Quantitative Structure−Property Relationships for the Prediction of Vapor Pressures of Organic Compounds from Molecular Structures

Heidi Engelhardt McClelland and Peter C. Jurs*

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

Received November 7, 1999

A quantitative structure−property relationship (QSPR) is developed to relate the molecular structures of 420 diverse organic compounds to their vapor pressures at 25 °C expressed as log(vp), where vp is in pascals. The log(vp) values range over 8 orders of magnitude from −1.34 to 6.68 log units. The compounds are encoded with topological, electronic, geometrical, and hybrid descriptors. Statistical and computational neural network (CNN) models are built using subsets of the descriptors chosen by simulated annealing and genetic algorithm feature selection routines. An 8-descriptor CNN model, which contains only topological descriptors, is presented which has a root-mean-square (rms) error of 0.37 log unit for a 65-member external prediction set. A 10-descriptor CNN model containing a larger selection of descriptor types gives an improved rms error of 0.33 log unit for the external prediction set.

INTRODUCTION

Vapor pressure is a property of interest for scientists studying the environmental fate of a compound, as well as for chemical engineers designing a process.[1] Substances with relatively high vapor pressures are likely to remain in the environment in a gaseous state and thus unlikely to persist in soil and in water. Conversely, substances with very low vapor pressures are less likely to vaporize.[2] In addition to being important in its own right, vapor pressure can be used to predict a number of other physical properties. From the vapor pressure of a substance and its water solubility, many other environmentally relevant quantities such as Henry's law constant, distribution coefficients for adsorption on soil, and partition coefficients can be calculated.[3]

Vapor pressure measurements can be difficult to make, and even unreliable in certain ranges. Equations of state for the calculation of vapor pressure exist, but these typically require a minimum of vapor pressure values from two temperatures.[4] In the absence of experimental data, the only option is to predict the property of interest, using a quantitative structure−property relationship (QSPR), for example.

In this paper, we present QSPR models for vapor pressure prediction developed for a data set of 420 industrially important organic compounds. These models rely on information derived from the compounds' molecular structures only. Vapor pressure prediction has been the topic of three other papers recently.[5−7] A comparison of the results of those studies and this study is made.

EXPERIMENTAL SECTION

The data set for this study came from DIPPR (Design Institute for Physical Property Data) Project 801.[8] From a set of over 1400 compounds, a set of 420 was chosen for this study. All data for inorganic compounds were removed from consideration. The data appearing in the database are ranked according to quality. All compounds with quality ratings of 2 or 3 (indicating less than 3% error) were considered for this study. The DIPPR database consists of a list of compounds, a set of coefficients for an equation to calculate the vapor pressure for each compound, and the maximum and minimum temperatures for which the equation is valid. A temperature of 25 °C was chosen for this study. Several of the compounds were eliminated from consideration, as 25 °C was out of the valid temperature range for them.

The data set contained a total of 420 organic compounds. The structures in the data set contained a diverse array of functional groups. Eighty-seven compounds contained oxygen, 38 contained nitrogen, 27 contained sulfur, 34 contained fluorine, 44 contained chlorine, 14 contained bromine, and 5 contained iodine. The molecular weight range was from 26 to 260 amu, with a mean of 106 amu. The number of carbon atoms in the structures ranged from 1 to 17. Table 1 lists all the compounds in the data set and the logarithms of their vapor pressures. The vapor pressures of the structures in the training set ranged from −1.34 for *n*-undecylbenzene to 6.68 for acetylene. The data set was split randomly into a training set (tset) of 290 compounds, a cross-validation set (cvset) of 65 compounds, and an external prediction set (pset) of 65 compounds. The training set was used to create multiple linear regression or CNN models, and the prediction set was used to validate the models. The cross-validation set was used to prevent overtraining of neural network models.

Sketches and preliminary three-dimensional models of all the compounds in the data set were generated using HyperChem (Hypercube, Inc., Waterloo, Ontario, Canada) running on a Pentium personal computer. The structures were transferred to a DEC ALPHAstation 500 workstation. Using MOPAC[9] with the PM3 Hamiltonian,[10] the geometries of

* Corresponding author. Phone: (814) 865−3739. E-mail: pcj@psu.edu.

