Prediction of Aqueous Solubility of Heteroatom-Containing Organic Compounds from Molecular Structure

Nathan R. McElroy and Peter C. Jurs*

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

Received April 15, 2001

The use of quantitative structure−property relationships (QSPRs) to predict aqueous solubilities (log S) of heteroatom-containing organic compounds from their molecular structure is presented. Three data sets are examined. Data set 1 contains 176 compounds having one or more nitrogen atoms with some oxygen (log S[mol/L] range is −7.41 to 0.96). Data set 2 contains 223 compounds having one or more oxygen atoms, with no nitrogen (log S[mol/L] range is −8.77 to 1.57). Data set 3 contains all 399 compounds from sets 1 and 2 (log S/mol/L] range is −8.77 to 1.57). After descriptor generation and feature selection, multiple linear regression (MLR) and computational neural network (CNN) models are developed for aqueous solubility prediction. The best results were obtained with nonlinear CNN models. Root-mean-square (rms) errors for training with the three data sets ranged from 0.3 to 0.6 log units. All models were validated with external prediction sets, with the rms errors ranging from 0.6 log units to 1.5 log units.

INTRODUCTION

Aqueous solubility is an important physical property describing the interaction of a solute, in this case organic liquids and solids, with water. Research in environmental, pharmaceutical, and chemical industries rely upon solubility data to help improve drug delivery systems, minimize environmental impact, and maximize process efficiency. The prediction of aqueous solubility and other physical properties as well as biological activities are even more crucial for improving industrial efficiency. Synthesis of, exposure to, and experiments on new chemicals can be reduced by avoiding all of those which do not fall into a target range of desired property or activity.

Previous aqueous solubility studies from this group[1,2] have used diverse sets of organic compounds including hydrocarbons, halogenated hydrocarbons, and compounds containing several different heteroatoms. These data sets contained large subgroups of compounds from a homologous series, such as hydrocarbon isomers or compounds sharing a common backbone with substituted functional groups. Models were constructed using both multiple linear regression (MLR) and nonlinear computational neural network (CNN) techniques, which have been shown to be very effective in predicting physical properties[3−8] and biological activities.[9,10]

Aqueous solubility studies published by others have employed several different methods. Examples include group-contribution approaches,[11,12] mobile order theory,[13,14] linear modeling using molecular descriptors and other physical data,[15−18] and computational neural network models.[19−21] Compound diversity ranged from very limited—such as modeling the solubility of polychlorinated biphenyls[15] or drug compounds[19,20]—to highly diverse, such as the modeling of over 1000 organic compounds.[21]

One goal of this study was to create a robust predictive model for a more diverse set of compounds—ranging from simple alcohols and amines to complex steroid and drug molecules—having only oxygen and nitrogen heteroatoms. This eliminated several hydrocarbons, halohydrocarbons, and sulfur- or phosphorus-containing compounds from consideration. Another goal was to model two subsets of compounds to discover the effect of separating compounds based simply on heteroatom content and to compare the predictive ability of a general model to a heteroatom-specific model.

EXPERIMENTAL SECTION

This study used three data sets: data set 1 (nitrogen), data set 2 (oxygen), and data set 3 (combined), where sets 1 and 2 were subsets of set 3. Aqueous solubility values for all compounds were taken from the literature,[2,22] and all values were converted from units of mg/L or ppm to units of log(molarity). The choice of compounds was limited to those whose data were experimentally measured at 25 ± 1 °C and whose molecular content fit the profile mentioned above.

Data set 1 contained 176 organic compounds with a minimum of one nitrogen atom, zero or more oxygen atoms, and zero or more halogens per molecule. Solubility values ranged from −7.41 to 0.96 log units (mean = −2.17 log units). Molecular weight ranged from 53.1 to 478.9 amu (mean = 202.8 amu). The compound with the lowest solubility value, 13H-dibenzo[a,i]carbazole, was removed from set 1 during model training but retained in data set 3 modeling. The solubility value of −7.41 log units for this compound was approximately 1.5 log units lower than the next grouping of higher values for the nitrogen set, causing a large gap in solubility value distribution. This value, however, fit well within the combined set data distribution. Data set 2 contained 223 organic compounds with a minimum of one oxygen atom, zero or more halogens, and no nitrogens. Solubility values ranged from −8.77 to 1.57 log units (mean = −2.09 log units). Molecular weight ranged

