb	5.48	5.11	5.24	5.12
76	2-Butyne	1.46	2.33	2.27	2.24
77	Azulene	3.20	3.70	2.71	3.87
78	Trifluoromethylthiobenzene	3.57	3.19	3.64	3.94
79	2,5-pcb	5.16	4.96	5.18	5.06
80	1,2,3-Trichlorocyclohexene	2.84	2.61	4.33	4.27
81	Biphenyl	4.09	3.88	4.02	4.30
82	*p*-Xylene	3.15	3.56	3.25	3.31
84	Thiophenol	2.52	2.99	2.70	2.56
85	Bromotrifluoromethane	1.86	1.42	2.11	2.35
86	9-Methylanthracene	5.07	4.46	4.32	4.42
87	Trichloroethylene	2.42	2.77	2.98	3.08
88	1,4-Dimethyltetrachlorocyclohexane	4.40	3.62	3.74	3.62
89	Propylene	1.77	1.68	2.24	2.35
90	Cyclohexene	2.86	2.82	2.24	2.39
91	Methylthiobenzene	2.74	3.18	2.82	2.82
93	γ-Phenylpropyliodide	3.90	3.25	3.41	3.41
94	2,3,4′-pcb	5.42	5.21	5.24	5.29
95	Fluoropentachlorocyclohexane	3.19	3.66	3.91	3.66
96	1,2,3,5-Tetrachlorobenzene	4.92	4.73	4.33	4.35
97	2,2′-pcb	4.90	5.05	5.19	5.20
98	1-Butene	2.40	2.59	2.27	2.29
99	1,3-Dimethylnaphthalene	4.42	4.38	4.21	4.38
100	1,7-Dimethylnaphthalene	4.44	4.36	4.20	4.37
101	1-Methylnaphthalene	3.87	4.38	4.24	4.14
102	2,6-pcb	4.93	5.19	5.18	5.03
103	α-Bromotoluene	2.92	3.12	2.66	2.87
104	2,2′,3′-Trichlorobiphenyl	5.31	5.19	5.31	5.25
105	Hexafluorobenzene	2.22	3.82	4.24	4.56
106	3-Bromothiophene	2.62	2.55	2.70	2.65
107	1,2,3,5-Tetramethylbenzene	4.17	4.06	4.11	3.89
108	Halothane	2.30	1.52	2.99	2.89
109	2,4,6-pcb	5.47	5.12	5.08	5.18
110	1,1-Dichloroethylene	2.13	2.91	1.26	1.88
111	*o*-Dibromobenzene	3.64	3.37	3.81	3.34
112	1,2,4,5-Tetramethylbenzene	4.00	4.14	3.56	3.77
113	1-Hexene	3.39	4.28	2.97	3.16
114	Neopentane	3.11	3.29	3.05	3.16
115	Chloroform	1.97	1.67	1.09	1.96
116	1-Fluorobutane	2.58	2.31	2.56	2.31
117	Pyrene	4.88	4.46	4.66	4.46
118	1,1-Dichloro-2,2-diphenylethane	4.51	4.51	5.10	5.19
119	Isobutylene	2.34	2.07	2.38	2.33
120	Diphenylmethane	4.14	4.62	4.68	4.63
121	Isopropylbenzene	3.66	3.34	3.48	3.63
122	Naphthalene	3.30	3.65	3.88	3.78
123	1-Heptene	3.99	3.98	3.43	3.98
124	2,2-Dimethylbutane	3.82	3.00	3.33	3.61
125	1-Fluoropentane	2.33	2.98	2.76	2.99

Table 1 (*Continued*)

ID no.	Compound name	log *P* (experimental)	Estimated log *P*		
			AP	TI 102	TI 202
126	*o*-Xylene	3.12	3.60	3.34	3.43
127	Ethylbenzene	3.15	3.23	2.94	3.08
128	Trichloromethylthiobenzene	3.78	4.07	3.97	4.07
129	Thiophene	1.81	2.69	2.50	2.34
130	Bromochloromethane	1.41	1.52	1.00	1.41
131	1,2-Dichlorotetrafluoroethane	2.82	2.08	2.55	2.68
132	2-Chlorobiphenyl	4.38	4.80	4.95	4.87
133	2,4′-Dichlorobiphenyl	5.10	5.22	5.20	5.29
134	1,3,5-Trichlorobenzene	4.15	4.26	4.04	4.17
135	1-Octene	4.57	4.57	4.43	4.57
137	Phenylethylsulfide	3.20	3.23	3.20	3.02
138	1-Ethyl-2-methylbenzene	3.53	3.14	3.67	3.80
139	Propylbenzene	3.72	3.71	3.55	3.09
140	Indane	3.18	3.25	2.97	2.58
141	2-Chloropropane	1.90	2.28	1.77	2.28
142	Phenylazide	2.59	2.99	2.67	3.09
143	2,4-Dibromotetrachlorocyclohexane	3.98	3.66	3.47	3.61
144	Tetrachloroethylene	3.40	3.60	3.52	3.59
145	1-Nonene	5.15	4.28	4.36	4.36
146	2,3-Dimethylbutane	3.85	3.29	3.50	3.70
147	Dichlorofluoromethane	1.55	1.44	1.49	0.86
148	1,1,2,2-Tetrachloroethane	2.39	2.69	4.04	4.31
149	1,2,4-Trimethylbenzene	3.78	3.58	3.79	3.60
150	Fluorobenzene	2.27	2.99	2.78	2.68
151	Butylbenzene	4.26	3.34	3.41	3.25
152	Ethylbromide	1.61	1.77	1.82	1.79
153	Tetrafluoromethane	1.18	1.76	2.01	1.76
154	*p*-Cymene	4.10	3.41	3.56	3.69
155	*p*-Chlorotoluene	3.33	4.03	2.92	2.95
156	1-Bromopropane	2.10	2.40	2.42	2.27
157	Bromocyclohexane	3.20	3.62	2.46	3.16
158	2-Methylthiophene	2.33	2.65	2.68	2.58
159	Diphenylsulfide	4.45	4.28	4.58	4.47
160	1,2,4,5-Tetrachlorobenzene	4.82	4.78	4.15	4.40
161	1,1,1-Trichloroethane	2.49	2.68	3.05	2.35
162	*p*-Dichlorobenzene	3.52	3.68	4.25	4.49
163	1-Bromobutane	2.75	2.74	2.76	2.74
164	*p*-Chlorobiphenyl	4.61	4.84	4.69	4.84
165	Cyclopropylbenzene	3.27	3.34	2.61	2.34
166	2,6-Dichlorotoluene	4.29	4.15	4.54	4.15
167	Allene	1.45	3.05	2.03	2.06
168	β-Phenylethylbromide	3.09	3.32	3.23	3.46
169	1,3-Butadiene	1.99	2.13	2.07	1.97
170	2-Chlorothiophene	2.54	2.54	2.57	2.69
171	1-Bromopentane	3.37	3.28	2.96	3.28
172	γ-Phenylpropylbromide	3.72	3.02	3.47	3.73
173	1,3-Cyclohexadiene	2.47	2.47	2.60	2.58
174	Pentamethylbenzene	4.56	4.14	3.98	4.14
175	*p*-Dibromobenzene	3.79	3.70	4.06	4.08
176	1,4-Pentadiene	2.48	2.10	2.84	2.69
178	1,1-Difluoroethane	0.75	0.86	1.72	0.94
179	1-Bromohexane	3.80	3.87	4.40	3.87
180	*m*-Xylene	3.20	3.53	3.17	3.14
181	Dibenzothiophene	4.38	4.32	4.14	4.05
182	Ethyliodide	2.00	1.90	1.69	1.59
183	Trifluoromethylbenzene	3.01	3.47	4.06	2.86
184	2,3,6-Trimethylnaphthalene	4.73	4.36	4.39	4.43
185	Difluoromethane	0.20	0.94	1.10	0.86
186	1,2,4-Trifluorobenzene	2.52	2.52	3.57	3.03
187	Bromobenzene	2.99	3.09	2.87	3.09
188	Hexachloro-1,3-butadiene	4.78	5.24	3.98	4.29

Table 1 (*Continued*)

