Prediction of Vapor Pressures of Hydrocarbons and Halohydrocarbons from Molecular Structure with a Computational Neural Network Model

Eric S. Goll and Peter C. Jurs*

152 Davey Laboratory, Chemistry Department, Penn State University, University Park, Pennsylvania 16802

Received May 10, 1999

Computational methods are used to link the molecular structures of 352 hydrocarbons and halohydrocarbons to their vapor pressures at 25 °C. The data are from the Design Institute for Physical Property Data (DIPPR) database. Vapor pressures of the compounds range from −1.016 log(VP) to +6.65 log(VP) with VP in pascals. Multiple linear regression was used to develop linear statistical models. A 7:3:1 computational neural network (CNN) was used to create a nonlinear model best suited for prediction of vapor pressure. The root-mean-square errors associated with the training, cross-validation, and prediction set compounds used for this CNN model were 0.163, 0.163, and 0.209 log units.

INTRODUCTION

Vapor pressure, the pressure exerted by a vapor in equilibrium with its liquid or solid phase, is an important property on the molecular level. The vapor pressure of a compound is directly related to the forces of intermolecular attraction. Weak intermolecular forces lead to compounds with high vapor pressures, and stronger intermolecular forces lead to compounds with low vapor pressures.

Many other physical properties are related to the vapor pressure of a given compound. Boiling point, for example, has a direct relationship with vapor pressure. A liquid boils when its vapor pressure equals the external pressure acting on the surface of the liquid. Varying the pressure thus causes the boiling point to vary. Other properties, including critical temperature, critical pressure, and heat of vaporization, are also related to vapor pressure.[1]

Knowing the vapor pressure of a compound can be extremely important. If there is a chemical spill in a public area, for example, the vapor pressure must be known in order to estimate its rate of evaporation. The absorption of insecticides and herbicides into soil is also dependent upon vapor pressure. As environmental laws become stricter, concerns over the lingering effects of environmental pollutants have become more widespread. Being aware of compounds' vapor pressures and knowing their concentrations in the atmosphere can be critical. Calculating the values, however, is often difficult as precisely measuring the other experimental parameters can be time-consuming, expensive, or potentially hazardous to workers. Often, values for other experimental parameters are not available. Prediction of vapor pressure by the quantitative structure−property relationship (QSPR) method is an attractive alternative.

In a QSPR study, the structures of a set of compounds are related to a specific physical property through use of a mathematical model. The method is inductive; the models are developed from a training set of compounds for which the property is known. The key assumption in any QSPR study is that molecular structure is closely correlated with the physical property being studied and that compounds with similar structures will have similar values for their physical properties. An advantage to such a study is that not only can a physical property be predicted but also a better knowledge of how the physical property and molecular structure are correlated may be obtained and thus prove beneficial for future studies.

QSPR methodology has been reported quite extensively in the literature to predict many physiochemical properties, such as boiling points,[2,3] chromatographic retention,[4,5] aqueous solubility,[6,7] and supercritical CO_2 solubility.[8]

QSPR is just one of the many methods used to predict physical properties such as vapor pressure. A benefit of the QSPR approach is that predictions of physical properties can be made solely on the basis of molecular structure. Other predictive methods exist that require experimental data. The disadvantages of these types of methods include additional research to find the necessary information if it is available, high costs for labor and compounds if experimentation must be performed, and the possibility of adding experimental error into the predictions if the data are incorrect.

Lyman et. al. recommend two methods that require a minimum of experimental data and are applicable to almost any organic material over a wide range of pressure.[1] The first method, the Antoine equation,[1] can generally be used over a range from 760 to 10^{-3} mmHg. The second method, the modified Watson correlation,[1] can be used over a pressure range from 760 to at least 10^{-7} mmHg. In both cases, the normal boiling point must be known.

The Antoine equation has the general form

$$\ln P_{\mathrm{vp}} = A_2 - \frac{B_2}{T - C_2}$$

where A_2, B_2, and C_2 are constants, T is temperature in kelvin, and P_{vp} is vapor pressure in millimeters of mercury. This equation should only be used to estimate the vapor pressure of chemicals that are either in the liquid or vapor state at the temperature desired.

10.1021/ci9900433 CCC: $18.00 © 1999 American Chemical Society
Published on Web 10/05/1999

The modified Watson correlation is expressed as follows:

$$\Delta H_v = \Delta H_{vb}\left(\frac{1 - T/T_c}{1 - T_b/T_c}\right)^m$$

where ΔH_v is the heat of vaporization in calories per mole, ΔH_{vb} is the heat of vaporization at the normal boiling point in calories per mole, T is the temperature in kelvin, T_b is the temperature of the normal boiling point in kelvin, T_c is the critical temperature in kelvin, and m is the exponent of the Watson correlation. The value of *m* depends on the physical state at the temperature of interest.

Burkhard et. al. studied correlative and noncorrelative predictive methods for a specific type of halocarbon, polychlorinated biphenyls.[9] They concluded that noncorrelative methods have poor predictive ability and as the vapor pressure decreases, the error increases. The best noncorrelative approach was presented by Mackay et al.[10] This method uses the following equation to predict vapor pressure of a solid:

$$\ln P_s = -(4.4 + \ln T_b)[1.803(T_b/T - 1) - 0.803 \ln(T_b/T)] - 6.8(T_m/T - 1)$$

where P_s is the vapor pressure of the solid, T_b is the temperature of the normal boiling point in kelvin, T is the temperature in kelvin, and T_m is the melting point of the solid. To predict the vapor pressure of a liquid, the final term in the above equation is ignored.

Correlative methods, which require a set of compounds with known vapor pressures, have much better predictive abilities than noncorrelative methods. Burkhard et. al. concluded that the best correlative method was one which created a relationship between ΔG_v, the Gibbs' free energy of vaporization, and gas−liquid chromatographic retention indices.[9] Given a set of compounds with known vapor pressures, the correlative relationship can be determined and vapor pressure predictions can be made.

METHODOLOGY

Overview of ADAPT. QSPR studies, as presented here, are performed using the ADAPT[11,12] software system (Automated Data Analysis and Pattern recognition Toolkit). After compiling a data set, ADAPT can be used for descriptor generation, feature selection, and model development. Descriptors are used to characterize different parts of the molecular structures of each of the compounds in the data set. Each descriptor is given a numerical value. Similar compounds should have similar descriptor values. Conversely, very different structures will have very different values. The closeness in descriptor value, in turn, can yield information about the property of interest.

Three types of models are generated in this work. A type I model is a linear model developed by traditional multiple linear regression analysis. A type II model simply utilizes the linear model descriptors and presents them to a neural network to develop a linear/nonlinear hybrid model. Finally, for a type III model, a reduced descriptor pool is analyzed with a genetic algorithm to seek the best subset of descriptors to form a nonlinear model. A computational neural network is used to analyze each subset of descriptors that is considered. The best nonlinear models are discovered, and further analysis characterizes their properties. Type I models are the least computationally intensive of the three, while type III models are the most computationally intensive. Computational time is a tradeoff, as models generated with a computational neural network are generally the most effective models.