10.1021/ci010035y CCC: $20.00
Published on Web 07/17/2001

from 72.1 to 959.2 amu (mean = 180.8 amu). Data set 3 contained all 399 compounds from set 1 and set 2. Solubility values ranged from −8.77 to 1.57 log units (mean = −2.11 log units). Molecular weight ranged from 53.1 to 959.2 amu (mean = 190.5 amu). A complete list of compounds with observed log S (mol/L) values are presented in Table 1. Compounds were in solid and liquid states, with approximately a three-to-one solid to liquid distribution.

All computations were performed on a DEC 3000 AXP Model 500 workstation running the Unix operation system. The Automated Data Analysis and Pattern recognition Toolkit (ADAPT) software package[23,24] was used for descriptor generation and formation of linear models. In-house simulated annealing,[25] genetic algorithm,[26] and CNN routines[27] were used to develop linear and nonlinear predictive models. QSPR models were developed in five steps: (1) structure entry-optimization, (2) descriptor generation-objective feature selection, (3) linear feature selection-linear model formation, (4) linear feature selection-non-linear model formation, and (5) nonlinear feature selection-non-linear model formation.

Structure Entry-Optimization. All compounds were sketched on a Pentium-III PC using HyperChem (Hypercube, Inc. Waterloo, ON, Canada), which stored information about bond lengths, bond angles, and atom types in connection tables for use in ADAPT. Several descriptors required accurate three-dimensional structure information, thus these preliminary two-dimensional structures were passed to the semiempirical molecular orbital package MOPAC.[28] We used the PM3 Hamiltonian[29] to find accurate three-dimensional geometries and the AM1 Hamiltonian[30] for molecular charge information because they are best suited for these tasks.[31]

Before descriptor generation and model building, compounds were placed into three subsets: a training set (tset), a cross-validation set (cvset), and an external prediction set (pset). The training set compounds were used in objective and subjective feature selection, to guide linear model building, and for adjusting the weights and biases of a CNN in nonlinear model training. Cross-validation set compounds were used in nonlinear modeling routines to prevent overtraining of the CNN. The prediction set compounds were never used in model building and were used only at the end of the process to demonstrate predictive ability of a model. Compounds were placed randomly into these sets with the condition that prediction set members adequately represented the range of the dependent variables without being the extreme value on either side of that range.

Data set 1 was split into a training set of 131 compounds and 22-member cross-validation and prediction sets. Data set 2 was split into a 167-member training set and 28-member cross-validation and prediction sets. Data set 3 contained a 298-member training set, a 50-member cross-validation set, and a 51-member prediction set.

Descriptor Generation-Objective Feature Selection. A total of 229 descriptors were calculated for each compound. Of those, 141 were topological, 30 were geometric, 10 were electronic, and 48 were hybrid descriptors. Topological descriptors encoded information about the atom types, bond types, and connectivity of the molecule without the need for optimized geometry. Examples included κ indices,[32−34] fragment counts, path counts,[35,36] and molecular connectivity.[37−40] Geometric descriptors gave information about molecular size and shape, thus required accurate three-dimensional coordinates of the optimized geometry. Examples included solvent accessible surface area and volume,[40] moments of inertia,[41] and shadow areas.[42,43] Electronic descriptors[44−46] provided information about the electronic environment of the molecule, such as the energy of the highest occupied molecular orbital or lowest unoccupied molecular orbital, and partial atomic charges. Hybrid descriptors were a combination of the other three types, which presented information about the partially charged surface areas[47] and hydrogen bonding characteristics of the molecule. Hydrogen bonding descriptors were calculated assuming a mixture of water and solute to describe the interactions between molecules.