ID no.	Compound name	log *P* (experimental)	Estimated log *P*		
			AP	TI 102	TI 202
189	Vinylbromide	1.57	1.78	1.26	1.70
190	*o*-Chlorotoluene	3.42	3.25	3.71	3.55
191	α-Chlorotoluene	2.30	2.94	3.00	3.04
192	1,4-Cyclohexadiene	2.30	2.82	2.83	2.67
193	1-Bromoheptane	4.36	4.35	4.22	4.35
194	Styrene	2.95	3.24	2.72	2.94
195	Chlorotrifluoromethane	1.65	2.01	2.18	2.01
196	(Dimethyl)phenylphosphine	2.57	3.89	3.35	3.42
197	Cycloocta-1,5-diene	3.16	2.58	2.89	2.87
198	Tetrachlorocyclohexane	2.82	4.22	4.50	4.19
199	1-Bromooctane	4.89	4.08	4.40	4.53
200	2-Methylnaphthalene	3.86	4.41	4.24	4.15
201	3-Methylthiophene	2.34	2.72	2.67	2.58
202	Methylenechloride	1.25	1.46	1.19	1.42
203	Hexachlorobenzene	5.31	4.98	4.15	5.12
204	Indene	2.92	3.18	3.12	3.20
205	*Tert*-butylbenzene	4.11	3.12	3.86	3.88
206	1,2-Dichloroethane	1.48	2.02	1.91	1.89
207	1,3,5-Trimethylbenzene	3.42	3.69	3.42	3.58
208	Phenanthrene	4.46	4.67	4.52	4.98
209	Benzene	2.13	2.79	2.90	2.65
210	3,3,3-Trifluoropropylbenzene	3.31	2.95	3.73	4.07
211	α-(2,2,2-Trichloroethyl)styrene	4.56	3.37	4.07	3.68
212	2,3-Dimethylnaphthalene	4.40	4.38	4.35	4.35
213	1,3-Dichloropropane	2.00	2.06	2.52	1.72
214	1,2,3,4-Tetramethylbenzene	4.11	4.09	4.13	4.37
215	Stilbene-*t*	4.81	4.43	4.46	4.47
216	Fluorene	4.18	4.15	4.25	4.15
217	2-Fluoro-3-bromotetrachlorocyclohexane	3.28	3.59	3.88	3.62
218	Allylbenzene	3.23	2.94	3.06	3.34
219	Carbontetrachloride	2.83	2.51	2.71	2.51

of the TIs were calculated using POLLY 2.3 [36], while the remaining four *J* indices were calculated using other in-house software. More information on the set of topological indices calculated by POLLY has been reported in earlier studies [11,37,38]. One-hundred additional indices, the real-number local vertex invariants (LOVIs) [39], were added to the set of topological indices to form the expanded set of 202 indices used in this study.

2.4. Calculation of atom pairs

Atom pairs (APs) were calculated using the method of Carhart et al. [40]. In this method, an atom pair is defined as a substructure composed of two non-hydrogen atoms, *i* and *j*, and their interatomic separation:

$$\langle \text{atom descriptor}_i \rangle – \langle \text{separation} \rangle – \langle \text{atom descriptor}_j \rangle$$

where ⟨atom descriptor⟩ contains information regarding the atom type, number of non-hydrogen neighbors, and the number of π electrons. Interatomic separation is defined as the number of atoms traversed in the shortest bond-by-bond path containing both atoms. The atom pairs were calculated using APProbe [41]. This program also generated the intermolecular similarity scores for both sets of compounds as detailed later in the section on similarity measures.

2.5. Data reduction

Initially, all TIs were transformed by the natural logarithm of the index plus one. Some TIs may be several orders of magnitude greater than others, so the scaling is conducted to minimize the effect of scale.

A principal component analysis (PCA) was used on the transformed indices to minimize the intercorrelation of indices. The PCA was conducted using the SAS procedure PRINCOMP [42]. Only PCs with eigenvalues greater than or equal to one have been retained for this study. A more detailed explanation of this approach has been provided in a previous study by Basak et al. [16]. These PCs were subsequently used as independent variables (in place of the TIs) to determine similarity scores in the Euclidean distance method described below.

2.6. Similarity measures

Intermolecular similarity was measured in three similarity spaces. The first two spaces used Euclidean distance (ED) within an n-dimensional PC space derived from the two sets of TIs to measure intermolecular similarity. ED between molecules i and j is defined as

$$\mathrm{ED}_{ij} = \left[\sum_{k=1}^{n} (D_{ik} - D_{jk})^2 \right]^{1/2} \tag{1}$$

where n equals the number of dimensions or PCs retained from the PCA. D_{ik} and D_{jk} are the data values of the kth dimension for molecules i and j, respectively.

The third space was created using the AP method, an associative measure described by Carhart et al. [40] based on atom pair descriptors. The measurement is the ratio of the number of shared atom pairs between two molecules over the total number of atom pairs present in the two molecules. Similarity (S) between molecules i and j is defined as:

$$S_{ij} = \frac{2C}{T_i + T_j} \tag{2}$$

where C is the number of atom pairs common to molecule i and j. T_i and T_j are the total number of atom pairs in molecule i and j, respectively. The numerator is multiplied by a factor of 2 to reflect the presence of shared atom pairs in both compounds. The similarity scores (S) were calculated for both sets of chemicals using APProbe [41].

2.7. Analog/K-nearest neighbor selection

Following the quantification of the intermolecular similarity of the molecules, analogs or nearest neighbors are determined on the basis of both ED and S. The ED method measures a distance between molecules. Thus, the lower the value of ED, the greater the similarity between two molecules. In the case of the AP method, two molecules are considered identical if $S = 1$, while they have no atom pairs in common if $S = 0$.

As mentioned earlier, many times chemical analogs are used in hazard and risk assessment, following the axiom that similar chemicals behave similarly. Many times this analog selection is handled by an expert; however, computerized methods speed up the process and create results that can be easily duplicated.

2.8. Property estimation

Property estimation was carried out using the K-nearest neighbor (KNN) method. For each compound, a number of similar chemicals ($K = 1$–25) are selected and the property of interest is estimated based on the values of these nearest neighbors. For instance, in estimating the $\log_{10}(p_{\mathrm{vap}})$ of the probe compound, the mean of the $\log_{10}(p_{\mathrm{vap}})$ for the