10.1021/ci990137c CCC: $19.00
Published on Web 05/25/2000

Table 1. Vapor Pressure Study Data Set

no	compound	log(vapor presure) (Pa, 25 °C)	log(vapor pressure) (Pa, 25 °C), topological type II model	log(vapor pressure) (Pa, 25 °C), best 10-descriptor type II model	no	compound	log(vapor presure) (Pa, 25 °C)	log(vapor pressure) (Pa, 25 °C), topological type II model	log(vapor pressure) (Pa, 25 °C), best 10-descriptor type II model
1	ethane[b]	6.622	6.222	6.350	71	ethyl vinyl ether	4.837	4.865	4.955
2	propane[b]	5.979	5.638	5.821	72	*n*-butyl ethyl ether	3.867	3.700	3.887
3	*n*-butane	5.387	5.097	5.327	73	*cis*-2-butene	5.331	5.312	5.357
4	*n*-pentane	4.835	4.626	4.840	74	*trans*-2-butene[a]	5.370	5.312	5.356
5	*n*-hexane[a]	4.307	4.185	4.347	75	*cis*-2-pentene	4.820	4.740	4.845
6	*n*-heptane	3.783	3.740	3.845	76	*trans*-2-pentene[a]	4.829	4.740	4.839
7	*n*-octane	3.272	3.270	3.349	77	*cis*-2-hexene	4.300	4.267	4.347
8	*n*-nonane	2.764	2.772	2.862	78	*trans*-2-hexene	4.316	4.267	4.336
9	*n*-decane[b]	2.258	2.255	2.376	79	*cis*-2-heptene	3.810	3.810	3.859
10	*n*-undecane	1.745	1.734	1.885	80	*cis*-3-heptene	3.850	3.811	3.826
11	*n*-dodecane[b]	1.252	1.231	1.370	81	*trans*-2-octene[b]	3.339	3.340	3.388
12	*n*-tridecane[a]	0.7551	0.7633	0.8452	82	*cis*-2-octene	3.331	3.340	3.388
13	*n*-tetradecane	0.2705	0.3468	0.3231	83	*trans*-3-octene	3.363	3.340	3.370
14	*n*-pentadecane[a]	−0.1838	−0.0108	−0.1479	84	*cis*-4-octene	3.381	3.342	3.381
15	*n*-hexadecane	−0.7004	−0.3081	−0.5267	85	*trans*-4-octene	3.375	3.342	3.360
16	3-methyl-1-butyne	4.943	5.075	4.605	86	*cis*-3-octene[b]	3.377	3.340	3.361
17	formic acid	3.755	3.140	3.698	87	dimethylamine	5.308	4.572	5.026
18	acetic acid	3.318	2.889	3.340	88	diethylamine[a]	4.497	4.107	4.252
19	propionic acid[a]	2.695	2.319	2.791	89	isobutylamine	4.269	4.344	4.281
20	*n*-pentanoic acid	1.417	1.287	1.516	90	isopropylamine	4.891	4.818	4.590
21	*n*-hexanoic acid	0.7631	0.8400	0.8333	91	di-*n*-butylamine	2.562	1.885	2.374
22	propylene	6.065	5.854	5.877	92	3-nitrobenzotrifluoride[b]	1.535	0.6471	−0.0557
23	1-butene	5.472	5.241	5.323	93	*o*-xylene	2.948	3.037	2.954
24	1-pentene	4.930	4.698	4.814	94	*m*-xylene	3.049	3.035	2.970
25	1-hexene	4.392	4.204	4.288	95	*p*-xylene	3.068	3.035	2.962
26	1-heptene	3.876	3.720	3.809	96	cumene[a]	2.784	2.576	2.617
27	1-octene	3.367	3.218	3.310	97	*o*-ethyltoluene	2.530	2.523	2.530
28	1-nonene[a]	2.857	2.692	2.846	98	*m*-ethyltoluene	2.628	2.523	2.560
29	1-decene	2.355	2.151	2.364	99	*p*-ethyltoluene[b]	2.600	2.522	2.563
30	1-undecene[b]	1.827	1.612	1.862	100	1,2,3-trimethylbenzene[a]	2.335	2.695	2.564
31	1-dodecene	1.423	1.099	1.317	101	1,2,4-trimethylbenzene[a]	2.458	2.693	2.581
32	1-tridecene[a]	0.8653	0.6314	0.7541	102	mesitylene	2.527	2.694	2.612
33	1-tetradecene	0.3726	0.2223	0.2119	103	isobutylbenzene[b]	2.423	2.102	2.159
34	1-pentadecene[a]	−0.1084	−0.1221	−0.2483	104	*sec*-butylbenzene[b]	2.350	2.097	2.101
35	1-hexadecene	−0.5778	−0.4035	−0.5997	105	*tert*-butylbenzene	2.463	2.462	2.249
36	carbon tetrachloride	4.182	4.243	4.313	106	*o*-cymene	2.342	2.257	2.177
37	methyl chloride[b]	5.759	5.855	5.438	107	*m*-cymene[b]	2.366	2.256	2.164
38	ethyl chloride	5.204	5.369	5.307	108	*p*-cymene	2.300	2.256	2.215
39	vinyl chloride	5.600	5.605	5.448	109	*o*-diethylbenzene	2.148	2.057	2.058
40	dichloromethane	4.766	5.177	5.055	110	*m*-diethylbenzene	2.205	2.056	2.101
41	chloroform[a]	4.418	4.580	4.300	111	*p*-diethylbenzene	2.136	2.056	2.106
42	1,1-dichloroethane	4.481	4.790	4.637	112	*p*-diisopropylbenzene	1.516	1.613	1.220
43	1,2-dichloroethane[a]	4.027	4.577	3.985	113	2-ethyl-*m*-xylene	1.990	2.207	2.153
44	1,1,2-trichloroethane	3.489	3.946	3.492	114	2-ethyl-*p*-xylene	2.101	2.207	2.157
45	1,1,1,2-tetrachloroethane[a]	3.205	3.518	3.174	115	4-ethyl-*m*-xylene	2.074	2.207	2.155
46	trichloroethylene[a]	3.992	4.197	3.762	116	4-ethyl-*o*-xylene	1.999	2.206	2.178
47	tetrachloroethylene[b]	3.393	3.541	3.932	117	3-ethyl-*o*-xylene[b]	1.918	2.207	2.147
48	*cis*-1,2-dichloroethylene	4.435	4.902	4.390	118	1-methyl-3-*n*-propylbenzene	2.183	2.048	2.134
49	methanol	4.226	4.692	4.460	119	1-methyl-4-*n*-propylbenzene	2.167	2.048	2.135
50	ethanol	3.899	4.388	3.906	120	methyl propionate[b]	4.055	3.826	4.220
51	1-propanol	3.449	3.826	3.458	121	ethyl propionate	3.690	3.524	3.559
52	1-butanol	2.951	3.264	3.057	122	*n*-propyl propionate	3.270	2.996	3.028
53	1-pentanol	2.521	2.692	2.648	123	vinyl propionate[a]	3.763	3.552	3.718
54	1-heptanol	1.439	1.566	1.672	124	methyl *n*-butyrate	3.633	3.277	3.670
55	1-octanol	1.024	1.054	1.057	125	ethyl *n*-butyrate	3.349	2.997	3.044
56	1-nonanol	0.4971	0.5980	0.4460	126	ethyl isobutyrate[b]	3.529	3.210	3.175
57	1-decanol[a]	0.0561	0.2075	−0.1114	127	cyclobutane	5.195	4.806	5.251
58	methylamine	5.548	5.338	5.257	128	cyclopentane	4.627	4.499	4.763
59	*n*-propylamine	4.616	4.502	4.479	129	cyclohexane[a]	4.115	4.134	4.235
60	*n*-butylamine[b]	4.093	4.008	4.085	130	cycloheptane	3.464	3.705	3.758
61	*n*-pentylamine	3.602	3.504	3.620	131	cyclooctane	2.876	3.217	3.226
62	benzene	4.102	3.920	3.900	132	cyclopentene	4.703	4.549	4.625
63	toluene[a]	3.580	3.384	3.364	133	cyclohexene	4.076	4.192	4.153
64	ethylbenzene	3.107	2.861	2.931	134	cycloheptene	3.524	3.774	3.708
65	*n*-propylbenzene	2.666	2.362	2.535	135	cyclohexanol	1.942	1.947	2.465
66	*n*-butylbenzene	2.156	1.857	2.084	136	1,4-dioxane[a]	3.707	4.149	4.153
67	*n*-undecylbenzene	−1.3380	−0.6693	−0.9605	137	ethylene oxide	5.243	4.935	4.795
68	diethyl ether	4.856	4.607	4.853	138	furan	4.903	4.758	5.058
69	diisopropyl ether	4.297	3.955	4.118	139	tetrahydrofuran	4.333	4.302	4.378
70	methyl *tert*-butyl ether	4.523	4.460	4.707	140	neopentane	5.234	5.085	5.096

Table 1. (Continued)