Before building models, it was desirable to remove descriptors that contained little or no information within the descriptor space, and those descriptors that were highly correlated with each other. Objective feature selection, which does not involve the dependent variable (log S), was performed using only the training set compounds. First, descriptors that contained little or redundant information across the entire range of observations were removed. Next, if a pairwise correlation between two descriptors was greater than 0.90, then one of those two descriptors was also removed. Following these criteria, the ratio of descriptors to training set observations was no greater than 0.6 for each reduced pool, which has been shown to reduce the likelihood of chance correlations in model formation.[48] The reduced pool for data set 1 contained 76 descriptors, while data set 2 had 100 descriptors, and data set 3 had 98 descriptors.

Linear Feature Selection/Linear Model Formation. The next step was to use subjective feature selection to form small subsets of descriptors from the reduced pools generated above. A simulated annealing routine[25] using multiple linear regression assessed the quality of the descriptors based on correlation coefficients (R), descriptor T-values, and minimization of the root-mean-square (rms) error for compounds in the training set. This process gave several linear type I models, which were investigated further using statistical tests.

A variance of inflation (VIF) factor was calculated for each descriptor by regressing that descriptor against all others in the model, except for the dependent variable. VIFs were calculated as $[1/(1 - R^2)]$, where R is a multiple correlation coefficient. Models were considered free of multicollinearities if the VIF values for all model descriptors were less than 10.

Models of various sizes were created and tested in the following manner. Starting with three descriptors, models were increased sequentially in size until no marked improvement in training set rms errors occurred with the addition of another descriptor. This ensured that a maximum amount of information was captured with a minimum number of descriptors. Final models were checked for the presence of outliers in the training set. Six standard statistical tests were performed to test for outliers: residuals, standardized residuals, studentized residuals, leverage points, DFFITS values, and Cook's distances.[49,50] Any observation that was flagged by four or more of these tests as problematic was considered for its effect on the model. Outliers were removed from the model and coefficients were recalculated. If by removing an observation, the rms error greatly improved, that observation was struck from the training set and moved to the