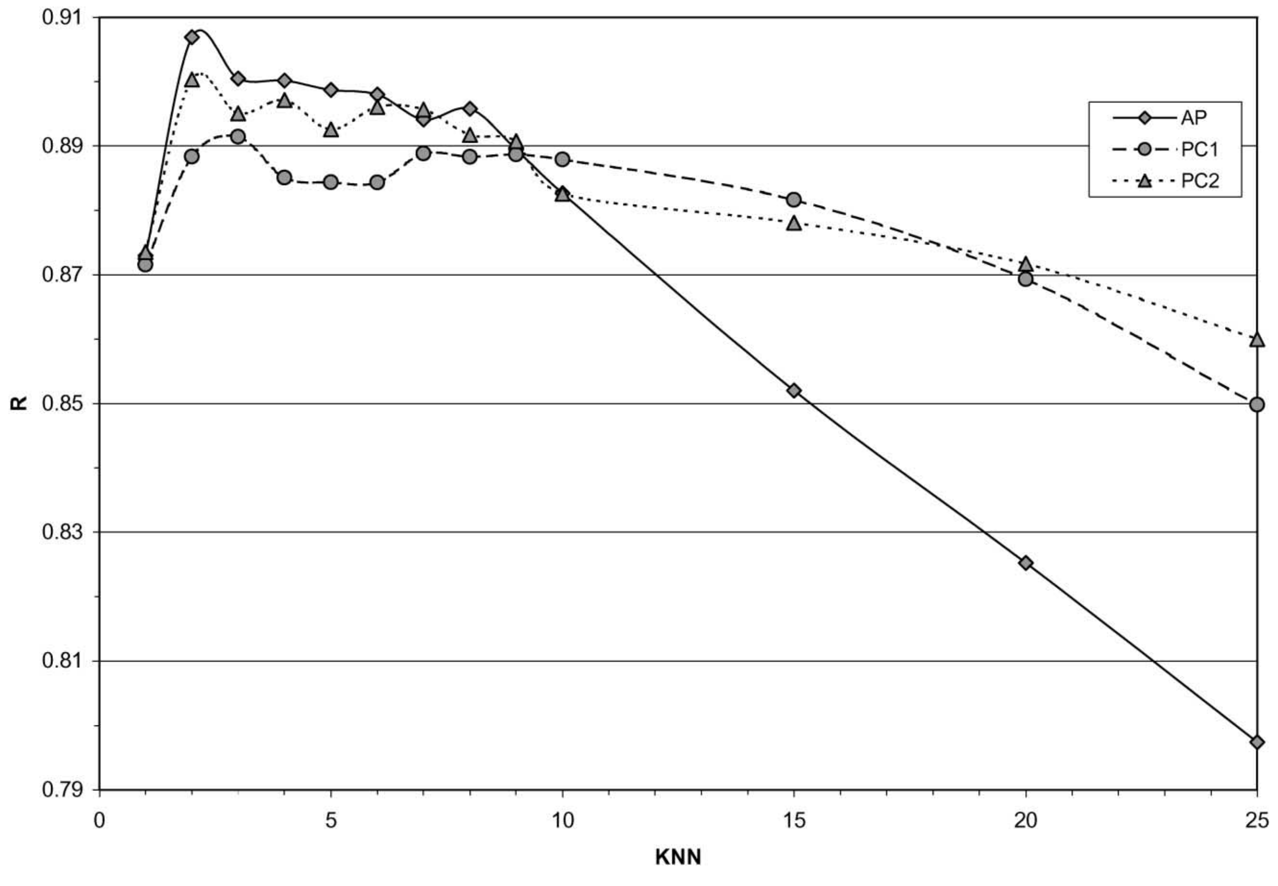

Fig. 1. Pattern of correlation, R, for the KNN estimation ($K = 1$–10, 15, 20, 25) of $\log P$ for 213 STARLIST chemicals.

K-nearest neighbors was used as the estimate. KNN estimation was carried out for all chemicals in both of the datasets, resulting in a full cross-validation. Thus, the correlation coefficients reported are the cross-validated correlation coefficients.

3. Results

3.1. Euclidean distance method

After screening for intercorrelations and indices with 0 values, a total of 99 TIs were retained for the 213 STARLIST compounds, and 96 TIs were retained for the set of 469 compounds. The PCA was carried out and PCs with eigenvalues greater than or equal to 1.0 were retained. In this manner, no PC is discarded which accounts for one full independent variable. This resulted in the retention of 10 PCs for the set of 213 compounds and 13 PCs for set of 469 compounds. The 10 PCs explained 94.6% of the variance within the STARLIST set of 213 chemicals and the 13 PCs explained 93.3% of the variance in the TSCA vapor pressure set.

With the addition of the LOVIs, PCA was again conducted to extract a small number of orthogonal PCs. This time, the combined set of TIs and LOVIs for the STARLIST set contained 99 TIs and 76 LOVIs for a total of 175 parameters. From the PCA, 12 PCs were retained with a total explained variance of 96.8%. Similarly, the descriptor set for the TSCA vapor pressure set was composed of the 96 TIs used earlier and 76 LOVIs for a total of 172 parameters. The PCA resulted in the retention of 13 PCs explaining a total of 95.1% of the variance in the data.

Table 2
Symbols and definitions of topological parameters

Index	Definition
$I_{\mathrm{D}}^{\mathrm{W}}$	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\bar{I}_{\mathrm{D}}^{\mathrm{W}}$	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
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
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
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
${}^{h}\chi$	Path connectivity index of order $h = 0$–6
${}^{h}\chi$	Cluster connectivity index of order $h = 3$–6
${}^{h}\chi_{\mathrm{Ch}}$	Chain connectivity index of order $h = 3$–6
${}^{h}\chi_{\mathrm{PC}}$	Path–cluster connectivity index of order $h = 4$–6
${}^{h}\chi^{\mathrm{b}}$	Bond path connectivity index of order $h = 0$–6
${}^{h}\chi_{\mathrm{C}}^{\mathrm{b}}$	Bond cluster connectivity index of order $h = 3$–6
${}^{h}\chi_{\mathrm{Ch}}^{\mathrm{b}}$	Bond chain connectivity index of order $h = 3$–6
${}^{h}\chi_{\mathrm{PC}}^{\mathrm{b}}$	Bond path–cluster connectivity index of order $h = 4$–6
${}^{h}\chi^{\mathrm{v}}$	Valence path connectivity index of order $h = 0$–6
${}^{h}\chi_{\mathrm{C}}^{\mathrm{v}}$	Valence cluster connectivity index of order $h = 3$–6
${}^{h}\chi_{\mathrm{Ch}}^{\mathrm{v}}$	Valence chain connectivity index of order $h = 3$–6
${}^{h}\chi_{\mathrm{PC}}^{\mathrm{v}}$	Valence path–cluster connectivity index of order $h = 4$–6
P_h	Number of paths of length $h = 0$–10
J	Balaban's J index based on distance
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
Triplet	Global invariants based on solutions of linear equation systems using the adjacency matrix (A), distance matrix (D), and column/row vectors: distance sums (S), atomic number (Z), number of non-hydrogen atoms (N and N^2), vertex degree (V), or numerical constants (1). Notation is described by triplets (e.g. AZV). Results are weightings for each atom in a molecule. These weights are combined by five possible formulas: 1 = Sum of weights:$\sum_i x_i$ 2 = Sum of squared weights $\sum_i x_i^2$ 3 = Sum of square root of weights $\sum_i x_i^{1/2}$ 4 = Sum of cross-product $\sum_i (x_i \times x_j)^{-1/2}$ 5 = Product of weights $N[\sum_i x_i]^{1/N}$

3.2. Comparison of methods

For the purpose of brevity and clarity, the models based on the initial set of 102 TIs will be referred to as PC1 and the models based on the set of TIs and LOVIs will be referred to as PC2. Fig. 1 presents the correlation coefficients between the KNN based estimated values and the calculated log P values for the diverse set of 213 STARLIST chemicals listed in Tables 1 and 2 for the three similarity methods. Fig. 2 shows the pattern of standard error of estimates for the same set. Fig. 1 demonstrates that the AP estimation method is marginally superior to the two methods based on topological indices and that there is some minimal improvement with the addition of the LOVIs in PC2. Figs. 3 and 4 present the patterns of correlation and standard errors of estimation of $\log_{10}(p_{\mathrm{vap}})$ for the set of 469 TSCA chemicals. In this set, PC1 estimation does significantly better than the AP method and slightly outperforms PC2.

Another important consideration is that the estimation from neighborhood means results in a smoothing of the data, an over-all loss of variance in our property of interest (dependent variable). The decrease in data variance in log P and $\log_{10}(p_{\mathrm{vap}})$ is summarized in Tables 3 and 4 and illustrated in Figs. 5 and 6, respectively. In both cases and for all models, it is clearly demonstrated that data variance rapidly decreases with increasing K.

4. Discussion

The goal of this paper was to evaluate the relative effectiveness of three QMSA methods in estimating log P and normal vapor pressure for two structurally diverse sets of chemicals. Of the two classes of methods, one based on atom pairs (substructures) and the other based on topological indices (mathematical structural invariants which are real numbers), neither emerged as a distinct method of choice. While the AP-based method was superior to the TI-based method for the log P dataset, the TI method was better than the AP method for estimation of normal vapor pressure. In several of our earlier papers [18,47], we have compared QMSA methods based on physicochemical properties versus calculated molecular descriptors for analog selection. The earlier study compared the degree of overlap in analog selection between a physicochemical-based QMSA and four other QMSA methods based on APs and TIs. The latter study compared QMSA models based on physicochemical properties with APs and TIs for the estimation of log P for two sets of compounds. It would be of interest to see how an expanded set of QMSA methods works in the estimation of properties such as log P and normal vapor pressure.