no	compound	log(vapor presure) (Pa, 25 °C)	log(vapor pressure) (Pa, 25 °C), topological type II model	log(vapor pressure) (Pa, 25 °C), best 10-descriptor type II model
141	2,2-dimethylbutane	4.631	4.559	4.555
142	2,3-dimethylbutane[a]	4.496	4.374	4.465
143	2,2-dimethylpentane[b]	4.147	4.105	4.068
144	2,3-dimethylpentane	3.962	3.952	3.917
145	2,4-dimethylpentane	4.118	3.956	3.983
146	3,3-dimethylpentane	4.042	4.107	4.017
147	2,2-dimethylhexane[a]	3.657	3.679	3.589
148	2,3-dimethylhexane[b]	3.495	3.535	3.445
149	2,4-dimethylhexane	3.607	3.536	3.462
150	2,5-dimethylhexane	3.606	3.540	3.495
151	3,3-dimethylhexane[b]	3.581	3.680	3.517
152	3,4-dimethylhexane	3.461	3.537	3.392
153	2,2-dimethyloctane[b]	2.686	2.821	2.626
154	2,6-dimethylheptane	3.094	3.112	3.011
155	1,3-butadiene	5.449	5.495	5.316
156	*cis*-1,3-pentadiene	4.704	4.852	4.761
157	*trans*-1,3-pentadiene[b]	4.739	4.852	4.793
158	1,4-pentadiene	4.991	4.847	4.969
159	isoprene	4.866	4.988	4.963
160	1,4-hexadiene	4.367	4.311	4.466
161	2,3-dimethyl-1,3-butadiene	4.305	4.518	4.619
162	1,3-cyclohexadiene	4.113	4.270	4.311
163	1-methylnaphthalene[a]	1.002	1.262	1.333
164	1,2,3,4-tetrahydronaphthalene[b]	1.695	1.957	1.913
165	2-ethyl-1-butene	4.369	4.293	4.325
166	2,3-dimethyl-1-butene[a]	4.527	4.403	4.468
167	3,3-dimethyl-1-butene	4.760	4.688	4.530
168	2,3-dimethyl-2-butene	4.222	4.421	4.758
169	2-ethyl-1-pentene	3.882	3.824	3.873
170	3-ethyl-1-pentene[b]	4.035	3.876	3.804
171	2,3,3-trimethyl-1-butene	4.172	4.219	4.130
172	2,4,4-trimethyl-1-pentene[b]	3.775	3.752	3.703
173	2,4,4-trimethyl-2-pentene[a]	3.680	3.776	3.727
174	2,3-dimethyl-1-hexene[b]	3.568	3.492	3.477
175	methyl acetate	4.456	4.376	4.819
176	ethyl acetate[a]	4.094	4.034	4.134
177	*n*-propyl acetate	3.649	3.521	3.578
178	*n*-butyl acetate	3.174	2.992	3.039
179	isobutyl acetate	3.376	3.203	3.228
180	isopropyl acetate	3.906	3.701	3.773
181	methyl fluoride	6.582	6.421	6.294
182	difluoromethane[b]	6.228	6.388	6.469
183	trifluoromethane	6.672	6.430	6.598
184	ethyl fluoride[b]	5.963	5.958	5.909
185	1,1,1-trifluoroethane	6.101	6.412	6.271
186	1,1-difluoroethylene[b]	6.605	6.129	6.204
187	tetrafluoroethylene	6.515	6.112	6.188
188	1,1-difluoroethane	5.776	6.017	6.112
189	pentafluoroethane	6.139	6.209	6.258
190	fluorobenzene	4.012	3.716	3.686
191	3,3,3-trifluoropropene	5.757	6.070	5.918
192	1,1,1,2,3,3-hexafluoropropane	5.313	5.654	5.726
193	1,1,1,2-tetrafluoroethane	5.823	6.158	6.093
194	octafluorocyclobutane	5.496	5.305	5.464
195	methyl formate[a]	4.892	4.809	5.188
196	ethyl formate[a]	4.513	4.427	4.383
197	*n*-propyl formate	4.041	3.895	3.758
198	*n*-butyl formate	3.585	3.358	3.201
199	isobutyl formate	3.738	3.579	3.389
200	methylcyclopentane	4.263	4.114	4.224
201	ethylcyclopentane	3.727	3.692	3.703
202	1,1-dimethylcyclopentane	4.006	3.985	3.914
203	*cis*-1,2-dimethylcyclopentane	3.799	3.847	3.779
204	*trans*-1,2-dimethylcyclopentane	3.931	3.847	3.808
205	*cis*-1,3-dimethylcyclopentane	3.945	3.845	3.813
206	*trans*-1,3-dimethylcyclopentane	3.934	3.845	3.828
207	*n*-propylcyclopentane[a]	3.217	3.216	3.249
208	isopropylcyclopentane	3.332	3.413	3.335
209	methylcyclohexane[b]	3.788	3.698	3.721
210	ethylcyclohexane	3.233	3.218	3.239
211	1,1-dimethylcyclohexane	3.480	3.577	3.429
212	*cis*-1,2-dimethylcyclohexane[b]	3.286	3.414	3.355
213	*trans*-1,2-dimethylcyclohexane	3.412	3.414	3.338
214	*cis*-1,3-dimethylcyclohexane	3.457	3.413	3.323
215	*trans*-1,3-dimethylcyclohexane[b]	3.371	3.413	3.333
216	*cis*-1,4-dimethylcyclohexane	3.379	3.412	3.328
217	*trans*-1,4-dimethylcyclohexane[a]	3.481	3.412	3.342
218	*n*-propylcyclohexane	2.747	2.697	2.797
219	*n*-butylcyclohexane	2.243	2.151	2.354
220	1,2-dichloropropane[a]	3.838	4.230	3.856
221	*n*-propyl chloride	4.662	4.825	4.760
222	*n*-butyl chloride	4.131	4.334	4.222
223	*sec*-butyl chloride	4.321	4.480	4.445
224	1-chloropentane	3.640	3.857	3.690
225	2,3-dichloropropene	3.820	4.246	4.071
226	benzoyl chloride	1.920	1.250	2.028
227	hexafluoroacetone[a]	5.830	5.466	5.792
228	*o*-chlorophenol	2.528	1.209	1.513
229	methyl chloroacetate[a]	3.004	3.452	3.732
230	indan	2.315	2.499	2.389
231	5-ethylidene-2-norbornene	2.898	3.045	2.871
232	acetone[b]	4.488	4.956	5.016
233	methyl ethyl ketone	4.090	4.396	4.479
234	3-pentanone	3.697	3.880	4.003
235	3-hexanone	3.268	3.365	3.467
236	2-pentanone	3.676	3.877	3.925
237	2-hexanone	3.191	3.361	3.393
238	2-heptanone	2.717	2.831	2.876
239	5-methyl-2-hexanone	2.841	3.044	3.047
240	diisobutyl ketone	2.351	2.257	2.452
241	isophorone[b]	1.766	2.135	2.198
242	cyclohexanone	2.742	3.001	3.059
243	acetophenone	1.722	1.691	2.093
244	formaldehyde[a]	5.715	6.148	6.074
245	1-propanal	4.628	4.833	4.766
246	1-butanal	4.175	4.275	4.174
247	2-methylpropanal	4.324	4.456	4.331
248	acrolein	4.563	4.780	4.583
249	isobutane	5.546	5.223	5.389
250	isopentane	4.962	4.729	4.875
251	2-methylpentane[b]	4.449	4.289	4.381
252	3-methylpentane	4.402	4.291	4.339
253	2-methylhexane	3.943	3.862	3.891
254	3-methylhexane[a]	3.913	3.863	3.850
255	2-methylheptane	3.439	3.423	3.403
256	3-methylheptane	3.417	3.423	3.366
257	4-methylheptane	3.436	3.423	3.361
258	3-methylnonane	2.421	2.481	2.404
259	2-methylnonane	2.400	2.480	2.441
260	4-methylnonane	2.490	2.481	2.406
261	5-methylnonane	2.468	2.481	2.406
262	2-methyloctane	2.927	2.962	2.923
263	3-methyloctane	2.921	2.962	2.885
264	4-methyloctane	2.959	2.962	2.883
265	isobutene	5.482	5.348	5.394
266	2-methyl-1-butene	4.910	4.786	4.869
267	3-methyl-1-butene[b]	5.080	4.838	4.908
268	2-methyl-2-butene	4.796	4.855	4.944
269	2-methyl-1-pentene[b]	4.416	4.291	4.390
270	4-methyl-1-pentene	4.559	4.341	4.370
271	2-methyl-2-pentene	4.323	4.333	4.432
272	4-methyl-*cis*-2-pentene[b]	4.512	4.380	4.411
273	2-methyl-1-hexene[b]	3.909	3.822	3.901
274	3-methyl-1-hexene	4.041	3.875	3.867
275	3-methyl-*trans*-2-pentene[a]	4.270	4.334	4.400
276	2-methyl-1-heptene	3.439	3.347	3.453
277	styrene	2.912	2.811	2.922
278	*o*-methylstyrene	2.392	2.471	2.647
279	*m*-methylstyrene	2.403	2.474	2.550
280	*p*-methylstyrene[b]	2.387	2.470	2.534