Table 1. 399 Organic Compounds Used in This Study

no.	CAS no.	set no.	obsd log (mol/L)	nitrogen CNN model[a]	combined CNN model[b]	oxygen CNN model[c]	no.	CAS no.	set no.	obsd log (mol/L)	nitrogen CNN model[a]	combined CNN model[b]	oxygen CNN model[c]
1	000050−06−6	nit-1	−2.32	−2.78	−2.57[cv]		76	000098−95−3	nit-76	−1.80	−1.94	−1.77[p]	
2	000050−11−3	nit-2	−2.00	−1.77	−1.97[cv]		77	000099−34−3	nit-77	−2.20	−1.52[p]	−2.08	
3	000050−47−5	nit-3	−3.66	−3.61	−3.60		78	000099−65−0	nit-78	−2.37	−3.02	−2.58	
4	000050−48−6	nit-4	−3.46	−3.65	−4.08		79	000100−02−7	nit-79	−0.90	−1.27	−1.75	
5	000050−49−7	nit-5	−4.19	−3.75[p]	−3.99		80	000100−25−4	nit-80	−3.39	−2.19	−2.89	
6	000052−31−3	nit-6	−1.46	−2.41	−2.53[p]		81	000100−26−5	nit-81	−2.13	−2.00[cv]	−1.55	
7	000052−43−7	nit-7	−1.97	−1.92[p]	−1.88[cv]		82	000100−61−8	nit-82	−1.28	−0.97	−1.10	
8	000053−86−1	nit-8	−5.58	−4.81	−4.16		83	000100−75−4	nit-83	−0.17	−0.50	−0.66[cv]	
9	000056−29−1	nit-9	−2.73	−2.34	−2.60		84	000101−27−9	nit-84	−4.37	−4.41[p]	−4.68	
10	000056−40−6	nit-10	0.52	0.57	0.47		85	000101−42−8	nit-85	−1.61	−1.65	−1.94	
11	000056−41−7	nit-11	0.27	0.39	0.31		86	000102−69−2	nit-86	−2.28	−2.11	−1.77	
12	000056−45−1	nit-12	0.61	0.19	0.32		87	000102−82−9	nit-87	−3.12	−2.96	−2.71	
13	000056−54−2	nit-13	−3.36	−3.33	−3.19		88	000103−84−4	nit-88	−1.29	−1.84	−1.63	
14	000056−84−8	nit-14	−1.42	−0.32[cv]	−0.02[cv]		89	000103−90−2	nit-89	−1.03	−1.12	−1.63	
15	000056−85−9	nit-15	−0.55	−0.54	−0.68[p]		90	000105−56−6	nit-90	−0.82	−0.57	−0.52	
16	000056−75−7	nit-16	−1.79	−1.47[cv]	−1.97		91	000107−13−1	nit-91	0.15	0.38	0.19	
17	000057−13−6	nit-17	0.96	0.49	0.43[p]		92	000107−95−9	nit-92	0.79	0.26	0.11	
18	000057−43−2	nit-18	−1.74	−2.25[cv]	−2.34		93	000108−47−4	nit-93	0.51	−0.41[p]	−0.79	
19	000057−44−3	nit-19	−1.39	−1.64	−1.75[cv]		94	000111−49−9	nit-94	−0.49	−0.55	−0.25[p]	
20	000057−53−4	nit-20	−1.82	−2.02	−1.54		95	000111−86−4	nit-95	−2.81	−3.14	−2.21[cv]	
21	000057−62−5	nit-21	−2.88	−2.92	−2.78		96	000115−58−2	nit-96	−2.18	−2.67	−2.43	
22	000058−08−2	nit-22	−0.95	−1.13	−1.76		97	000119−90−4	nit-97	−3.61	−3.87	−4.25	
23	000058−55−9	nit-23	−1.39	−1.29	−1.56		98	000119−91−5	nit-98	−5.40	−4.74	−4.82	
24	000059−67−6	nit-24	−0.84	−1.32[cv]	−0.85		99	000119−93−7	nit-99	−2.21	−3.51	−3.64	
25	000060−18−4	nit-25	−2.57	−1.92	−1.44		100	000120−72−9	nit-100	−1.52	−1.80	−1.43	
26	000060−32−2	nit-26	0.59	−0.41	−1.06		101	000121−73−3	nit-101	−2.76	−3.05[cv]	−2.47	
27	000060−54−8	nit-27	−3.28	−2.59[p]	−2.72[p]		102	000121−82−4	nit-102	−3.57	−3.11	−2.91[cv]	
28	000060−80−0	nit-28	−0.56	−1.64[cv]	−2.13		103	000121−92−6	nit-103	−1.67	−1.26[p]	−2.23	
29	000061−82−5	nit-29	0.52	−0.76[p]	−0.55		104	000125−40−6	nit-104	−2.18	−1.90[p]	−2.05	
30	000061−90−5	nit-30	−0.79	−0.12	−0.63		105	000125−42−8	nit-105	−2.43	−2.09[p]	−2.34[cv]	
31	000062−53−3	nit-31	−0.41	−0.73	−0.42		106	000126−98−7	nit-106	−0.42	0.16[cv]	−0.23	
32	000062−57−7	nit-32	0.21	0.24[cv]	−0.10[cv]		107	000129−66−8	nit-107	−1.10	−1.59[p]	−1.95	
33	000063−91−2	nit-33	−0.92	−1.66[cv]	−1.37		108	000131−89−5	nit-108	−4.25	−3.57[cv]	−2.13[p]	
34	000065−45−2	nit-34	−1.82	−1.85	−1.12[p]		109	000132−60−5	nit-109	−3.19	−3.64[cv]	−3.32	
35	000065−49−6	nit-35	−1.96	−1.98	−1.18		110	000132−66−1	nit-110	−3.16	−3.99	−3.28	
36	000065−71−4	nit-36	−1.52	−0.92[cv]	−1.05		111	000133−90−4	nit-111	−2.47	−2.96[p]	−2.42[p]	