In hazard assessment, the two widely used methods are: (a) class-specific QSAR; and (b) analog-based hazard evaluation. Neither of the two datasets analyzed here fell into

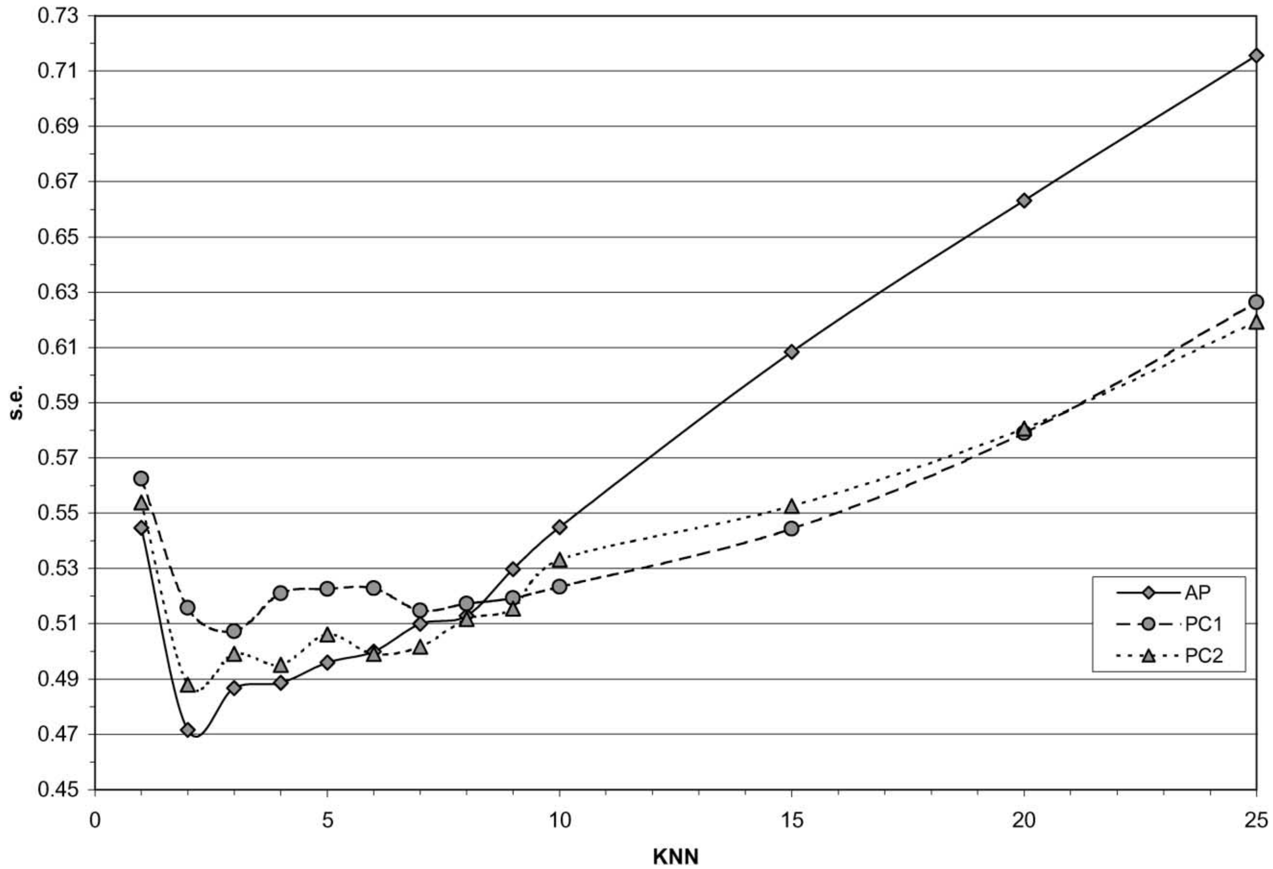

Fig. 2. Pattern of standard error, S.E., for the KNN estimation ($K = 1$–10, 15, 20, 25) of log P for 213 STARLIST chemicals.

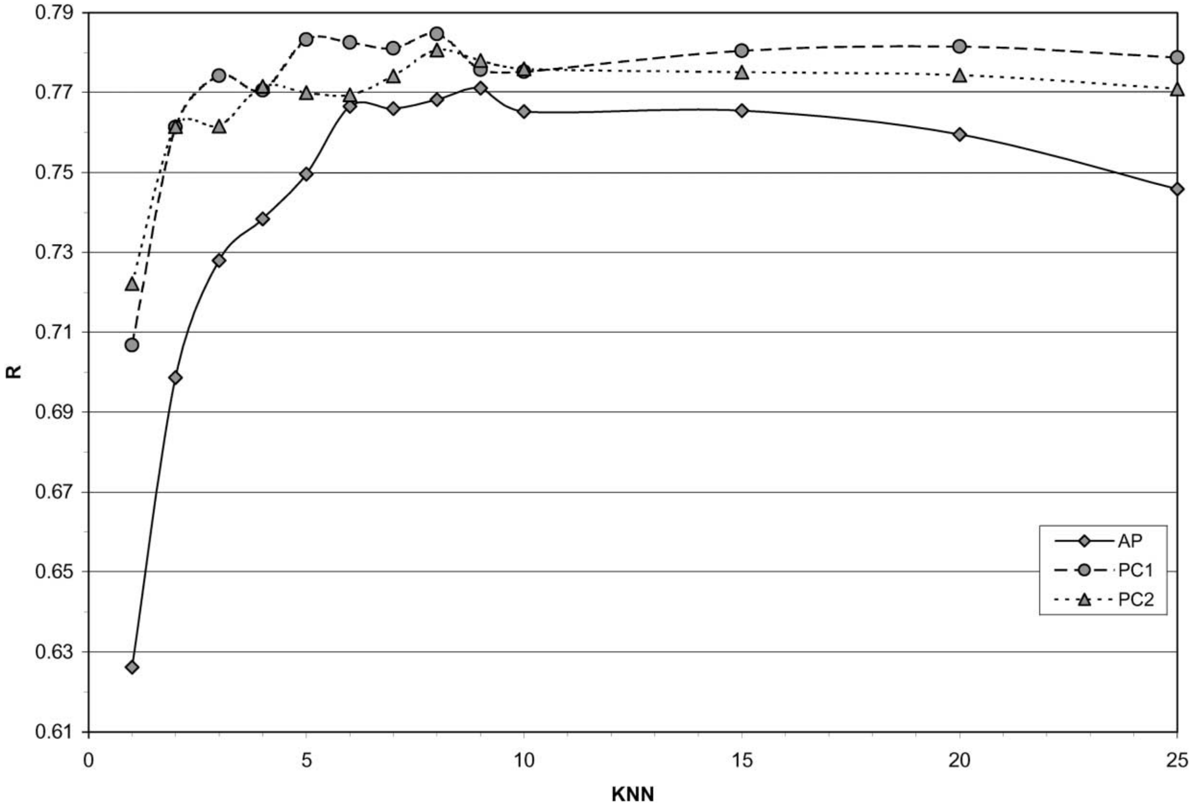

Fig. 3. Pattern of correlation, *R*, for the KNN estimation ($K = 1$–10, 15, 20, 25) of normal vapor pressure ($\log_{10}(p_{\text{vap}})$) for 469 TSCA chemicals.