Data Set. The data set was comprised of 352 hydrocarbons and halohydrocarbons. The compounds and experimental vapor pressures were provided by the Design Institute for Physical Property Data (DIPPR). The compounds have molecular weights ranging from 30 to 346. The dependent variable used in this study was the base 10 logarithm of the vapor pressure expressed in pascals at 298 K. The vapor pressures of the 352 compounds in the data set ranged from −1.006 log units (for *n*-decylbenzene) to 6.6505 log units (for trifluoromethane). The data set is comprised of compounds containing only carbon, hydrogen, and halogens in various bonding configurations. Table 1 presents a complete listing of the data set compounds and their corresponding experimentally determined vapor pressures.

The 352-compound data set was split randomly into a 300-member training set (tset) and 52-member prediction set (pset). The tset was used to develop type I models and was further divided into a 270-member tset and 30-member cross-validation set (cvset) set for type II and type III model construction.

Molecular Modeling. The HyperChem Molecular Modeling software package was used to enter the molecular structures. After each of the 352 structures was drawn, three-dimensional models were built and their geometry was optimized by HyperChem. To generate the lowest energy conformations of the three-dimensional structures, MOPAC,[13] a semiempirical molecular orbital modeling routine, was utilized. Three-dimensional molecular models are needed for each compound in order to allow the calculation of descriptors based on geometry. If the structures were not in their lowest energy conformations (i.e. the structure reached a local minima), descriptor generation would be hindered as the process is directly related to molecular structure and geometry. A PM3 Hamiltonian[14] was utilized for geometry optimization.

Descriptor Generation. Once all structures in the data set were in their lowest energy conformations, descriptor generation was performed using ADAPT. Four types of descriptors were calculated to encode information about each of the structures. These types include topological, geometric, electronic, and hybrid descriptors. Topological descriptors yield information about the connectivity of compounds. Such descriptors encode κ indices,[15−17] path lengths of molecules,[18,19] count of atom types, types of bonds, hybridization of carbon atoms, molecular distance-to-edge indices,[20,21] and molecular paths and connectivity.[22−27] Geometric descriptors provide information about the shape of the compound. Examples include principal moments of inertia,[28] gravitational indices,[29] shadow-area projections,[30,31] and solvent-accessible molecular surface area and volume.[32] Charge information is encoded by electronic descriptors.[33,34] Charges on the most positive or negative atoms, dipole moments, σ charges, electron density, the energies of the highest occupied molecular orbital and the lowest unfilled molecular orbital, electronegativity, and molecular hardness are all encoded

Table 1. Data Set

		log(VP)	
	compound name	exptl	calcd
1	bromochlorodifluoromethane	5.45	5.41
2	bromotrichloromethane	3.68	3.62
3	bromotrifluoromethane	6.23	6.01
4	dibromodifluoromethane	4.93	5.04
5	chlorotrifluoromethane	6.54	6.28
6	dichlorodifluoromethane	5.84	5.79
7	trichlorofluoromethane	5.04	5.08
8	bromodifluoromethane	5.65	5.57
9	tribromomethane	2.87	3.00
10	chlorodifluoromethane	5.95	5.95
11	dichlorofluoromethane	5.26	5.33
12	chloroform	4.27	4.47
13	trifluoromethane	6.65	6.32
14	bromochloromethane	4.30	4.35
15	dibromomethane	3.77	3.73
16	chlorofluoromethane	5.54	5.78
17	dichloromethane	5.06	5.03
18	difluoromethane	6.20	6.26
19	diiodomethane	2.38	2.72
20	methyl bromide	5.33	5.22
21	methyl chloride	5.75	6.01
22	methyl fluoride	6.59	6.52
23	methyl iodide	4.75	4.57
24	bromotrifluoroethylene	5.42	5.06
25	1,2-dibromotetrafluoroethane	4.72	4.78
26	chlorotrifluoroethylene	5.81	5.44
27	chloropentafluoroethane	5.95	5.81
28	1,1-dichlorotetrafluoroethane	5.33	5.29
29	1,2-dichlorotetrafluoroethane	5.31	5.28
30	1,1,1-trichlorotrifluoroethane	4.47	4.52
31	1,1,2-trichlorotrifluoroethane	4.66	4.66
32	tetrachloroethylene	3.41	3.22
33	1,1,2,2-tetrachlorodifluoroethane	3.96	3.94
34	tetrafluoroethylene	6.53	5.84
35	halothane	4.63	4.65
36	2-chloro-1,1-difluoroethylene	5.69	5.54
37	2-chloro-1,1,1,2-tetrafluoroethane	5.62	5.62
38	1,2-dichloro-1,1,2-trifluoroethane	4.94	5.06
39	2,2-dichloro-1,1,1-trifluoroethane	4.95	4.96
40	2,2-dichloro-1,1,2-trifluoroethane	4.94	4.77
41	trichloroethylene	3.98	3.95
42	pentachloroethane	2.67	2.70
43	pentafluoroethane	6.12	5.97
44	1,1,2,2-tetrabromoethane	0.77	1.29
45	2-chloro-1,1,1-trifluoroethane	5.31	5.52
46	1,1-dichloroethylene	4.88	4.69
47	*cis*-1,2-dichloroethylene	4.43	4.51
48	*trans*-1,2-dichloroethylene	4.76	4.48
49	1,1,1-trichlorofluoroethane	3.89	3.69
50	1,1,1,2-tetrachloroethane	3.20	3.19
51	1,1,2,2-tetrachloroethane	2.87	2.83
52	1,1-difluoroethylene	6.56	6.07
53	1,1,1,2-tetrafluoroethane	5.81	5.97
54	1,1,2,2-tetrafluoroethane	5.81	5.48
55	vinyl bromide	5.13	4.71
56	vinyl chloride	5.54	5.42
57	1-chloro-1,1-difluoroethane	5.53	5.75
58	1,1-dichloro-1-fluoroethane	4.89	5.07
59	1,1,1-trichloroethane	4.20	4.16
60	1,1,2-trichloroethane	3.50	3.47
61	vinyl fluoride	6.44	6.08
62	1,1,1-trifluoroethane	6.11	6.23
63	1,1,2-trifluoroethane	5.33	5.62
64	1,1-dibromoethane	3.52	3.60
65	1,2-dibromoethane	3.26	2.98
66	1,1-dichloroethane	4.48	4.70
67	1,2-dichloroethane	3.92	4.06
68	1,1-difluoroethane	5.76	6.09
69	1,2-difluoroethane	5.22	5.61
70	bromoethane	4.78	4.74
71	ethyl chloride	5.21	5.48
72	ethyl fluoride	5.94	6.17
73	ethyl iodide	4.26	4.16
74	ethane	6.63	6.50
75	1,3-dichlorohexafluoropropane	4.83	5.02
76	hexafluoropropylene	5.80	5.68
77	octafluoropropane	5.92	5.83
78	1,1,1,2,3,3,3-heptafluoropropane	5.66	5.63
79	1,1,1,2,3,3-hexafluoropropane	5.25	5.42
80	3,3,3-trifluoropropene	5.75	5.87
81	1,1,1,2,2-pentafluoropropane	5.67	5.98
82	2,3-dichloropropene	3.81	3.85
83	2-chloropropene	5.02	4.94
84	3-chloropropene	4.66	4.69
85	1,2,3-trichloropropane	2.69	2.43
86	propylene	6.06	5.91
87	1,1-dichloropropane	3.96	4.29
88	1,2-dichloropropane	3.85	3.84
89	1,3-dichloropropane	3.39	3.44
90	1-bromopropane	4.26	4.24
91	2-bromopropane	4.48	4.41
92	isopropyl chloride	4.86	5.04
93	*n*-propyl chloride	4.64	4.92
94	isopropyl iodide	3.95	3.96
95	*n*-propyl iodide	3.76	3.71
96	propane	5.98	6.01
97	decafluorobutane	5.44	5.68
98	1,3-dichloro-*trans*-2-butene	3.16	3.36
99	1,4-dichloro-*cis*-2-butene	2.74	2.79
100	1,4-dichloro-*trans*-2-butene	2.66	2.78
101	3,4-dichloro-1-butene	3.35	3.08
102	1-butene	5.48	5.37
103	*cis*-2-butene	5.31	5.40
104	*trans*-2-butene	5.36	5.40
105	isobutene	5.48	5.42
106	1,2-dichlorobutane	3.35	3.47
107	1,4-dichlorobutane	2.68	2.93
108	2,3-dichlorobutane	3.47	3.61
109	1-bromobutane	3.81	3.78
110	2-bromobutane	3.89	3.99
111	*n*-butyl chloride	4.14	4.38
112	*sec*-butyl chloride	4.34	4.57
113	*tert*-butyl chloride	4.62	4.68
114	isobutyl chloride	4.28	4.48
115	*n*-butyl iodide	3.35	3.28
116	*n*-butane	5.36	5.46
117	isobutane	5.38	5.53
118	2-methyl-1-butene	4.77	4.89
119	2-methyl-2-butene	4.77	4.93
120	3-methyl-1-butene	5.10	4.89
121	1-pentene	4.94	4.82
122	*cis*-2-pentene	4.79	4.87
123	*trans*-2-pentene	4.85	4.87
124	1,5-dichloropentane	2.09	2.41
125	1-chloropentane	3.62	3.89
126	isopentane	4.94	4.97
127	neopentane	5.26	5.18
128	*n*-pentane	4.80	4.90
129	2,3-dimethyl-1-butene	4.51	4.43
130	2,3-dimethyl-2-butene	4.21	4.11
131	3,3-dimethyl-1-butene	4.75	4.55
132	2-ethyl-1-butene	4.35	4.39
133	1-hexene	4.37	4.30
134	*cis*-2-hexene	4.29	4.35
135	*trans*-2-hexene	4.31	4.36
136	*cis*-3-hexene	4.32	4.36
137	*trans*-3-hexene	4.33	4.37
138	2-methyl-1-pentene	4.42	4.37
139	2-methyl-2-pentene	4.32	4.43
140	3-methyl-1-pentene	4.55	4.37
141	3-methyl-*cis*-2-pentene	4.31	4.41
142	3-methyl-*trans*-2-pentene	4.26	4.42
143	4-methyl-1-pentene	4.52	4.39
144	4-methyl-*cis*-2-pentene	4.52	4.43