37	000066−02−4	nit-37	−2.86	−2.82	−2.10		112	000139−13−9	nit-112	−0.51	−0.77	0.00	
38	000066−22−8	nit-38	−1.49	−0.78	−1.52		113	000150−68−5	nit-113	−2.92	−2.62	−2.51	
39	000066−71−7	nit-39	−1.83	−2.52	−2.68[cv]		114	000239−64−5	nit-114	−7.41	*d*	−7.01	
40	000067−20−9	nit-40	−3.48	−3.32	−2.93		115	000260−94−6	nit-115	−3.60	−3.45[p]	−3.55	
41	000068−94−0	nit-41	−2.37	−2.08[cv]	−1.54		116	000314−40−9	nit-116	−2.52	−2.36	−2.52	
42	000070−34−8	nit-42	−2.67	−2.84	−2.85		117	000315−30−0	nit-117	−2.38	−2.04	−1.58	
43	000070−47−3	nit-43	−0.66	−0.55	−0.37		118	000327−57−1	nit-118	−0.90	−0.23	−0.61[cv]	
44	000071−00−1	nit-44	−0.53	−0.60[p]	−1.24		119	000330−54−1	nit-119	−3.76	−3.63	−3.51	
45	000071−30−7	nit-45	−1.14	−1.02	−0.64		120	000330−55−2	nit-120	−3.52	−3.39	−3.51[p]	
46	000072−18−4	nit-46	−0.30	−0.01	−0.26		121	000426−13−1	nit-121	−4.10	−4.10	−3.66	
47	000072−19−5	nit-47	−0.09	0.00	0.18		122	000439−14−5	nit-122	−3.76	−3.87	−3.48[cv]	
48	000073−22−3	nit-48	−1.18	−1.56	−1.93		123	000443−48−1	nit-123	−1.26	−1.13	−1.44	
49	000073−24−5	nit-49	−2.12	−2.72	−1.82		124	000461−58−5	nit-124	−0.31	−0.29	−0.36	
50	000073−32−5	nit-50	−0.58	−0.23	−0.30		125	000485−71−2	nit-125	−3.17	−3.14	−3.02[p]	
51	000074−79−3	nit-51	0.02	−0.71	−1.25		126	000490−11−9	nit-126	−1.85	−1.72	−0.88	
52	000075−52−5	nit-52	0.26	0.24	−0.38		127	000495−69−2	nit-127	−1.68	−1.51	−1.75	
53	000076−24−4	nit-53	−2.23	−2.59	−2.29		128	000499−80−9	nit-128	−1.83	−1.96	−1.27[cv]	
54	000076−73−3	nit-54	−2.23	−2.34	−2.31		129	000499−81−0	nit-129	−2.19	−1.91	−1.33	
55	000076−74−4	nit-55	−2.52	−2.22	−2.25		130	000509−86−4	nit-130	−3.00	−2.70	−2.50[cv]	
56	000076−76−6	nit-56	−2.21	−1.75[cv]	−2.05		131	000528−29−0	nit-131	−3.10	−2.28[cv]	−2.13	
57	000077−02−1	nit-57	−1.71	−1.88	−1.98		132	000530−78−9	nit-132	−4.49	−3.70	−3.79	
58	000077−21−4	nit-58	−4.36	−4.26[p]	−3.12		133	000552−16−9	nit-133	−1.33	−1.37[p]	−1.35[p]	
59	000077−28−1	nit-59	−1.64	−2.11	−2.21		134	000552−89−6	nit-134	−3.88	−2.41	−1.84	
60	000079−06−1	nit-60	0.95	0.19	0.31		135	000553−26−4	nit-135	−1.54	−1.72	−1.79	
61	000079−07−2	nit-61	−0.02	−0.16[cv]	0.04		136	000554−84−7	nit-136	−1.01	−0.99	−1.46	
62	000079−57−2	nit-62	−3.17	−3.24[cv]	−2.69		137	000555−37−3	nit-137	−4.83	−4.26	−3.92	
63	000080−60−4	nit-63	0.31	0.12	0.03		138	000556−88−7	nit-138	−1.37	−1.20	−0.76	
64	000082−71−3	nit-64	−1.66	−1.95[p]	−1.74		139	000578−06−3	nit-139	−4.22	−3.40	−3.41	
65	000083−67−0	nit-65	−2.74	−1.84[p]	−2.18[cv]		140	000601−77−4	nit-140	−1.00	−0.67[cv]	−0.82	
66	000083−88−5	nit-66	−3.65	−2.94	−2.50[cv]		141	000603−11−2	nit-141	−1.02	−1.70	−1.44	
67	000085−41−6	nit-67	−2.61	−2.68[cv]	−2.32[cv]		142	000610−30−0	nit-142	−1.07	−1.45	−1.79[cv]	
68	000087−17−2	nit-68	−3.59	−2.85	−2.75[cv]		143	000621−64−7	nit-143	−1.12	−0.68	−0.89	
69	000087−33−2	nit-69	−2.63	−3.65	−2.74		144	000628−05−7	nit-144	−1.95	−1.71	−1.78[cv]	
70	000088−75−5	nit-70	−1.80	−1.14	−1.38		145	000645−05−6	nit-145	−3.36	−2.97	−3.33[p]	
71	000088−89−1	nit-71	−1.26	−1.52[cv]	−1.88[cv]		146	000738−70−5	nit-146	−2.86	−3.26	−3.72	
72	000089−00−9	nit-72	−1.18	−1.71	−0.76		147	000938−91−0	nit-147	−1.67	−1.25	−1.53	
73	000092−87−5	nit-73	−2.76	−3.42	−2.80		148	000957−51−7	nit-148	−2.98	−3.28	−2.93	
74	000096−88−8	nit-74	−1.55	−2.35	−2.50[cv]		149	001198−37−4	nit-149	−1.94	−1.72	−2.67	
75	000097−56−3	nit-75	−4.51	−3.45	−3.03[p]		150	001689−83−4	nit-150	−3.61	−3.50	−3.16	