Table 1. (Continued)

no	compound	log(vapor presure) (Pa, 25 °C)	log(vapor pressure) (Pa, 25 °C), topological type II model	log(vapor pressure) (Pa, 25 °C), best 10-descriptor type II model	no	compound	log(vapor presure) (Pa, 25 °C)	log(vapor pressure) (Pa, 25 °C), topological type II model	log(vapor pressure) (Pa, 25 °C), best 10-descriptor type II model
281	3-methoxypropionitrile	2.420	2.854	2.926	351	2-ethoxyethyl acetate	2.450	2.754	2.773
282	acetonitrile	4.084	4.679	4.935	352	bicyclohexyl[a]	1.156	1.186	1.274
283	propionitrile	3.800	4.043	3.944	353	benzyl chloride	2.240	2.445	2.502
284	methacrylonitrile	3.977	3.587	3.380	354	*o*-dichlorobenzene	2.257	2.477	2.422
285	*n*-butyronitrile	3.415	3.441	3.167	355	*m*-dichlorobenzene[a]	2.457	2.457	2.341
286	valeronitrile	2.988	2.844	2.589	356	*o*-chlorotoluene	2.672	2.730	2.712
287	trans-crotonitrile	3.360	3.433	3.151	357	2,4-dichlorotoluene	1.786	2.051	1.949
288	cis-crotonitrile[a]	3.629	3.433	3.296	358	1,2,4-trichlorobenzene[b]	1.759	1.756	1.514
289	cyanogen	5.758	5.069	5.457	359	2,4-xylenol[a]	1.153	1.357	1.314
290	vinylacetonitrile[a]	3.391	3.461	3.594	360	*m*-cresol[b]	1.267	1.652	1.610
291	3-ethylpentane	3.890	3.864	3.817	361	quinoline[b]	1.047	0.9403	1.945
292	2,2,3-trimethylbutane	4.135	4.173	4.119	362	phenylhydrazine	0.5354	0.9710	0.8562
293	3-ethylhexane	3.428	3.424	3.340	363	pyridine[b]	3.443	3.348	2.927
294	2-methyl-3-ethylpentane	3.503	3.537	3.391	364	aniline[b]	1.955	3.163	2.059
295	3-methyl-3-ethylpentane	3.486	3.683	3.487	365	*N*-methylaniline[a]	1.776	2.239	2.165
296	2,2,3-trimethylpentane	3.631	3.748	3.602	366	*N*,*N*-dimethylaniline	1.988	2.402	2.594
297	2,2,4-trimethylpentane	3.818	3.751	3.646	367	2-methylpyridine	3.179	3.002	2.952
298	2,3,3-trimethylpentane	3.556	3.749	3.564	368	2,6-dimethylpyridine	2.888	2.654	2.788
299	2,3,4-trimethylpentane	3.558	3.618	3.551	369	3-methylpyridine	2.908	3.011	2.739
300	2,2,5-trimethylhexane	3.347	3.347	3.161	370	4-methylpyridine	2.885	2.995	2.704
301	3,3,5-trimethylheptane	2.746	2.939	2.568	371	methyl mercaptan[a]	5.304	4.995	5.285
302	3,3-diethylpentane	2.988	3.260	2.954	372	ethyl mercaptan[a]	4.847	4.695	4.800
303	2,2,3,3-tetramethylpentane[b]	3.103	3.527	3.249	373	*n*-propyl mercaptan[a]	4.313	4.183	4.372
304	2,2,4,4-tetramethylpentane	3.427	3.534	3.386	374	*tert*-butyl mercaptan[b]	4.383	4.454	4.562
305	2,2,3,3-tetramethylhexane[a]	2.730	3.137	2.770	375	isobutyl mercaptan	3.967	3.958	4.096
306	2,2,5,5-tetramethylhexane	3.072	3.146	2.887	376	*sec*-butyl mercaptan	4.032	3.970	4.179
307	2,2-dimethyl-3-ethylpentane	3.177	3.344	3.099	377	isopropyl mercaptan	4.567	4.464	4.609
308	2,4-dimethyl-3-ethylpentane[b]	3.126	3.212	2.993	378	cyclohexyl mercaptan[b]	2.745	2.349	3.065
309	isopropanol	3.782	4.169	3.690	379	*n*-pentyl mercaptan	3.265	3.099	3.399
310	2-methyl-1-propanol	3.146	3.618	3.191	380	*n*-butyl mercaptan	3.791	3.656	3.918
311	2-butanol	3.383	3.626	3.395	381	phenyl mercaptan	2.305	2.386	2.342
312	2-pentanol	2.915	3.062	2.962	382	methyl ethyl sulfide	4.329	4.398	4.219
313	2-methyl-2-butanol	3.347	3.666	3.217	383	methyl *n*-propyl sulfide	3.832	3.902	3.619
314	2-methyl-1-butanol	2.687	3.054	2.743	384	methyl *tert*-butyl sulfide[a]	3.803	3.896	3.584
315	2-ethyl-1-hexanol	1.283	1.335	1.138	385	di-*n*-propyl sulfide	2.949	3.137	2.788
316	2-octanol	1.509	1.341	1.672	386	diethyl sulfide	3.904	4.077	3.850
317	2-ethyl-1-butanol	2.310	2.472	2.239	387	ethyl *n*-octyl sulfide[b]	1.084	1.118	0.6948
318	propargyl alcohol[a]	3.319	4.022	3.624	388	dimethyl sulfide	4.810	4.773	4.745
319	pyrrole[b]	3.040	4.277	3.485	389	thiophene	4.022	4.181	4.326
320	ethylenediamine	3.222	3.006	3.144	390	diethyl disulfide[a]	2.754	3.597	3.045
321	ethyleneimine	4.449	4.415	4.530	391	dimethyl disulfide	3.583	4.487	4.357
322	piperidine	3.632	3.038	3.470	392	tetrahydrothiophene[b]	3.388	3.831	3.421
323	pyrrolidine[a]	3.924	3.549	3.837	393	methyl *n*-butyl sulfide	3.318	3.402	3.083
324	*n*-methylpyrrolidine[a]	4.129	3.971	4.380	394	ethyl *tert*-butyl sulfide[b]	3.415	3.576	3.297
325	pyrimidine	3.352	3.496	3.473	395	trimethylene sulfide[b]	3.845	4.237	3.751
326	benzyl ethyl ether[a]	2.091	1.771	2.393	396	2-methylthiophene[a]	3.519	3.747	3.770
327	anisole	2.684	2.712	3.083	397	3-methylthiophene	3.469	3.734	3.760
328	phenetole[a]	2.314	2.467	2.691	398	methyl acrylate	4.065	3.660	4.258
329	dichlorodifluoromethane	5.814	5.596	5.588	399	*n*-butyl acrylate[b]	2.861	2.186	2.667
330	trichlorofluoromethane	5.026	4.843	4.853	400	ethyl methacrylate	3.433	2.973	3.304
331	chlorodifluoromethane	6.016	5.842	6.021	401	methyl bromide[b]	5.339	5.242	5.258
332	chlorotrifluoromethane	6.551	6.264	6.266	402	bromoethane	4.799	4.701	4.729
333	1,2-dichlorotetrafluoroethane	5.332	5.365	5.160	403	1-bromopropane	4.266	4.159	4.205
334	1,2-dibromotetrafluoroethane[b]	4.657	3.799	4.066	404	2-bromopropane[a]	4.460	4.314	4.513
335	bromodifluoromethane	5.639	5.133	5.391	405	1-bromoheptane	2.229	2.078	2.293
336	1,1,1-trichlorotrifluoroethane	4.681	4.696	4.463	406	1,1-dibromoethane[a]	3.532	3.266	3.599
337	1,1-dichlorotetrafluoroethane[a]	5.338	5.334	5.185	407	1,2-dibromoethane[b]	3.250	3.171	2.983
338	2,2-dichloro-1,1,1-trifluoroethane[a]	4.961	5.049	4.741	408	bromobenzene[a]	2.754	2.531	2.715
339	dichlorofluoromethane[a]	5.262	5.160	5.121	409	dibromomethane	3.780	3.765	3.791
340	bromotrichloromethane	3.716	3.390	3.868	410	methyl iodide	4.732	4.931	4.683
341	2-chloro-1,1,1,2-tetrafluoroethane[b]	5.584	5.599	5.588	411	ethyl iodide	4.254	4.193	4.203
342	1,1-dichloro-1-fluoroethane	4.895	5.072	4.871	412	*n*-propyl iodide	3.759	3.656	3.786
343	1,1,2-trichlorotrifluoroethane[b]	4.648	4.630	4.391	413	iodobenzene	2.151	2.029	2.428
344	bromochlorodifluoromethane	5.440	4.812	5.023	414	*n*-butyl iodide	3.258	3.134	3.389
345	bromotrifluoromethane[b]	6.210	5.567	5.822	415	acetylene	6.688	6.697	6.637
346	chloropentafluoroethane[b]	5.958	6.016	5.910	416	methylacetylene[b]	5.764	6.099	5.924
347	1-chloro-1,1-difluoroethane[a]	5.528	5.792	5.625	417	ethylacetylene	5.275	5.482	5.202
348	ethylene glycol	1.071	2.488	1.161	418	dimethylacetylene[a]	4.974	5.530	5.682
349	2,3-butanediol[b]	1.385	2.158	2.089	419	1-pentyne	4.764	4.923	4.583
350	*n*-methyl-2-pyrrolidone[a]	1.662	2.345	2.483	420	2-hexyne	4.028	4.397	4.715