Table 1. (Continued)

		log(VP)	
	compound name	exptl	calcd
145	4-methyl-*trans*-2-pentene	4.48	4.44
146	2,2-dimethylbutane	4.60	4.65
147	2,3-dimethylbutane	4.51	4.51
148	*n*-hexane	4.46	4.35
149	2-methylpentane	4.46	4.44
150	3-methylpentane	4.40	4.44
151	2-ethyl-1-pentene	3.88	3.89
152	3-ethyl-1-pentene	4.02	3.91
153	1-heptene	3.77	3.80
154	*cis*-2-heptene	3.81	3.85
155	*trans*-2-heptene	3.78	3.86
156	*cis*-3-heptene	4.08	3.87
157	*trans*-3-heptene	3.82	3.87
158	2-methyl-1-hexene	3.94	3.87
159	3-methyl-1-hexene	4.00	3.89
160	4-methyl-1-hexene	4.00	3.89
161	5-methyl-1-hexene	4.03	3.89
162	2,3,3-trimethyl-1-butene	4.19	3.89
163	1-bromoheptane	2.38	2.39
164	2,2-dimethylpentane	4.15	4.15
165	2,3-dimethylpentane	3.94	3.99
166	2,4-dimethylpentane	4.12	4.03
167	3,3-dimethylpentane	4.06	4.12
168	3-ethylpentane	3.92	3.89
169	*n*-heptane	3.76	3.82
170	2-methylhexane	3.91	3.92
171	3-methylhexane	3.91	3.91
172	2,2,3-trimethylbutane	4.13	4.20
173	2,3-dimethyl-1-hexene	3.56	3.47
174	2-ethyl-1-hexene	3.43	3.40
175	2-methyl-1-heptene	3.44	3.40
176	6-methyl-1-heptene	3.52	3.41
177	1-octene	3.18	3.32
178	*cis*-2-octene	3.09	3.36
179	*trans*-2-octene	3.37	3.37
180	*cis*-3-octene	3.37	3.38
181	*trans*-3-octene	3.35	3.39
182	*cis*-4-octene	3.40	3.38
183	*trans*-4-octene	3.39	3.39
184	2,4,4-trimethyl-1-pentene	3.78	3.61
185	2,4,4-trimethyl-2-pentene	3.69	3.67
186	2,2-dimethylhexane	3.57	3.68
187	2,3-dimethylhexane	3.54	3.48
188	2,4-dimethylhexane	3.89	3.52
189	2,5-dimethylhexane	3.57	3.52
190	3,3-dimethylhexane	3.64	3.64
191	3,4-dimethylhexane	3.33	3.46
192	3-ethylhexane	3.32	3.36
193	2-methyl-3-ethylpentane	3.56	3.46
194	3-methyl-3-ethylpentane	3.49	3.56
195	2-methylheptane	3.31	3.42
196	3-methylheptane	3.35	3.40
197	4-methylheptane	3.31	3.40
198	*n*-octane	3.28	3.30
199	2,2,3-trimethylpentane	3.63	3.73
200	2,2,4-trimethylpentane	3.80	3.78
201	2,3,3-trimethylpentane	3.55	3.68
202	2,3,4-trimethylpentane	3.55	3.58
203	2-methyl-1-octene	2.94	2.90
204	7-methyl-1-octene	3.10	2.93
205	1-nonene	2.79	2.77
206	3,3-diethylpentane	3.01	3.02
207	2,2-dimethyl-3-ethylpentane	3.18	3.22
208	2,4-dimethyl-3-ethylpentane	3.10	3.05
209	2,2-dimethylheptane	3.20	3.21
210	2,6-dimethylheptane	2.88	3.02
211	3-ethylheptane	2.96	2.82
212	2-methyloctane	2.93	2.90
213	3-methyloctane	2.91	2.88
214	4-methyloctane	2.98	2.88
215	*n*-nonane	2.56	2.77
216	1-decene	2.38	2.23
217	*cis*-2-decene	2.23	2.36
218	*trans*-2-decene	2.27	2.36
219	*n*-decane	2.48	2.23
220	1-undecene	2.01	1.78
221	*n*-undecane	1.69	1.67
222	1-dodecene	1.42	1.26
223	*cis*-2-dodecene	1.21	1.31
224	*trans*-2-dodecene	1.21	1.31
225	*n*-dodecane	1.00	1.12
226	1-tridecene	0.87	0.74
227	*n*-tridecane	0.73	0.59
228	1-tetradecene	0.16	0.28
229	*n*-tetradecane	0.00	0.12
230	octafluorocyclobutane	5.50	4.64
231	cyclobutane	5.15	5.24
232	hexachlorocyclopentadiene	1.06	0.91
233	cyclopentadiene	4.77	4.57
234	cyclopentene	4.71	4.66
235	cyclopentane	4.60	4.68
236	*n*-decylbenzene	−1.01	−0.51
237	*n*-decylcyclohexane	−0.72	−0.65
238	1,2,4-trichlorobenzene	1.76	1.65
239	1,3,5-triethylbenzene	1.31	1.26
240	*m*-dibromobenzene	1.49	1.45
241	*m*-dichlorobenzene	2.46	2.38
242	*o*-dichlorobenzene	2.25	2.39
243	1,2,3-triethylbenzene	1.26	1.20
244	bromobenzene	2.73	2.67
245	monochlorobenzene	3.19	3.18
246	fluorobenzene	4.00	4.01
247	iodobenzene	2.16	2.28
248	benzene	4.09	3.98
249	1,3-cyclohexadiene	4.11	4.06
250	1,4-cyclohexadiene	3.94	4.01
251	methylcyclopentadiene	4.24	4.13
252	cyclohexene	3.99	4.16