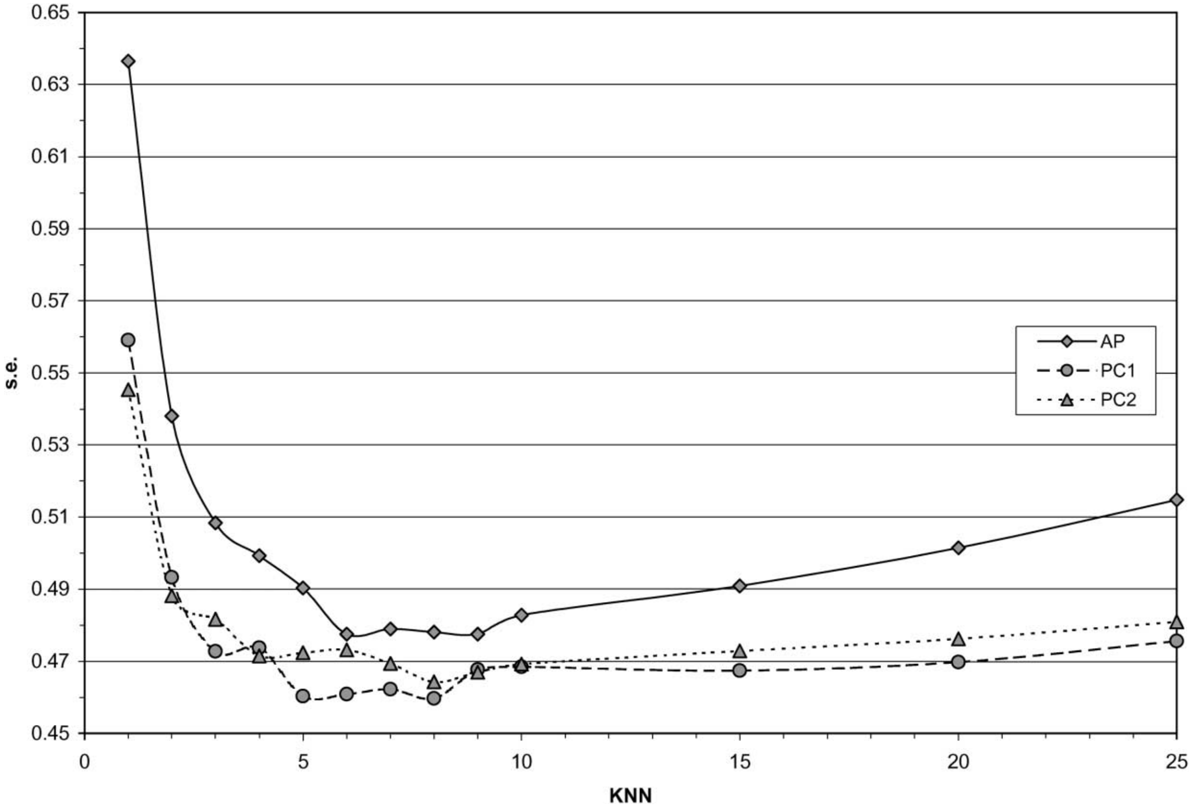

Fig. 4. Pattern of standard error, S.E., for the KNN estimation ($K = 1$–10, 15, 20, 25) of normal vapor pressure ($\log_{10}(p_{\text{vap}})$) for 469 TSCA chemicals.

Table 3
Variance of the dataset, before and after smoothing for the STARLIST set[a]

KNN	AP	AP_{SC}	PC1	$PC1_{SC}$	PC2	$PC2_{SC}$
1	1.1574	0.92	1.2089	0.97	1.1633	0.93
2	1.0104	0.81	1.0781	0.86	1.0795	0.86
3	0.9987	0.80	1.0114	0.81	0.9881	0.79
4	0.9420	0.75	0.9632	0.77	0.9515	0.76
5	0.8733	0.70	0.9522	0.76	0.9208	0.74
6	0.8442	0.67	0.9425	0.75	0.9118	0.73
7	0.8212	0.66	0.9018	0.72	0.8823	0.70
8	0.7700	0.61	0.8765	0.70	0.8559	0.68
9	0.7350	0.59	0.8381	0.67	0.8340	0.67
10	0.7147	0.57	0.8097	0.65	0.8119	0.65
15	0.6232	0.50	0.7304	0.58	0.7158	0.57
20	0.5224	0.42	0.6426	0.51	0.6156	0.49
25	0.4297	0.34	0.5478	0.44	0.5223	0.42

[a] The absolute variance in the data is 1.2524 and the second column of each pair shows the variance rescaled between 0 and 1.

Table 4
Variance of the dataset, before and after smoothing for the TSCA set[a]

KNN	AP	AP_{SC}	PC1	$PC1_{SC}$	PC2	$PC2_{SC}$
1	0.5356	0.98	0.5218	0.95	0.5169	0.94
2	0.3745	0.68	0.4268	0.78	0.4584	0.84
3	0.3226	0.59	0.3675	0.67	0.4046	0.74
4	0.3081	0.56	0.3524	0.64	0.3758	0.69
5	0.2882	0.53	0.3253	0.59	0.3464	0.63
6	0.2726	0.50	0.3049	0.56	0.3338	0.61
7	0.2627	0.48	0.2977	0.54	0.3294	0.60
8	0.2555	0.47	0.2902	0.53	0.3080	0.56
9	0.2442	0.45	0.2851	0.52	0.3045	0.55
10	0.2393	0.44	0.2793	0.51	0.2983	0.54
15	0.2007	0.37	0.2542	0.46	0.2630	0.48
20	0.1785	0.33	0.2374	0.43	0.2428	0.44
25	0.1635	0.30	0.2262	0.41	0.2245	0.41

[a] The absolute variance in the data is 0.5480 and the second column of each pair shows the variance rescaled between 0 and 1.

any structural or biological class. But our research group has developed hierarchical QSARs for both sets of these structurally diverse chemicals. Tables 5 and 6 summarize the results of hierarchical QSAR vis-à-vis QMSA-based (three methods) estimation of $\log P$ and normal vapor pressure, respectively. Data on Table 6 shows that the best predictions for $\log_{10}(p_{\mathrm{vap}})$ from the three similarity methods are very close, whereas the result from the hierarchical QSAR is much superior to those derived from QMSA techniques.

In the case of the $\log P$ data (Table 5), however, the predictions obtained from the three QMSA methods are only slightly inferior to those estimated by hierarchical QSARs. For details on the hierarchical QSAR method and results from a number of our studies, see our recent book chapter reviewing the method [48].

Many times when conducting practical hazard assessment, one has to do quick estimation of a large number of relevant properties. Similarity methods reported in this paper might

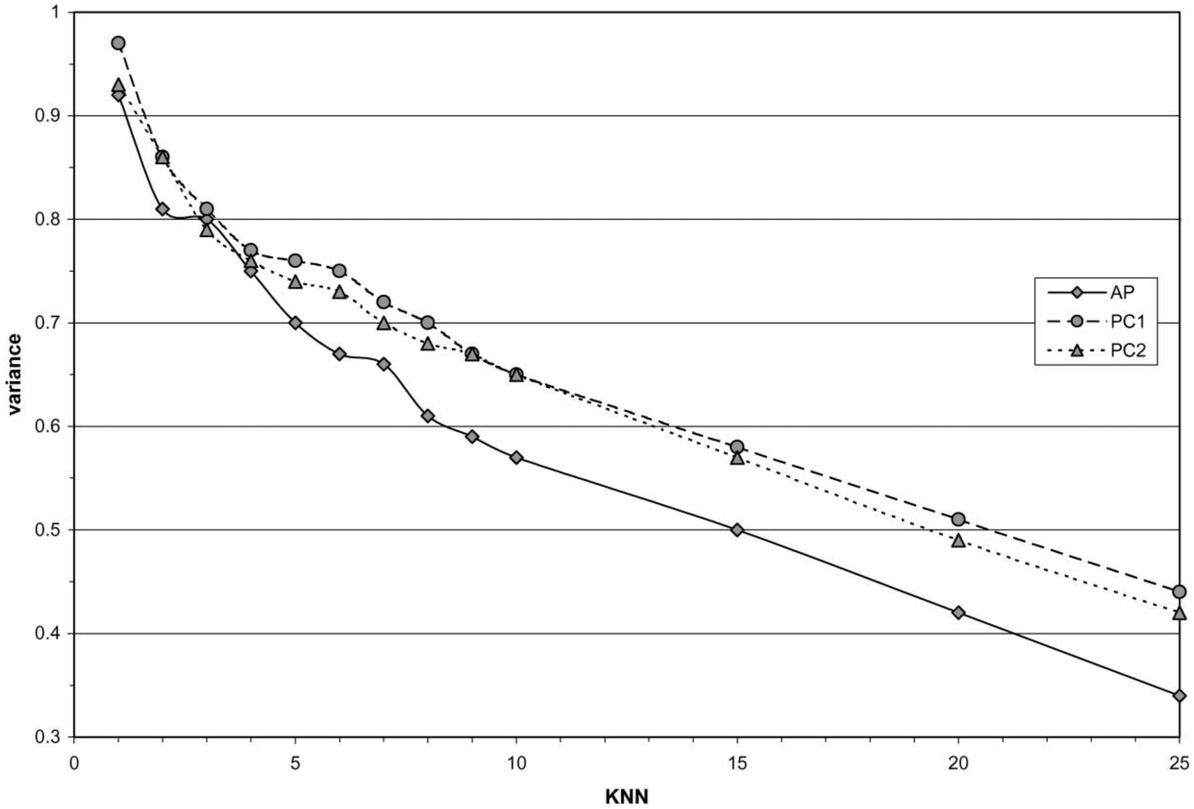

Fig. 5. Pattern of smoothing of data variance as a consequence of the KNN approach for the STARLIST $\log P$ data.