[a] cvset. [b] pset.

all 420 structures were optimized. The ADAPT[11] software system on the DEC workstation was used to perform the remainder of the study, with the exception of neural network models. ADAPT is a software system which performs all the tasks which are part of a QSPR study. The neural network programs used in this study were written in our group.

Following the optimization of the geometries of the structures, more than 200 descriptors meant to encode features of the structures were calculated. The descriptors were of four types: topological, electronic, geometric, and hybrid. Topological descriptors[12−16] contain information about the number, type, and connectivities of atoms in the molecule. Such descriptors include counts of functional groups in a molecule, atom and bond counts, and path counts.[12,13] Information about the electronic aspects of the structures is encoded by electronic descriptors.[17] Examples of such descriptors include the dipole moment and the sum of the negative charges in the molecule. Geometric descriptors[18−20] are used to encode information about the three-dimensional structure of a compound. One such descriptor is the molecular volume.[19] Finally, hybrid descriptors such as charged partial surface area (CPSA) descriptors[21] combine information about both the geometric and electronic aspects of a molecule. By focusing on solely those parts of a molecule which participate in hydrogen bonding, hydrogen-bonding descriptors can be calculated in the same manner as CPSA descriptors.

Objective descriptor reduction, which does not involve the use of the dependent variable, was performed using various criteria. Reducing the pool of descriptors eliminates those descriptors which contribute no information or whose information content is redundant. It is also necessary to reduce the descriptor pool to eliminate the possibility of chance correlations[22] in the multiple linear regression model. Descriptors which had identical or zero values for greater than 90% of the compounds were eliminated. One of each pair of descriptors with pairwise correlation coefficients exceeding 0.95 was eliminated. The reduced pool contained the remaining 99 descriptors.

Multiple linear regression models, type I models, were created from the reduced pool of descriptors with a simulated annealing algorithm using a multiple linear regression analysis routine to assess the quality of each model.[23] The root-mean-square (rms) error of the training set was used as the measure of the quality of each model. Each model was also checked for its statistical validity by the routine. The routine checked that the *P*-statistic was no higher than 1×10^{-5} and that the *T*-statistic was no lower than 4. The statistics check did not affect the course of the simulated annealing; however, models with unsatisfactory statistics were not included in the final output seen by the user. To determine the optimal model size (number of descriptors), the algorithm was used to generate models of varying sizes, with between three and twelve descriptors. The average rms error for 10 models of each subset size was calculated. The rms error of the models with fewer descriptors was compared to that of the models with more descriptors. The subset size which was chosen was the model size before the point at which the rms error did not decrease significantly with the addition of another descriptor.

The models were examined for multicollinearities among the descriptors in the model. A variance inflation factor (VIF) was calculated for each descriptor in the model. The VIF is defined as $1/(1 - R^2)$, where R is the multiple correlation coefficient for one descriptor regressed against all the other descriptors in the model.[24] Models with a VIF greater than 10 were eliminated from consideration.

The final model was checked for the presence of outliers in the training set. Six standard statistical tests were used to detect outliers.[25] The statistics used were the residuals, standardized residuals, studentized residuals, leverage points, DFITS values, and Cook's distances. Any observation flagged as an outlier by at least four of the tests was considered suspect, and its effect upon the model was investigated. Outliers were removed from the models, and the coefficients were recalculated. If the models did not change substantially when the outliers were removed, the suspect compounds were left in the data set.

The external prediction set was not used at any stage in the model creation process. The compounds in the prediction set were held out until the models were complete. The vapor pressures of the external prediction set of 65 compounds were then predicted as a means of externally validating the model.

If the relationship between the descriptors and the property of interest, in this case vapor pressure, is nonlinear, the nonlinear nature of neural networks may yield models that better predict the property. Descriptors chosen for a multiple linear regression model can be used to make computational neural network models, resulting in a type II model. Descriptors can also be chosen by a feature selection algorithm such as simulated annealing using a neural network fitness function in analogy to the way that multiple linear regression models are chosen using the same algorithm, which results in a type III model.