253	1-methylcyclopentene	4.17	4.21
254	3-methylcyclopentene	4.35	4.22
255	4-methylcyclopentene	4.34	4.21
256	cyclohexane	4.11	4.16
257	methylcyclopentane	4.28	4.22
258	*p-tert*-butylethylbenzene	1.69	1.27
259	benzyl chloride	2.27	2.52
260	*o*-chlorotoluene	2.67	2.72
261	*p*-chlorotoluene	2.46	2.73
262	toluene	3.55	3.58
263	cycloheptene	3.51	3.72
264	cycloheptane	3.48	3.64
265	1,1-dimethylcyclopentane	4.05	3.93
266	*cis*-1,2-dimethylcyclopentane	3.72	3.80
267	*trans*-1,2-dimethylcyclopentane	3.89	3.77
268	*cis*-1,3-dimethylcyclopentane	4.07	3.81
269	*trans*-1,3-dimethylcyclopentane	3.70	3.78
270	ethylcyclopentane	3.85	3.69
271	methylcyclohexane	3.78	3.71
272	styrene	2.98	2.92
273	ethylbenzene	3.12	3.13
274	*m*-xylene	3.03	3.07
275	*o*-xylene	2.90	3.02
276	*p*-xylene	3.06	3.07
277	1,5-cyclooctadiene	2.82	3.14
278	vinylcyclohexene	3.59	3.04
279	cyclooctene	3.01	3.27
280	cyclooctane	2.85	3.15
281	1,1-dimethylcyclohexane	3.48	3.44
282	*cis*-1,2-dimethylcyclohexane	3.31	3.35
283	*trans*-1,2-dimethylcyclohexane	3.41	3.25
284	*cis*-1,3-dimethylcyclohexane	3.47	3.38
285	*trans*-1,3-dimethylcyclohexane	3.37	3.27
286	*cis*-1,4-dimethylcyclohexane	3.38	3.38
287	*trans*-1,4-dimethylcyclohexane	3.47	3.26
288	ethylcyclohexane	3.24	3.17

Table 1. (Continued)

		log(VP)	
	compound name	exptl	calcd
289	isopropylcyclopentane	3.33	3.28
290	1-methyl-1-ethylcyclopentane	3.41	3.40
291	*n*-propylcyclopentane	3.08	3.16
292	α-methylstyrene	2.56	2.46
293	*m*-methylstyrene	2.41	2.48
294	*o*-methylstyrene	2.38	2.45
295	*p*-methylstyrene	2.65	2.46
296	*cis*-1-propenylbenzene	2.31	2.52
297	*trans*-1-propenylbenzene	2.27	2.53
298	cumene	2.92	2.77
299	*m*-ethyltoluene	2.61	2.65
300	*o*-ethyltoluene	2.51	2.60
301	*p*-ethyltoluene	2.46	2.64
302	mesitylene	2.53	2.53
303	*n*-propylbenzene	2.61	2.66
304	1,2,3-trimethylbenzene	2.30	2.44
305	1,2,4-trimethylbenzene	2.44	2.47
306	1-methyl-4-vinylcyclohexene	2.76	2.63
307	*n*-butylcyclopentane	2.69	2.62
308	isopropylcyclohexane	2.87	2.74
309	*n*-propylcyclohexane	2.76	2.63
310	*m*-divinylbenzene	1.97	1.78
311	*n*-butylbenzene	2.14	2.17
312	*sec*-butylbenzene	2.49	2.28
313	*tert*-butylbenzene	2.67	2.49
314	*m*-cymene	2.35	2.31
315	*o*-cymene	2.37	2.25
316	*p*-cymene	2.11	2.30
317	*m*-diethylbenzene	2.20	2.20
318	*o*-diethylbenzene	2.16	2.14
319	*p*-diethylbenzene	2.27	2.18
320	2-ethyl-*m*-xylene	2.13	2.05
321	2-ethyl-*p*-xylene	2.11	2.07
322	3-ethyl-*o*-xylene	1.93	2.05
323	4-ethyl-*m*-xylene	2.08	2.07
324	4-ethyl-*o*-xylene	2.14	2.06
325	5-ethyl-*m*-xylene	1.87	2.13
326	isobutylbenzene	2.40	2.33
327	1-methyl-2-*n*-propylbenzene	2.12	2.14
328	1-methyl-3-*n*-propylbenzene	2.16	2.19
329	1-methyl-4-*n*-propylbenzene	2.14	2.17
330	1,2,3,4-tetramethylbenzene	1.63	1.90
331	1,2,3,5-tetramethylbenzene	1.84	1.93
332	1,2,4-triethylbenzene	1.40	1.20
333	α-phellandrene	2.30	2.41
334	*β*-phellandrene	2.32	2.35
335	α-terpinene	2.38	2.30
336	*γ*-terpinene	2.17	2.19
337	terpinolene	2.07	2.08
338	*n*-butylcyclohexane	2.23	2.09
339	1,1-diethylcyclohexane	2.23	2.31
340	1,2,3,4-tetramethylcyclohexane	2.27	2.32
341	*p*-isopropenylstyrene	1.15	1.65
342	1-ethyl-2-isopropylbenzene	2.03	1.83
343	*p*-diisopropylbenzene	1.51	1.52
344	*n*-pentylbenzene	1.62	1.66
345	*p*-*tert*-butylstyrene	1.28	1.49
346	*n*-hexylbenzene	1.17	1.15
347	1,2,3,5-tetraethylbenzene	0.36	0.34
348	*n*-heptylbenzene	0.74	0.66
349	*n*-octylbenzene	0.24	0.21
350	pentaethylbenzene	−0.34	−0.31
351	4-isobutylstyrene	0.69	1.49
352	*n*-nonylbenzene	−0.25	−0.19