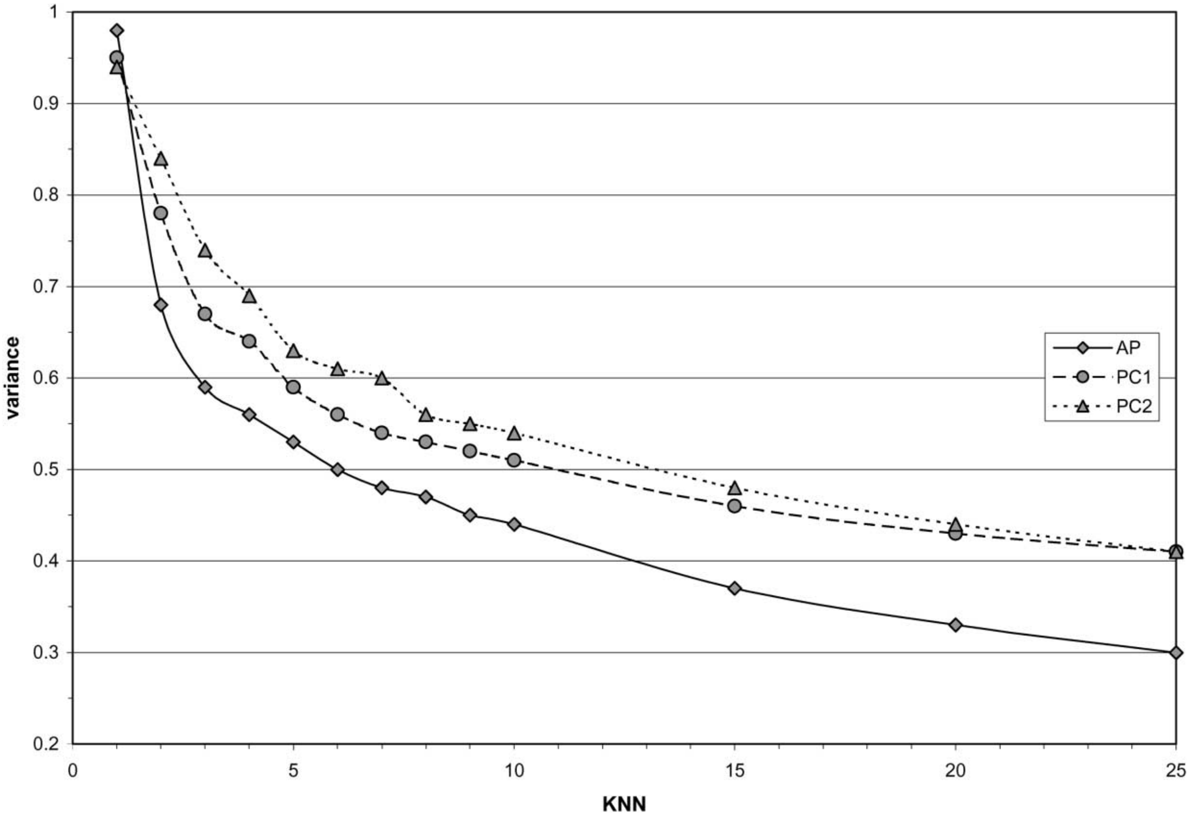

Fig. 6. Pattern of smoothing of data variance as a consequence of the KNN approach for the TSCA normal vapor pressure data.

Table 5
Model results for each of the three similarity methods and results from an earlier QSAR modeling study of the STARLIST log P data

Method	N	R	S.E.
Atom pair			
$K = 1$	213	0.8730	0.555
$K = 2$	213	0.9069	0.472
$K = 3$	213	0.9005	0.487
$K = 4$	213	0.9002	0.489
$K = 5$	213	0.8987	0.496
PC1			
$K = 1$	213	0.8716	0.562
$K = 2$	213	0.8884	0.516
$K = 3$	213	0.8914	0.507
$K = 4$	213	0.8851	0.521
$K = 5$	213	0.8844	0.523
PC2			
$K = 1$	213	0.8736	0.554
$K = 2$	213	0.9003	0.488
$K = 3$	213	0.8951	0.499
$K = 4$	213	0.8971	0.495
$K = 5$	213	0.8926	0.506
QSAR	219	0.9529	0.36

Table 6
Model results for each of the three similarity methods and results from an earlier QSAR modeling study of the TSCA normal vapor pressure data

Method	N	R	S.E.
Atom pair			
$K = 4$	469	0.7384	0.499
$K = 5$	469	0.7496	0.490
$K = 6$	469	0.7665	0.478
$K = 7$	469	0.7660	0.479
$K = 8$	469	0.7682	0.478
PC1			
$K = 6$	469	0.7694	0.473
$K = 7$	469	0.7741	0.469
$K = 8$	469	0.7806	0.464
$K = 9$	469	0.7780	0.467
$K = 10$	469	0.7760	0.469
PC2			
$K = 3$	469	0.7742	0.473
$K = 4$	469	0.7705	0.474
$K = 5$	469	0.7832	0.460
$K = 6$	469	0.7825	0.461
$K = 7$	469	0.7811	0.462
QSAR			
Training	342	0.8967	0.35
Test	134	0.9203	0.28

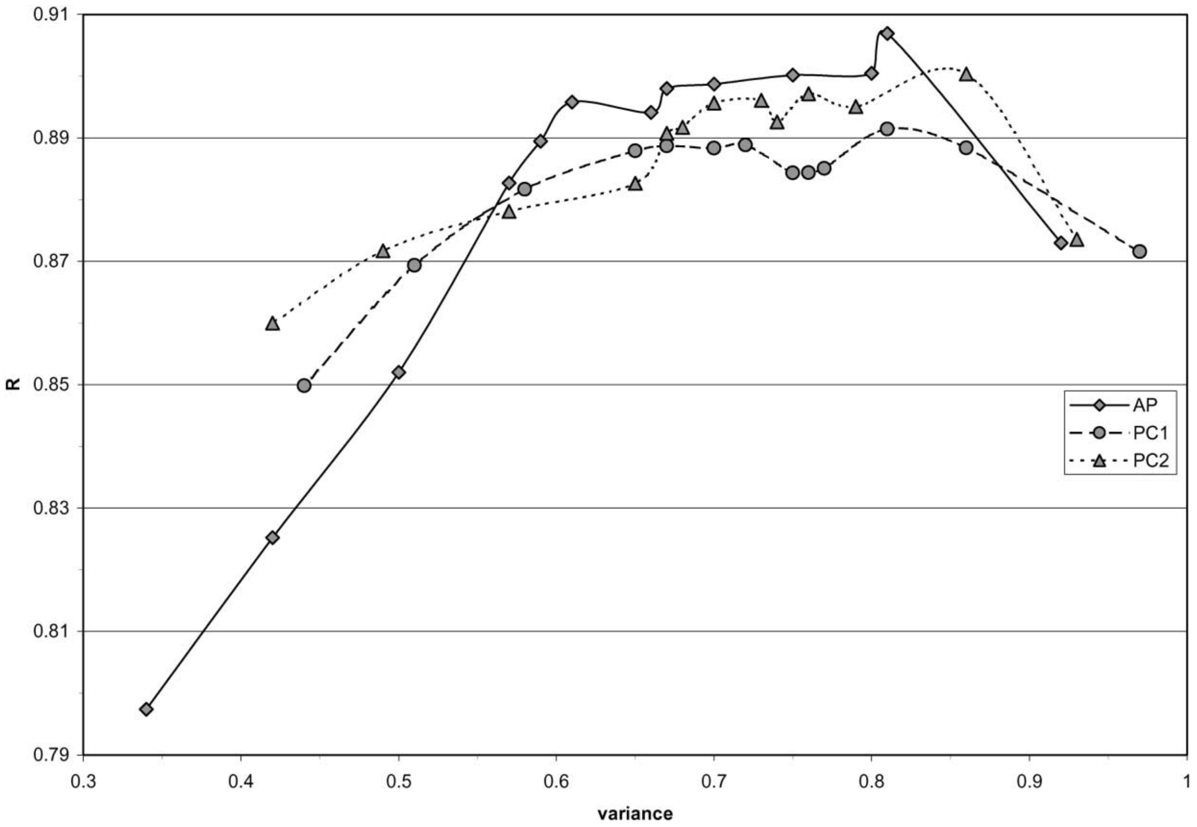

Fig. 7. Pattern of smoothing of data variance vs. the cross-validated correlation coefficient (R) for the STARLIST $\log P$ data.

give quick, first estimates for such properties. These methods can be developed and used with much less commitment of time and resources as compared to the hierarchical QSAR models.