The descriptors in the best linear model were used as input for a three-layer, fully connected, feed-forward computational neural network. The number of input neurons was equal to the number of descriptors in the linear model, and the output layer consisted of one neuron. The number of neurons in the hidden layer was optimized by varying the number of hidden layer neurons and then selecting the network architecture which resulted in the lowest rms errors for the training and cross-validation sets. Optimal starting weights and biases were selected using a simulated annealing optimization algorithm.[26] A Broyden−Fletcher−Goldfarb−Shanno[27] optimization algorithm is used to train the neural networks. The neural networks used in this study have been explained in more detail in a previous paper.[28] The ratio of training set observations to adjustable parameters in the neural network models was allowed to be no less than 2 to reduce the possibility of models arising solely from chance correlations.[29]

The neural network models were trained using the 290-member training set. As a neural network is trained, it reaches a point at which it loses its ability to generalize and begins to memorize the training set data. The 65-member cross-validation set was used to monitor the training of the network, and training was terminated at the point where the error of the cross-validation set reached a minimum and began to increase. The neural network model which had the lowest rms error for both the training and cross-validation sets was chosen to be the final model. The vapor pressures of the 65 prediction set compounds were predicted to externally

Table 2. Descriptors in the Best Eight Topological Descriptor Model

label	coeff	SE coeff	range	description[a]
QNEG	4.918	0.2011	−0.494 to −0.0806	most negative charge on an atom
V0	−0.8785	0.0175	1.16 to 11.9	valence zero order molecular connectivity
N3C	0.1033	0.0100	0 to 16	number of third-order clusters
CTAA	−1.344	0.072	0 to 2	count of acceptor atoms
RDTA	0.8407	0.0747	0 to 2	ratio of donors to acceptors
NSB	0.1334	0.0106	0 to 15	number of single bonds
NBR	−0.6160	0.0405	0 to 2	number of basis rings
1SP3	0.1751	0.0164	0 to 6	count of primary sp^3-hybridized carbons
CONST	8.186	0.071		

[a] QNEG, charge on the most negatively charged atom; V0, zero-order molecular connectivity;[15,30,31] N3C, number of third-order clusters;[15,30,31] CTAA, count of hydrogen-bonding acceptor atoms; RDTA, ratio of hydrogen-bonding donors to acceptors; NSB, number of single bonds; NBR, number of basis rings; 1SP3, number of sp^3-hybrized carbon atoms bound to only one other carbon atom.

validate the model as the final step in the model-building process.

RESULTS AND DISCUSSION

The first multiple linear regression model generated had eight descriptors. Coincidentally, none of the descriptors in this first model required geometric data for their calculation, despite the fact that descriptors of all types were screened. The most computationally expensive part of making a linear regression model is the geometric optimization, so this is a welcome discovery.

Five compounds were flagged by at least four of the outlier detection measures. The influence of formic acid (17), acetic acid (18), 1,1,1,2,3,3-hexafluoropropane (192), octafluorocyclobutane (194), and ethylene glycol (348) upon the multiple linear regression model was investigated. Five training sets were made excluding one outlier at a time, and the model coefficients were recomputed with each of the five new training sets. The model coefficients were not substantially affected by the removal of any of the five outliers, and so they were all retained in the training set.

The eight descriptors in the model are listed in Table 2. The training, cross-validation, and prediction set rms errors were 0.26, 0.32, and 0.37 log unit, respectively. Pairwise correlations among the eight descriptors ranged from 0.075 to 0.854, with an average of 0.273.

The descriptors in this model encoded solely topological information, meaning in this case that only a 2-D sketch of the molecule is required. While there is no causal relationship between the descriptors in the model and the log(vp), we can still relate the model descriptors to the intermolecular forces they are likely encoding.

One charge descriptor appeared in the model. QNEG is the charge on the most negatively charged atom in the molecule. This could be a measure of the ability of the molecules to participate in dipole−dipole and −induced dipole interactions.

The model contained two molecular connectivity descriptors. V0 is the zero-order molecular connectivity index.[15,30,31] N3C is the number of third-order clusters in the molecule.[15,30,31] These descriptors are encoding information about the size and the degree of branching in the molecule, which is directly related to vapor pressure.

Hydrogen bonding is one of the intermolecular forces which influence vapor pressure. Two hydrogen-bonding descriptors appeared in the model. CTAA is a count of the number of hydrogen-bonding acceptor atoms in the molecule.

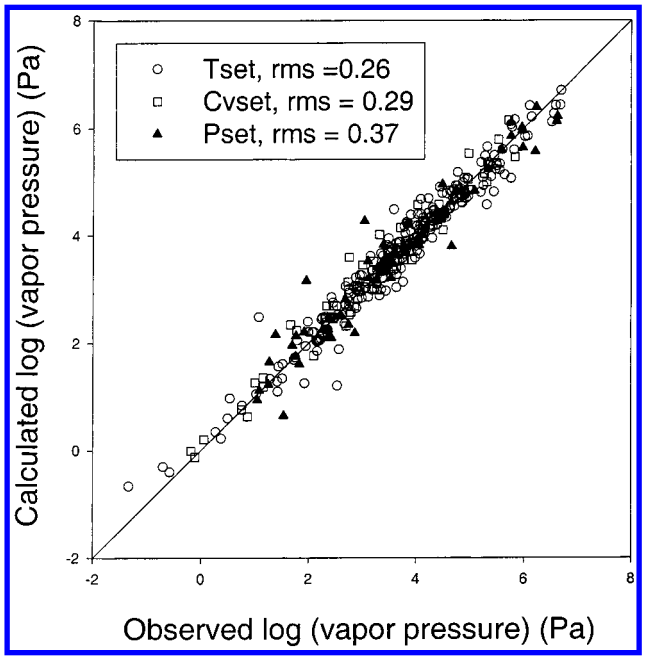

Figure 1. Log(vapor pressure) calculated by the best eight-descriptor type II model vs the observed log(vapor pressure) for all compounds in the data set.

RDTA is the ratio of the number of hydrogen-bonding donors to the number of acceptors in the molecule.

Finally, several counts of molecular fragments appeared in the model. NSB is the number of single bonds. This could be encoding the size of the molecule, as the more single bonds the molecule has, the more atoms it will naturally have. NBR is a count of the number of basis rings in the molecule. Among alkanes, ring structures have lower vapor pressures than their straight-chain counterparts. Finally, 1SP3 is the number of sp^3-hybridized carbon atoms in the molecule which are only bound to one other carbon atom (methyl groups). This is probably encoding branching, as the more methyl groups a compound has, the more branching it has.

To seek a superior model based on these same eight descriptors, they were submitted to a computational neural network. Various CNN architectures were tested, and the optimal architecture was found to be 8−3−1. The calculated values from the resulting model are listed in Table 1. A plot of the calculated versus observed log(vp) for the entire data set is shown as Figure 1. The rms errors were 0.26, 0.29, and 0.37 log unit for the training, cross-validation, and prediction sets, respectively.