with electronic descriptors. There are some descriptors which encode information about more than just one of the aforementioned types; these have been classified as hybrid or combination descriptors. These are developed by combining electronic and geometric information. Hybrid descriptors include charged partial surface area[35] and hydrogen bonding descriptors.

Descriptor Reduction. The descriptor development routines are capable of generating approximately 220 descriptors for each compound. The ratio of descriptors to number of compounds in the data set should be less than 0.6.[36] The overall pool of descriptors is reduced by two methods of objective feature selection to generate a subset of the most information rich descriptors, without the use of the dependent variable. The first is identical testing, which was used to remove any descriptor with greater than 85% identical values. This is important because little useful information would be obtained from a descriptor if each value of that descriptor was identical for most of the compounds in the data set. For example, if each structure had an imbedded aromatic ring, the topological descriptor, the number of rings, would not be useful. The second is pairwise correlation, which was used to remove one of two descriptors that provide very similar information (a pairwise correlation greater than 0.95). For example, if the electronic descriptors, the charge on the most positive atom and the charge on the most negative atom, differed for each compound in the data set, but had identical values to each other for every structure, only one of the two would be useful. Keeping descriptors with similar information in the final reduced pool would be redundant because the key is to obtain different information.

Subjective feature reduction is also important to apply to any QSPR study. In this type of procedure, the dependent variable was used to find subsets of descriptors that correlate best with the physical property being studied. A genetic algorithm with normal crossover mating was applied to reduce the pool of descriptors further and to allow the user to select the number of descriptors desired in the model.

The *T*-values for each of the descriptors was examined to ensure that the model coefficients were meaningful. A total of 84 descriptors, which were generated for each compound in the data set, comprised the reduced pool of descriptors. Models comprised of different numbers of descriptors were studied. Examining the root-mean-square (rms) errors for models containing three descriptors through models containing 10 descriptors, the optimum number of descriptors was determined to be seven.

Multiple Linear Regression Analysis. A type I model is the simplest of the models that can be generated and follows traditional multiple linear regression analysis. Two optimization procedures, a genetic algorithm[37] and simulated annealing,[38] were used to evaluate the reduced pool. Once the most potentially useful subset of descriptors has been selected, multiple linear regression can be used to create a linear model.

Neural Networks. The computational neural networks used in this study were three-layer, fully connected, feed-forward networks. The first layer is an input layer, where a linear transformation of the descriptor values is performed. The number of neurons in the input layer is equal to the number of descriptors selected for the model. The transformed values are then passed to the hidden layer. The input value of a hidden layer neuron is the summation of the products of the weights (neuron connections) times the corresponding outputs of the previous input layer plus a bias term. The summation is put through a nonlinear transfer function, here a sigmoid, and then the resulting values are passed to the output layer, which contains a single neuron which represents the predicted vapor pressure values.

Computational neural networks were used to achieve better models with smaller rms errors than supplied by linear methods. Type II and III models are developed using neural networks. Two important characteristics of neural networks are that they possess many readily available adjustable parameters and are of a nonlinear nature. This combination allows them to create more sophisticated models by minimizing the rms error of the training and prediction sets while maintaining the same structure–property relationship discovered by multiple linear regression.

Training of the neural network was done by use of a quasi-Newton Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm.[37] Training the network involves adjusting the weights and biases in an attempt to match the experimental values of the physical property for the data set. Because of the large number of adjustable parameters, it is possible to overtrain the network. If overtraining does occur, contributions of a small subset of the training set compounds may be considered as a major contribution, thus hindering the ability of the network to accurately predict the physical property in question. To avoid overtraining, the data set is split into a tset and a cvset set. Each connection in the network is made up of a weighting factor and a bias term. The weights and biases are changed during training based on the rms error of the tset; the corresponding values are then calculated for the cvset set. Training involves finding a good set of weights and biases. Training should be stopped when the rms error of the cvset set is at a minimum, because it is believed that overtraining occurs when the rms error begins to rise.[38] At this point, the values of the weights and biases are not changed further.

This study was done on a DEC 3000 AXP Model 500 and DEC Alpha Station 500 workstation running under UNIX. In addition to the ADAPT routine, the study also used two feature selection routines—genetic algorithm[37] and simulated annealing[38]—and computational neural network (CNN) routines developed at Penn State University.

RESULTS AND DISCUSSION

To start modeling, a series of linear type I regression models containing varying numbers of descriptors were examined to determine which subset of descriptors would lead to the best model. Two criteria to use in choosing the best number of descriptors are small rms error with as few descriptors as possible. There was a large difference in rms error between models containing two and three, three and four, four and five, and five and six descriptors. The difference in rms error between models began to tail off between six- and seven-descriptor models. Therefore, it was determined that six- and seven-descriptor models were best. In both cases, all T-values were greater than 4. The best linear six- and seven-descriptor models determined by multiple linear regression analysis were used to build nonlinear type II CNN models. The optimal network architecture is one that has a low rms error and few adjustable parameters. The number of adjustable parameters can be calculated by the following: $AP = IL \times HL + HL \times OL + HL + OL$, where AP is the number of adjustable parameters, IL is the number of neurons in the input layer, HL is the number of neurons in the hidden layer, and OL is the number of neurons in the output layer. To avoid overtraining of the network, the ratio of training set compounds to the number of adjustable parameters should be kept above 2.0.