Also, a similarity-based method using the KNN approach will be effective only when the structure space is such that chemicals are in close proximity to each other, i.e. one has a dense dataset. Neither of the two datasets analyzed in this paper are very large. On the other hand, they contain rather diverse structures that will make nearest neighbors not very close in any scale of distance. It will be interesting to see how the QSAR and QMSA methods compare with each other when the datasets used are more dense.

For both sets of chemicals analyzed in this paper, the nearest neighbor ($K = 1$) does not give the best estimate in any of the three methods. Initially, there is an increase in the quality of estimated property with increasing value of K. The correlation reaches a maximum around $K = 2$–8 for the STARLIST $\log P$ data and $K = 5$–9 for the TSCA vapor pressure data. Beyond 8–10 nearest neighbors, there is a progressive deterioration in the value of r. This is in line with our previous results where we found that $K = 5$–10 gave the best results in the characterization of neighborhoods for the prediction of properties in datasets ranging from small and structurally related sets of congeners to diverse sets consisting of nearly 3000 TSCA chemicals [5–8,10,43–46]. Additionally, as was demonstrated in Figs. 5 and 6, decreasing data variance should also be a concern when selecting the "optimal" number of neighbors for property estimation. For the STARLIST data, variance does not drop below 50% until $K = 15$ for the AP method and around $K = 20$ for PC1 and PC2. However, decreasing data variance is much more evident in the TSCA vapor pressure set where data variance has decreased to 50% at $K = 6$ for the AP method, approximately 11 for PC1, and around 13 for PC2.

So how many neighbors should be used? The best approach to this question is to look at estimation results (R) versus decreasing variance. Figs. 7 and 8 show the trends and the potential hazards of picking the value for K with the highest correlation coefficient. Ideally, the selection of data ranges would come from the upper right-hand quadrant of the graph. In this way, a maximum amount of data variance is retained while also retaining a high correlation coefficient. In the instance of the STARLIST data (Fig. 7), this concept works. If we choose $K = 1$ for the AP method or $K = 1$ or 2 for PC1 and PC2, the data variance remains above 85% of the original variance. However, this becomes a larger issue with the TSCA vapor pressure data (Fig. 8). In this case, our best selection is the PC2 method with $K = 2$, a reasonable value for R (0.7615) and the data variance is at 84% of the original. The highest values for the correlation coefficient fall in a range where the data variance has

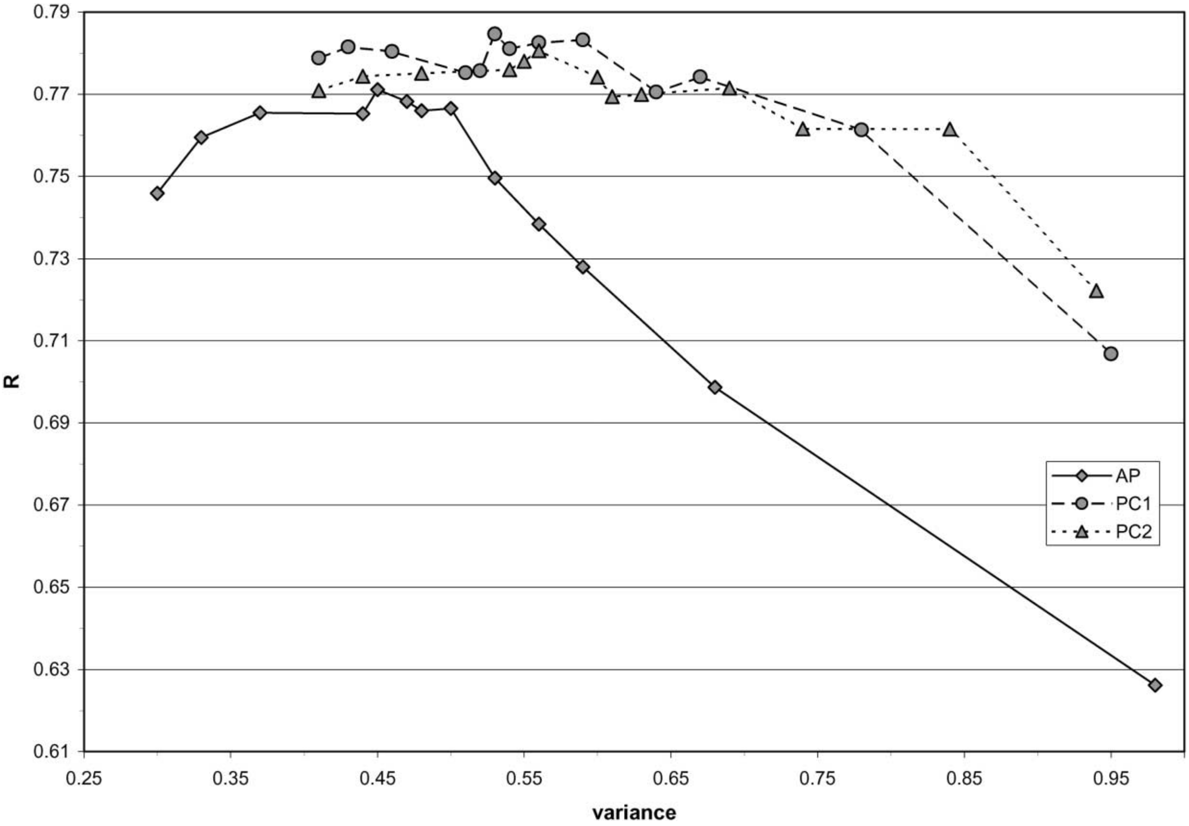

Fig. 8. Pattern of smoothing of data variance vs. the cross-validated correlation coefficient (*R*) for the TSCA normal vapor pressure data.

already been reduced to between 55 and 65% of its original value. Thus, while we would be most likely to choose *K* between 5 and 10 nearest neighbors for this set, realistically we should restrict ourselves to much lower values of *K*.

Our research has shown that, in general, one should use physicochemical properties, as compared to calculated TIs or APs, for the best results in property estimation based on chemicals analogs [15,47]. But such an approach in hazard assessment is impractical at present since experimental physicochemical properties are not available for the majority of chemicals. Molecular similarity methods based on descriptors that can be calculated directly from molecular structure give reasonable estimates of properties.

Acknowledgements

This is contribution number 297 from the Center for Water and the Environment of the Natural Resources Research Institute. Research reported in this paper was supported by grant F49620-98-1-0015 from the United States Air Force. The authors would like to thank Denise Mills and Gregory Grunwald for their continued assistance.

References

[1] C. Auer, J.V. Nabholz, K.P. Baetcke, Mode of action and the assessment of chemical hazards in the presence of limited data: use of structure–activity relationships (SAR) under TSCA, Section 5, Environ. Health Perspect. 87 (1990) 183–197.

[2] S.C. Basak, G.D. Grunwald, G.E. Host, G.J. Niemi, S.P. Bradbury, A comparative study of molecular similarity, statistical and neural network methods for predicting toxic modes of action of chemicals, Environ. Toxicol. Chem. 17 (1998) 1056–1064.

[3] M. Johnson, G.M. Maggiora (Eds.), Concepts and Applications of Molecular Similarity, Wiley, New York, 1990.

[4] M. Johnson, S.C. Basak, G. Maggiora, A characterization of molecular similarity methods for a property prediction, Math. Comput. Model. 11 (1988) 630–634.

[5] S.C. Basak, G.D. Grunwald, Molecular similarity and risk assessment: analog selection and property estimation using graph invariants, SAR QSAR Environ. Res. 2 (1994) 289–307.

[6] S.C. Basak, G.D. Grunwald, Molecular similarity and estimation of molecular properties, J. Chem. Inform. Comput. Sci. 35 (1995) 366–372.

[7] S.C. Basak, G.D. Grunwald, Estimation of lipophilicity from molecular structural similarity, New J. Chem. 19 (1995) 231–237.

[8] S.C. Basak, G.D. Grunwald, Predicting mutagenicity of chemicals using topological and quantum chemical parameters: a similarity based study, Chemosphere 31 (1995) 2529–2546.

[9] S.C. Basak, B.D. Gute, Use of graph-theoretic parameters in predicting inhibition of microsomal *p*-hydroxylation of aniline by alcohols: a molecular similarity approach, in: B.L. Johnson, C. Xintaras, J.S. Andrews Jr. (Eds.), Hazardous Waste: Impacts on

Human and Ecological Health, Princeton Scientific Publishing Co., Inc., Princeton, NJ, 1997, pp. 492–504.