Table 3. Descriptors in the Best 10-Descriptor Model

label	coeff	SE coeff	range	description[a]
FNSA 2	−3.561	0.299	−0.9193 to −0.0228	fractional negative SA
WNSA 3	0.2797	0.0184	−10.1600 to −0.2911	weighted negative SA
RNCG 1	−1.526	0.231	0.0830 to 0.9218	ratio negative charge
V1	−0.8006	0.0225	0.3333 to 7.9140	valence first-order molecular connectivity
NAB 1	−0.1439	0.0102	0 to 6	number of aromatic bonds
SCDH 3	−43.45	2.64	0 to 0.07997	sum (SAxQ) of donor H/total SA
RDTA 0	0.5989	0.0693	0 to 2	ratio of donors to acceptors
HPOS	−3.432	0.427	0 to 0.2171	most positive charge on the hydrogen atom
TPIV	−47.75	6.42	0.01069 to 0.03718	polarizability index
RWTP	4.334	0.599	0.02971 to 0.3333	reciprocol of the molecular ID
CONST	7.831	0.198		

[a] FNSA 2, sum of negative SA, times the total negative charge, divided by the total SA;[21] WNSA 3, sum of negative SA times the negative charge, times the total SA, divided by 1000;[21] RNCG 1, charge on the most negative atom divided by the total negative charge;[21] V1, valence first-order molecular connectivity;[15,30,31] NAB 1, number of aromatic bonds; SCDH 3, sum of the surface area times the charge of donatable hydrogens/total surface area of the molecule; RDTA 0, number of H-bond donors/no. of H-bond acceptors; HPOS, charge on the most positively charged H, from MOPAC; TPIV, polarizability, from MOPAC, divided by the volume of the molecule;[34] RWTP, 1 divided by the molecular ID.[14]

In an effort to generate an improved model and a model possibly containing geometry-dependent descriptors, more descriptors were generated and added to the pool of descriptors. Five quantum-chemical descriptors calculated by MOPAC were added to the descriptor pool. In addition, the reciprocals of several descriptors which appeared in previously published boiling point studies[32,33] were added to the pool of descriptors. Also, because boiling point and vapor pressure are related, objective feature reduction was performed again on the entire pool of descriptors, including the additional descriptors, and where a descriptor was to be eliminated by pairwise correlation, those descriptors which appeared in boiling point models were preferentially retained in the descriptor pool.

A 10-descriptor multiple linear regression model was found by the simulated annealing algorithm. As was the case with the eight-descriptor model, coincidentally, five compounds were flagged as outliers by at least four outlier tests. The compounds identified as outliers were *n*-hexadecane (15), *o*-chlorophenol (228), acetonitrile (282), ethylene glycol (348), and phenylhydrazine (362). The only outlier appearing in both the 8-descriptor and 10-descriptor models is ethylene glycol. Model coefficients were recomputed using training sets from which one outlier was excluded at a time. None of the five outliers were found to be exerting undue influence upon the model coefficients, and so they were retained in the training set.

This model contains descriptors of all types which are listed in Table 3. The training, cross-validation, and prediction set rms errors were 0.25, 0.33, and 0.36 log unit, respectively. The largest correlation among the descriptors was 0.843, the smallest was 0, and the average correlation among the descriptors was 0.321.

This model has only one descriptor in common with the eight-descriptor model, RDTA. Three CPSA descriptors appear in the model. These descriptors have been shown in several studies to effectively encode intermolecular interactions. FNSA 2 is the sum of the negative surface area of the molecule, times the total negative charge, divided by the total surface area of the molecule.[21] WNSA 3 is the sum of the negative surface area of the molecule, times its total negative charge, times the total surface area, divided by 1000.[21] RNCG 1 is the charge on the most negatively charged atom divided by the total negative charge of the molecule.[21]

Two topological descriptors appear in the model. V1 is the valence first-order molecular connectivity.[15,30,31] Molecular connectivity descriptors encode information about the size and degree of branching of a molecule, which is directly related to vapor pressure. NAB 1 is a count of the number of aromatic bonds present in the molecule.

Two hydrogen-bonding descriptors are included in the model. They are calculated in a manner analogous to that used to calculate CPSA descriptors, but by taking only atoms which can participate in hydrogen bonding into account in the calculations, hydrogen-bonding descriptors result. SCDH 3 is the sum of the molecular surface area times the charge of donatable hydrogens divided by the total surface area of the molecule. RDTA 0 is the ratio of the number of hydrogen bond donors to the number of hydrogen bond acceptors in the molecule.

Two quantum-chemical descriptors including information calculated by MOPAC are incorporated into the model. HPOS is the charge on the most positively charged H, taken from MOPAC. It is likely encoding the ability of a molecule to hydrogen bond. TPIV is a polarizability index, calculated by taking the polarizability calculated by MOPAC, and dividing by the volume of the molecule.[34]

A final topological descriptor is included in the model. RWTP is the reciprocal of the molecular ID.[14] This descriptor was included because the molecular ID was selected for boiling point models. The molecular ID is a measure of the size and branching of a molecule. The reciprocal was calculated for inclusion in the descriptor pool because the relationship between boiling point and vapor pressure is nonlinear, and that is a nonlinear variant of the original descriptor.

Again the descriptors from the multiple linear regression model were submitted to neural networks. Various CNN architectures were tried, and a neural network with a 10−4−1 architecture was found to be optimum. Figure 2 shows a plot of the log(vp) values calculated by the model vs the observed values. The training, cross-validation, and prediction set rms errors were 0.19, 0.24, and 0.33 log unit, respectively.

The prediction set compound which lies noticeably further from the ideal 1:1 correlation line is 3-nitrobenzotrifluoride (92). It is the only molecule in the data set with a nitro group. The descriptor values for this compound were compared with

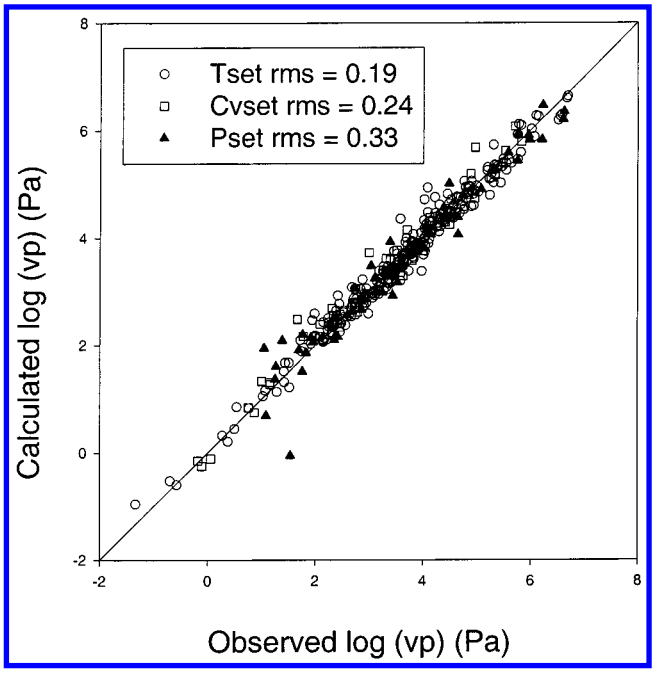

Figure 2. Log(vapor pressure) calculated by the best 10-descriptor type II model vs the observed log(vapor pressure) for all compounds in the data set.

the descriptor ranges for the training set data as shown in Table 3. All but two descriptors were within the ranges. FNSA 2, the fractional negative surface area, was −1.278, as compared to the range for the training set compounds of −0.9193 to −0.0228. WNSA 3, the weighted negative surface area, was −23.78, as compared to the training set range of −10.1600 to −0.2911. The recomputed rms error with that molecule excluded from the prediction set is 0.28 log unit.