Seven network architectures containing six descriptors were created; all contained six input neurons and one output neuron, and the number of neurons in the hidden layer was varied. After training the network, the rms errors of the tset and cvset were compared. If the errors were dissimilar, an additional hidden layer was added and the new network was trained. This process was repeated until the errors were similar and low. For some networks, similar and low rms errors for the tset and cvset are not easily achievable. This was the case for the six-descriptor type II models attempted.

Models containing seven descriptors were then created. All contained seven input neurons and one output neuron, and once again, the number of neurons in the hidden layer was varied. A type II study revealed pset, cvset, and tset rms errors of 0.221, 0.171, and 0.171, respectively, for a 7:3:1 (input:hidden:output neurons) CNN architecture which has only 28 adjustable parameters. Since the rms error of the pset was low compared to other architectures, and since the rms errors of the cvset and tset were similar in comparison to each other, the 7:3:1 network was selected as the best architecture for further investigation.

The seven descriptors used in the best type I model and the numerical ranges for the compounds in the data set are shown in Table 2. Of the seven descriptors chosen as the best model, five were topological descriptors and two were combination descriptors. The five topological descriptors came from only two different descriptor programs. Three of them were simple feature counts of compounds—the number of single bonds in a given compound, the number of ring atoms (saturated or unsaturated rings), and the number of fluorine atoms in a given compound. Of the 352 compounds in the data set, all but one contained at least one single bond. Nearly 35% of the compounds in the data set contained at least one ring atom. Nearly 15% of the compounds contained at least one fluorine atom. The values of these three descriptors for each of the compounds yield information such as electronegativity, strength of intermolecular forces, and molecular weight, all of which play a role in a compound's characteristic vapor pressure. The other two topological descriptors are molecular connectivity descriptors: zero-order molecular connectivity and third-order clustering. Molecular connectivity, in general, yields information about the size and the degree of branching in a compound with bond type taken into account. Third-order clustering involves counting the number of connections or paths. Highly branched compounds and those that have high path counts have high vapor pressure. The two combination descriptors included the difference in charged partial surface area and the relative charge weighted surface area. The latter two descriptors encode features of molecules which may be responsible for polar intermolecular interactions by combining surface area and charge information, both of which can also contribute to a characteristic vapor pressure.

Table 2. Descriptors Used in the Linear Type I Model for Predicting Vapor Pressure

label	coeff	SE coeff	range	description[a]
V0	-0.670	1.23×10^{-2}	1.38 to 11.6	zero-order molecular connectivity
NF	0.204	1.80×10^{-2}	0 to 10	no. of fluorine atoms
NSB	5.47×10^{-2}	7.24×10^{-3}	0 to 16	no. of single bonds
NRA	-0.121	4.05×10^{-3}	0 to 8	no. of ring atoms
DPSA	-6.35×10^{-2}	2.31×10^{-3}	8.11 to 50.5	difference in partial surface area
N3C	0.117	7.10×10^{-3}	0 to 16	third-order clustering
RPCG	0.518	6.69×10^{-2}	3.37×10^{-2} to 1	relative charge weighted surface area
CONS	8.15	4.98×10^{-2}		*y*-intercept

[a] V0, zero-order molecular connectivity via a valence molecular connectivity term;[26] NF, a count of the number of fluorine atoms that appear in the compound; NSB, a count of the number of single bonds that are used to connect atoms within the compound; NRA, a count of the number of atoms that appear in any saturated or unsaturated ring system; DPSA, the difference between partial positive surface area and partial negative surface area, the summation of the product of the surface area, and the charge of the *i*th negatively charged atom subtracted from the summation of the product of the surface area and charge of the *i*th positively charged atom; N3C, third-order cluster which involves counting the number of connections or paths; and RPCG, relative positive charge, which is the ratio of the charge of the most positive atom to the summation of the total positive charge of the compound.

Table 3. Descriptors Used in Type III Nonlinear Computational Neural Network Model for Predicting Vapor Pressure

label	range	description[a]
MOLC 7	0 to 4.5	cluster three molecular connectivity
NF	0 to 10	no. of fluorine atoms
MDE 44	0 to 9.04	molecular distance to edge of quaternary-to-quaternary carbons
GRVH 3	4.99 to 11.3	cube root of the gravitation index
QPOS	0.0366 to 0.513	charge on most positive atom
DPSA	8.11 to 50.5	difference in partial surface area
ENEG	3.42 to 7.1	electronegativity

[a] MOLC 7, cluster three molecular connectivity;[25−27] NF, a count of the number of fluorine atoms that appear in the compound; MDE 44, molecular distance to edge of quaternary-to-quaternary carbons;[20,21] GRVH 3, cube root of the gravitation index of all atoms including hydrogens, or the summation over all bonds *x* (mass1 × mass2/(distance)^2); QPOS, the charge on the most positive atom; DPSA, the difference between partial positive surface area and partial negative surface area, the summation of the product of the surface area, and the charge of the *i*th negatively charged atom subtracted from the summation of the product of the surface area and the charge of the *i*th positively charged atom;[35−37] and ENEG, the electronegativity of a compound.

The multiple correlation coefficient (R^2) for this seven-descriptor model was 0.983, and the rms error was 0.186 for the 352 compounds in the data set. The pairwise correlation values for the seven descriptors ranged from 0.027 to 0.674, with an average of 0.322.

Type II models of various architectures were examined; the number of hidden layer neurons ranged from 2 to 5. The best network found had a 7:3:1 architecture, which afforded rms errors of 0.221, 0.171, and 0.171 for the pset, cvset, and tset, respectively.

Next, a genetic algorithm coupled with a CNN was used to seek the best seven-descriptor nonlinear type III model with a 7:3:1 architecture. After creating the nonlinear type III model, the rms errors of each of the sets were reduced to 0.209, 0.163, and 0.163 for the pset, cvset, and tset, respectively. Table 3 lists the seven descriptors supporting the best nonlinear type III CNN model. In contrast to the type I model, seven different descriptor programs contributed to the best model, and the four different types of descriptors were represented. Of the seven descriptors, three were topological, two were electronic, one was geometric, and one was combination. The topological descriptors included the number of fluorine atoms in the compound, the molecular distance to edge, and cluster three molecular connectivity. Molecular distance-to-edge descriptors are concerned with the types of carbon atoms appearing in the molecule.[20,21] Four types of carbon atoms exist: primary, secondary, tertiary, and quaternary. The descriptor of this type that was selected represented the distance-to-edge descriptor between quaternary and other quaternary carbons. The latter descriptor mentioned is a graph-theoretic property used to measure the degree of branching in a given compound. Cluster three molecular connectivity specifically refers to a three-membered straight chain which has at least one atom branched from the chain. Both the distance-to-edge descriptor and the cluster three molecular connectivity descriptor encode information due to branching and possible steric interactions.