[10] S.C. Basak, B.D. Gute, G.D. Grunwald, Characterization of the molecular similarity of chemicals using topological invariants, in: R. Carbo-Dorca, P.G. Mezey (Eds.), Advances in Molecular Similarity, Vol. 2, JAI Press, Stanford, Connecticut, 1998, pp. 171–185.

[11] S.C. Basak, B.D. Gute, Use of graph invariants in QMSA and predictive toxicology, in: P. Hansen, P. Fowler, M. Zheng (Eds.), Discrete Mathematical Chemistry: DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Vol. 51, American Mathematical Society, Providence, Rhode Island, 2000, pp. 9–24.

[12] J.C. Arcos, Structure–activity relationships: criteria for predicting the carcinogenic activity of chemical compounds, Environ. Sci. Technol. 21 (1987) 743–745.

[13] J. Ashby, R.W. Tennant, Definitive relationships among chemical structure, carcinogeniocity, and mutagenicity for 301 chemicals tested by the U.S. NTP, Mutat. Res. 257 (1991) 229–306.

[14] D.M. Sanderson, C.G. Earnshaw, Computer prediction of possible toxic action from chemical structure: the DEREK system, Hum. Exp. Toxicol. 10 (1991) 261–273.

[15] S.C. Basak, G.D. Grunwald, Use of topological space and property space in selecting structural analogs, Math. Model. Sci. Comput. 4 (1994) 464–469.

[16] S.C. Basak, V.R. Magnuson, G.J. Niemi, R.R. Regal, Determining structural similarity of chemicals using graph-theoretic indices, Discrete Appl. Math. 19 (1988) 17–44.

[17] S.C. Basak, B.D. Gute, G.D. Grunwald, Development and application of molecular similarity methods: using nonempirical parameters, Math. Model Sci. Comput., 2000, in press.

[18] S.C. Basak, B.D. Gute, G.D. Grunwald, Quantitative comparison of five molecular structure spaces in selecting analogs of chemicals, Math. Model. Comput. Sci., 2000, in press.

[19] S.C. Basak, B.D. Gute, G.D. Grunwald, A comparative study of topological and geometrical parameters in estimating normal boiling point and octanol/water partition coefficient, J. Chem. Inform. Comput. Sci. 36 (1996) 1054–1060.

[20] A. Leo, D. Weininger, CLOGP Version 3.2 User Reference Manual, Medicinal Chemistry Project, Pomona College, Claremont, CA, 1984.

[21] G.J. Niemi, S.C. Basak, G.D. Veith, G.D. Grunwald, Prediction of octanol/water partition coefficient (K_{OW}) with algorithmically derived variables, Environ. Toxicol. Chem. 11 (1992) 893–898.

[22] S.C. Basak, B.D. Gute, G.D. Grunwald, Use of topostructural, topochemical, and geometric parameters in the prediction of vapor pressure: a hierarchical QSAR approach, J. Chem. Inform. Comput. Sci. 37 (1997) 651–655.

[23] C.L. Russom, E.B. Anderson, B.E. Greenwood, A. Pilli, ASTER: an integration of the AQUIRE data base and the QSAR system for use in ecological risk assessments, Sci. Total Environ. 109–110 (1991) 667–670.

[24] H. Wiener, Structural determination of paraffin boiling points, J. Am. Chem. Soc. 69 (1947) 17–20.

[25] M. Randić, On characterization of molecular branching, J. Am. Chem. Soc. 97 (1975) 6609–6615.

[26] L.B. Kier, W.J. Murray, M. Randić, L.H. Hall, Molecular connectivity V: connectivity series applied to density, J. Pharm. Sci. 65 (1975) 1226–1230.

[27] L.B. Kier, L.H. Hall, Molecular Connectivity in Structure–activity Analysis, Research Studies Press, Letchworth, Hertfordshire, UK, 1986.

[28] D. Bonchev, N. Trinajstić, Information theory, distance matrix and molecular branching, J. Chem. Phys. 67 (1977) 4517–4533.

[29] C. Raychaudhury, S.K. Ray, J.J. Ghosh, A.B. Roy, S.C. Basak, Discrimination of isomeric structures using information theoretic topological indices, J. Comput. Chem. 5 (1984) 581–588.

[30] S.C. Basak, A.B. Roy, J.J. Ghosh, Study of the structure-function relationship of pharmacological and toxicological agents using information theory, in: X.J.R. Avula, R. Bellman, Y.L. Luke, A.K. Rigler (Eds.), Proceedings of the 2nd International Conference on Mathematical Modelling, Vol. 2, University of Missouri-Rolla, Rolla, Missouri, 1980, pp. 851–856.

[31] S.C. Basak, V.R. Magnuson, Molecular topology and narcosis: a quantitative structure–activity relationship (QSAR) study of alcohols using complementary information content (CIC), Arzneim. Forsch. 33 (1983) 501–503.

[32] A.B. Roy, S.C. Basak, D.K. Harriss, V.R. Magnuson, Neighborhood complexities and symmetry of chemical graphs and their biological applications, in: X.J.R. Avula, R.E. Kalman, A.I. Lipais, E.Y. Rodin (Eds.), Mathematical Modelling in Science and Technology, Pergamon Press, New York, 1984, pp. 745–750.

[33] A.T. Balaban, Highly discriminating distance-based topological index, Chem. Phys. Lett. 89 (1982) 399–404.

[34] A.T. Balaban, Topological indices based on topological distances in molecular graphs, Pure Appl. Chem. 55 (1983) 199–206.

[35] A.T. Balaban, Chemical graphs. Part 48. Topological index J for heteroatom-containing molecules taking into account periodicities of element properties, Math. Chem. (MATCH) 21 (1986) 115–122.

[36] S.C. Basak, D.K. Harriss, V.R. Magnuson, POLLY 2.3: Copyright of the University of Minnesota, Minnesota, 1988.

[37] S.C. Basak, B.D. Gute, Characterization of molecular structures using topological indices, SAR QSAR Environ. Res. 7 (1997) 1–21.

[38] S.C. Basak, Information theoretic indices of neighborhood complexity and their applications, in: J. Devillers, A.T. Balaban (Eds.), Topological Indices and Related Descriptors in QSAR and QSPR, Gordon & Breach Science Publishers, The Netherlands, 1999, pp. 563–593.

[39] P.A. Filip, T.S. Balaban, A.T. Balaban, A new approach for devising local graph invariants: derived topological indices with low degeneracy and good correlation ability, J. Math. Chem. 1 (1987) 61–83.

[40] R.E. Carhart, D.H. Smith, R. Venkataraghavan, Atom pairs as molecular features in structure–activity studies: definition and applications, J. Chem. Inform. Comput. Sci. 25 (1985) 64–73.

[41] S.C. Basak, G.D. Grunwald, APProbe: Copyright of the University of Minnesota, 1993.

[42] SAS Institute Inc., SAS/STAT User's Guide, Release 6.03 Edition, SAS Institute Inc., Cary, NC, 1988, Chapter 34, pp. 751–771.

[43] S.C. Basak, G.D. Grunwald, Tolerance space and molecular similarity, SAR QSAR Environ. Res. 3 (1995) 265–277.

[44] S.C. Basak, B.D. Gute, G.D. Grunwald, Estimation of normal boiling points of haloalkanes using molecular similarity, Croat. Chem. Acta 69 (1996) 1159–1173.

[45] S.C. Basak, G.D. Grunwald, Use of graph invariants, volume and total surface area in predicting boiling point of alkanes, Math. Model. Sci. Comput. 2 (1993) 735–740.

[46] S.C. Basak, S. Bertelsen, G.D. Grunwald, Use of graph theoretic parameters in risk assessment of chemicals, Toxicol. Lett. 79 (1995) 239–250.

[47] B.D. Gute, G.D. Grunwald, D. Mills, S.C. Basak, Molecular similarity based estimation of properties: a comparison of structure spaces and property spaces, SAR QSAR Environ. Res. 11 (2001) 363–382.

[48] S.C. Basak, B.D. Gute, G.D. Grunwald, A hierarchical approach to the development of QSAR models using topological, geometrical and quantum chemical parameters, in: J. Devillers, A.T. Balaban (Eds.), Topological Indices and Related Descriptors in QSAR and QSPR, Gordon & Breach Science Publishers, The Netherlands, 1999, pp. 675–696.