Two other prediction set compounds had descriptor values which were outside the range of the training set values for the same descriptors. Methyl chloride (37), with a residual of 0.32, had a value of 0.9891 for RNCG 1. Quinoline (361), with a residual of 0.898, had a value of 11 for NAB1.

A genetic algorithm[35,36] with a computational neural network fitness function was used to select descriptors for a fully nonlinear type III model. A 10−4−1 architecture was used to construct the models. The models found by the algorithm did not have lower training and cross-validation set rms errors than the type II models, however, and are not presented here.

Other studies with similarly sized data sets using similar methodology for the prediction of vapor pressure at 25 °C have appeared in the literature recently. Our model compares favorably with the other published models. Although the units used for vapor pressure in the studies differ, the comparison below is insensitive to differences in units.

Up to this point, the errors for our models have been reported as rms errors; however, the other authors have used standard errors. The rms error of 0.33 log unit for the prediction set for the best, 10-descriptor type II model translates to a standard error of 0.36 log unit.

Liang and Gallagher[7] published a study using a data set of 479 compounds. They screened 25 descriptors and developed a 7-descriptor linear regression model (neural network analysis did not result in an improvement) which yielded a cross-validated standard error of 0.534 log unit. Katritzky et al.[6] published a study using a data set of 411 compounds, in which 800 descriptors were screened. A five-descriptor model with a standard error of 0.33 log unit for the training set resulted. Cross-validated R^2_{cv} compared favorably with R^2. Basak, Gute, and Grunwald[5] published a study on a data set of 476 compounds. The vapor pressures for the compounds in their data set ranged from 2.60 to 6.12 log Pa, which is smaller by about 3 orders of magnitude than the range of vapor pressures in our data set. They screened 92 descriptors and developed a 10-descriptor model with a standard error of 0.29 log unit for the test set of 134 compounds.

A logical extension of this work would be to create models using more homogeneous subsets of compounds drawn from the same DIPPR Project 801 database. One such study on hydro- and halocarbons has been carried out in our group,[37] and similar studies could of course be done for compounds with other functional groups.

CONCLUSION

In this study we have created models for the prediction of the log(vp) of a wide variety of organic compounds. An eight-descriptor CNN model with an 8−3−1 architecture and based entirely on topological information predicted the vapor pressure of an external prediction set with an rms error of 0.37 log unit. A 10-descriptor CNN model including geometry-dependent descriptors with a 10−4−1 architecture yielded an improved prediction set rms error of 0.33 log unit. Our models perform as well as or better than published models using data sets of a similar size and scope.

ACKNOWLEDGMENT

We thank Professor Thomas Daubert for supplying the data from DIPPR (Design Institute for Physical Property Data) Project 801.

REFERENCES AND NOTES

(1) Daubert, T. E.; Jones, D. K. Project 821: Pure Component Liquid Vapor Pressure Measurements. *AIChE Symp. Ser.* **1990**, *86*, 29−39.
(2) Verschueren, K. *Handbook of Environmental Data on Organic Chemicals*; Van Nordstrand Reinhold: New York, 1996.
(3) Site, A. D. The Vapor Pressure of Environmentally Significant Organic Chemicals: A Review of Methods and Data at Ambient Temperature. *J. Phys. Chem. Ref. Data* **1997**, *26*, 157−193.
(4) 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.
(5) Basak, S. C.; Gute, B. D.; Grunwald, G. D. Use of Topostructural, Topochemical, and Geometric Parameters in the Prediction of Vapor Pressure: A Hierarchical QSAR Approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 651−655.
(6) Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T. QSPR Studies on Vapor Pressure, Aqueous Solubility, and the Prediction of Water−Air Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720−725.
(7) Liang, C. K.; Gallagher, D. A. QSPR Prediction of Vapor Pressure from Solely Theoretically-Derived Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 321−324.
(8) Buck, E.; Daubert, T. E. Project 801: The DIPPR Data Compilation Project. *AIChE Symp. Ser.* **1990**, *86*, 5−14.
(9) Stewart, J. P. P. MOPAC 6.0, Quantum Chemistry Program Exchange, Indiana University, Bloomington, IN, Program 455.
(10) Stewart, J. P. P. Mopac: A semiempirical orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−103.
(11) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley-Interscience: New York, 1979.
(12) Wiener, H. J. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

(13) 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.
(14) Randic, M. On Molecular Identification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164−175.
(15) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Wiley: New York, 1986.
(16) Kier, L. B. Distinguishing Atom Differences in a Molecular Graph Shape Index. *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.* **1986**, *5*, 7−12.
(17) Dixon, S. L. Ph.D. Thesis, The Pennsylvania State University, University Park, PA, 1994.
(18) Goldstein, H. *Classical Mechanics*; Addison-Wesley: Reading, MA, 1950.
(19) Perlman, R. S. In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker Inc.: New York, 1980, Chapter 10.
(20) 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.
(21) 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.
(22) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.
(23) Sutter, J. M.; Jurs, P. C. *Adaptation of Simulated Annealing to Chemical Optimization Problems*. In Data Handling in Science & Technology; Kalivas, J. H., Ed.; Elsevier: Amsterdam, 1995; Vol. 15, Chapter 5.
(24) Stanton, D. T.; Jurs, P. C.; Hicks, M. G. Computer-Assisted Prediction of Normal Boiling Points of Furans, Tetrahydrofurans, and Thiophenes. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 301−310.
(25) Belsey, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics*; Wiley: New York, 1980.
(26) 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.
(27) Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. Quantitative Structure−Property Relationships for Toxicity of Phenols Using Regression Analysis and Computational Neural Networks. *Environ. Toxicol. Chem.* **1994**, *13*, 841−851.
(28) Wessel, M. D.; Jurs, P. C. Prediction of Reduced Ion Mobility Constants from Structural Information Using Multiple Linear Regression Analysis and Computational Neural Networks. *Anal. Chem.* **1994**, *66*, 2480−2487.
(29) Livingstone, D. J.; Manallack, D. T. Statistics Using Neural Networks: Chance Effects. *J. Med. Chem.* **1993**, *36*, 1295−1297.
(30) 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.
(31) Kier, L. B.; Hall, L. H. Molecular Connectivity VII: Specific Treatment to Heteroatoms. *J. Pharm. Sci.* **1976**, *65*, 1806−1809.
(32) Egolf, L. E.; Wessel, M. D.; Jurs, P. C. Prediction of Boiling Points and Critical Temperatures of Industrially Important Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 947−956.
(33) Wessel, M. D.; Jurs, P. C. Prediction of Normal Boiling Points for a Diverse Set of Industrially Important Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 841−850.
(34) Famini, G. R.; Wilson, L. Y. Using Theoretical Descriptors in Structure−Activity Relationships: Solubility in Supercritical CO_2. *J. Phys. Org. Chem.* **1993**, *6*, 539−544.
(35) Lucasius, C. B.; Kateman, G. Understanding and Using Genetic Algorithms Part 1. Concepts, Properties and Context. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 1−33.
(36) Hibbert, D. B. Genetic Algorithms in Chemistry. *Chemom. Intell. Lab. Syst.* **1993**, *19*, 277−293.
(37) Goll, E. S.; Jurs, P. C. Prediction of Vapor Pressures of Hydrocarbons and Halohydrocarbons from Molecular Structure with a Computational Neural Network Model. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1081−1089.

CI990137C