The electronic descriptors measured electronegativity of the compound and the charge on the most positive atom. Although these two descriptors reveal different information, they are somewhat related. So it appears that electronic considerations must be taken into account to predict vapor pressure as well.

The geometric descriptor selected in the best model was the cube root of the gravitation index with hydrogen atoms included. Katritzky et al. found the cube root of the gravitation index, which they described as a “bulk cohesiveness descriptor,” to be very useful in QSPR studies of boiling points.[29] The combination descriptor was the difference in charged partial surface area, the difference between partial positive surface area, and partial negative surface area, a CPSA descriptor. CPSA descriptors combine molecular surface area and partial atomic charge information.[35] It is interesting to note that the CPSA descriptor and the number of fluorine atoms in each compound appeared in both type I and type III models. The pairwise correlation values for the seven descriptors ranged from 0.022 to 0.871, with an average of 0.353.

The seven descriptors chosen in the best model fit well for the types of compounds in the study. Since 53 of the compounds—15% of the compounds in the data set—contain at least one fluorine atom, the number of compounds containing fluorine would seem to be a good descriptor for characterizing the data set. And since fluorine is the most electronegative element, it would be reasonable to assume that electronegativity would appear in the model. Similarly, the charge on the most positive atom could be involved. The difference in partial charged surface area also involves

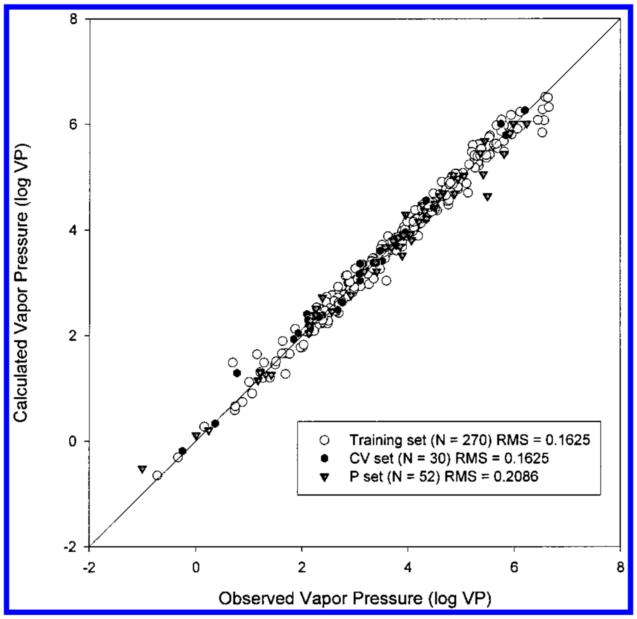

Figure 1. Calculated vs. observed vapor pressure for nonlinear computational neural network model.

charges and is a reasonable descriptor for compounds containing halogens. The cube root of the gravitation index is the cube root of the summation of the product of the masses of the atoms (including hydrogen atoms) in a compound divided by the square of the bond length. Compounds with many atoms (or more halogens) are going to have a larger gravitation index value. The product of the distances between quaternary and quaternary carbons encoded topological information about the types of carbons present in the compounds and the degree of branching. The more quaternary carbons there are, the more non-hydrogen atoms there are, and the higher the vapor pressure. Cluster-three molecular connectivity also describes the structural branching. Again, the more branching there is, the higher the vapor pressure will be. Collectively, the seven descriptors are effective in producing a powerful model that is relatively accurate in predicting the vapor pressure of halocarbons and hydrocarbons.

Figure 1 shows a plot of calculated versus observed log(VP) for the compounds in the entire data set. As the plot shows, most of the compounds fall on or near the one-to-one ideal line, and the model predicts vapor pressure relatively closely. Two pset compounds, however, were observed as possible outliers, octafluorocyclobutane and *n*-decylbenzene, neither of which is representative of a majority of the compounds in the data set. An experiment was performed to investigate the effect of manipulating the data sets by switching these two compounds into the tset and replacing them with two compounds that lay relatively close to the ideal line in the plot, namely, *m*-dibromobenzene and pentaethylbenzene. The altered tset was submitted to a genetic algorithm program. Since manipulation of the training and psets removes the randomness from the study, a better model was found, although it is not reported as the overall best model. The rms errors obtained were 0.171, 0.163, and 0.165 for the tset, cvset, and pset, respectively. The descriptors selected in this model included the following: the Weiner number,[19] zero-order molecular connectivity,[25] second-order molecular connectivity,[25] the number of chlorine atoms in the molecule, the standardized shadow area projected onto the *yz* plane, the partial positive surface area, and the cube root of the gravitational index. Although only two compounds were different from the original sets, only the latter descriptor was selected in both models. Moreover, this model contained four topological, two geometric, and one combination descriptor.

This experiment showed that even minor changes could lead to a different model. Although only one identical descriptor appeared in both models, similar descriptor types (from the same routines) did, however, appear in both models. A side experiment was performed to see the effect of randomizing the cross-validation set. Ten random cross-validation sets were formed, and a genetic algorithm was run with each. Again, some of the same descriptors appeared in several models. But more specifically, similar descriptor types (from the same routines) appeared in almost every model. This suggests that even though some randomness does enter QSPR studies, the effect is minimal in terms of the types of descriptors chosen in models.

A different kind of genetic algorithm was also explored. In this alternative genetic algorithm, 50% of the father chromosome and 50% of the mother chromosome are mated to form the daughter chromosome, instead of crossover mating. No significantly improved models were developed when using the original sets. However, when the two compounds were switched from the tset into the pset (and vice versa), a good model was created. The rms errors of the tset, cvset, and pset were 0.157, 0.153, and 0.157, respectively. The descriptors selected in this model included the following: the number of fluorine atoms in the molecule, the number of single bonds in the molecule, the molecular distance to edge of primary to quaternary carbons, the cube root of the gravitational index, the difference of the partial surface area, third-order clustering, and the relative positive charge (charge of the most positive atom/Σ (total positive charge)). Two of these descriptors appear in the model for the original sets. Four of these descriptors are topological, two are combination, and one is geometric.

Last, the results obtained have been tested to ensure that chance correlations played no role in the models found. A Monte Carlo experiment was performed in which the dependent variables were scrambled. Since the physical property is linked to molecular structure, scrambling the dependent variable should result in high error. As expected, the models generated by the genetic algorithm yielded high rms errors for the three sets. The best model generated by ADAPT had rms errors of 1.20, 1.24, and 1.84 for the set, cvset, and pset, respectively. These errors are an order of magnitude larger than the models generated when the dependent variable was not scrambled. Thus, only the correct dependent variable can be used to generate reasonable models, and the results obtained were not due to chance.

CONCLUSION

A data set comprised of 352 hydrocarbons and halohydrocarbons was used to form accurate models to predict vapor pressures expressed as log(VP) with VP in pascals. The advantage of the QSPR study performed here as compared with other methods is that no experimental parameters are required. In the best seven-descriptor model, all four descriptor types were utilized. Three descriptors were topological, two were electronic, one was geometric, and one was geometric.

The rms error associated with the prediction set was 0.209 log units. This is a significant improvement over previously reported vapor pressure models. These previously reported models, however, were for more diverse data sets consisting of not only hydrocarbons and halohydrocarbons but also O-containing and N-containing compounds as well. For example, Liang and Gallagher utilized a data set comprised of 479 such compounds. Their best model contained seven descriptors and had an rms error of 0.534.[39] Katritzky et. al. used a data set comprised of 411 compounds similar to those used by Liang and Gallagher. Their best model contained five descriptors and had an rms error of 0.331.[40]

It can be seen that the less diverse the types of compounds in the data set are, the better the model appears to be. In the future, other smaller subsets of compounds can be used to generate type-specific models rather than broader models which characterize more diverse data sets.

ACKNOWLEDGMENT

This work was supported by the Design Institute for Physical Property Data (DIPPR) Project 931: Data Prediction Methods.

REFERENCES AND NOTES

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

(2) Wessel, M. D.; Jurs, P. C. Prediction of Normal Boiling Points for a Diverse Set of Industrially Important Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 841−850.

(3) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. Normal Boiling Points for Organic Compounds: Correlation and Prediction by a Quantitative Structure−Property Relationship. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 28−41.

(4) Sutter, J. M.; Peterson, T. A.; Jurs, P. C. Prediction of Gas Chromatographic Relative Retention Times of Alkylbenzenes. *Anal. Chim. Acta* **1997**, *342*, 113−122.

(5) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S. Prediction of Gas Chromatographic Retention Times and Response Factors Using a General Quantitative Structure−Property Relationship Treatment. *Anal. Chem.* **1994**, *66*, 1799−1807.

(6) Sutter, J. M.; Jurs, P. C. Prediction of Aqueous Solubility for a Diverse Set of Heteroatom-Containing Organic Compounds Using a Quantitative Structure−Property Relationship. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 100−107.

(7) Mitchell, B. E.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489−496.

(8) Engelhardt, H. L.; Jurs, P. C. Prediction of Supercritical Carbon Dioxide Solubility of Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 478−484.

(9) Burkhard, L. P.; Andren, A. W.; Armstrong, D. E. Estimation of Vapor Pressures for Polychlorinated Biphenyls: A comparison of Eleven Predictive Methods. *Environ. Sci. Technol.* **1985**, *19*, 500−507.

(10) Mackay, D.; Bobra, A.; Chand, D. W.; Shiu, W. Y. Vapor Pressure Correlations for Low-Volatility Environmental Compounds. *Environ. Sci. Technol.* **1982**, *16*, 645−649.

(11) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley: New York, 1979.

(12) Jurs, P. C.; Chou, J. T.; Yuan, M. In *Computer-Assisted Drug Design*; Olson, E. C., Christoffersen, R. E., Eds.; The American Chemical Society: Washington, D.C., 1979; pp 103−129.

(13) Stewart, J. P. P. *MOPAC 6.0, Quantum Chemistry Program Exchange*; Indiana University, Bloomington, IN, Program 455.

(14) Stewart, J. P. P. MOPAC: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−105.

(15) Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct.−Act. Relat. Pharmacol., Chem. Biol.* **1985**, *4*, 109−116.

(16) Kier, L. B. Shape Indices for Orders One and Three from Molecular Graphs. *Quant. Struct.−Act. Relat. Pharmacol., Chem. Biol.* **1986**, *5*, 1−7.

(17) Kier, L. B. Distinguishing Atom Differences in Molecular Graph Shape Index. *Quant. Struct.−Act. Relat. Pharmacol., Chem. Biol.* **1986**, *5*, 7−12.

(18) Randic, M.; Brissey, G. M.; Spencer, R. B.; Wilkins C. L. Search for All Self-Avoiding Paths for Molecular Graphs. *Comput. Chem.* **1979**, *3*, 5−13.

(19) Weiner, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

(20) Mitchell, B. Ph.D. Thesis, Penn State, 1997.

(21) Liu, S.; Cao, C.; Li, Z. Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance-Edge (MDE) Vector, λ. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387−394.

(22) Kier, L. B.; Hall, L. H.; Murray, W. J.; Randic, M. Molecular Connectivity I: Relationship to Nonspecific Local Anesthesia. *J. Pharm. Sci.* **1975**, *64*, 1971−1974.

(23) Kier, L. B.; Hall, L. H.; Molecular Connectivity VII: Specific Treatment to Heteroatoms. *J. Pharm. Sci.* **1976**, *65*, 1806−1809.

(24) Randic, M. On Molecular Identification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164−175.

(25) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(26) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies Press Ltd., John Wiley & Sons: New York, 1986; pp 18−20.

(27) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

(28) Goldstein, H. *Classical Mechanics*; Addison-Wesley: Reading, MA, 1950; pp 144−156.

(29) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Organics. *J. Phys. Chem.* **1996**, *100*, 10400−10407.

(30) Stouch, T. R.; Jurs, P. C. A Simple Method for the Representation, Quantification, and Comparison of the Volumes and Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 4−12.

(31) Rohrbaugh, R. H.; Jurs, P. C. Molecular Shape and the Prediction of High-Performance Liquid Chromatographic Retention Indexes of Polycyclic Aromatic Hydrocarbons. *Anal. Chem.* **1987**, *59*, 1048.

(32) Pearlman, R. S. Molecular Surface Area and Volumes and Their Use in Structure/Activity Relationships. In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker: New York, 1980; Chapter 10.

(33) Abraham, R. J.; Smith, P. E. Charge Calculations in Molecular Mechanics IV: A General Method for Conjugated Systems. *J. Comput. Chem.* **1987**, *9*, 288−297.

(34) Dixon, S. L.; Jurs, P. C. Atomic Charge Calculations for Quantitative Structure-Property Relationships. *J. Comput. Chem.* **1992**, *13*, 492−504.

(35) Stanton, D. T.; Jurs, P. C. Development and Use of Charged Partial Surface Area Structural Descriptors in Computer-Assisted Quantitative Structure−Property Relationship Studies. *Anal. Chem.* **1990**, *62*, 2323−2329.

(36) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238.

(37) Luke, B. T. Evolutionary Programming Applied to the Development of Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279−1287.

(38) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure−Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.

(39) Liang, C.; Gallagher, D. A. QSPR Prediction of Vapor Pressure from Solely Theoretically-Derived Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 321−324.

(40) Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T.; Karelson, M. QSPR Studies on Vapor Pressure, Aqueous Solubility, and the Prediction of Water−Air Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720−725.

CI9900433