Table 1 (Continued)

no.	CAS no.	set no.	obsd log (mol/L)	nitrogen CNN model[a]	combined CNN model[b]	oxygen CNN model[c]	no.	CAS no.	set no.	obsd log (mol/L)	nitrogen CNN model[a]	combined CNN model[b]	oxygen CNN model[c]
151	001918−02−1	nit-151	−2.75	−2.91	−2.95		226	000090−64−2	ox-50	0.08		−0.62	−0.56
152	001982−49−6	nit-152	−4.11	−3.52	−2.67		227	000092−69−3	ox-51	−3.48		−3.04	−3.51
153	002164−08−1	nit-153	−4.59	−4.08	−2.78		228	000093−09−4	ox-52	−3.89		−2.94	−3.83
154	002164−17−2	nit-154	−3.42	−2.99[p]	−3.57[cv]		229	000093−72−1	ox-53	−3.28		−3.30	−2.86
155	002180−92−9	nit-155	−2.08	−3.27	−2.90		230	000093−76−5	ox-54	−2.97		−3.46[p]	−3.32
156	002516−95−2	nit-156	−1.32	−1.96	−2.03		231	000093−89−0	ox-55	−2.32		−2.66	−1.97
157	002623−33−8	nit-157	−1.91	−3.16	−2.83[cv]		232	000094−13−3	ox-56	−2.62		−2.25	−1.90
158	002631−37−0	nit-158	−3.35	−2.77	−3.05[cv]		233	000094−26−8	ox-57	−2.86		−3.13	−2.73
159	002813−95−8	nit-159	−2.11	−2.66	−2.89		234	000094−74−6	ox-58	−2.23		−2.93	−2.84[cv]
160	003625−25−0	nit-160	−2.77	−3.20	−2.73		235	000094−75−7	ox-59	−2.52		−2.39	−2.13
161	005902−51−2	nit-161	−2.48	−2.38[p]	−2.76		236	000094−81−5	ox-60	−3.68		−3.37[cv]	−3.40[p]
162	006153−64−6	nit-162	−3.14	−3.20[cv]	−2.69		237	000094−96−2	ox-61	−0.54		−0.74[p]	−0.77
163	010004−44−1	nit-163	−0.07	−0.24	−0.38		238	000095−48−7	ox-62	−0.62		−0.73[p]	−0.69
164	010118−90−8	nit-164	−0.94	−1.58	−2.74[p]		239	000095−65−8	ox-63	−1.41		−1.61	−1.35
165	012771−68−5	nit-165	−2.60	−1.72[p]	−2.79		240	000095−87−4	ox-64	−1.54		−1.55	−1.54
166	018530−56−8	nit-166	−3.17	−3.40	−2.77[p]		241	000096−22−0	ox-65	−0.25		−0.49	−0.26
167	019666−30−9	nit-167	−5.69	−4.15	−4.54		242	000097−23−4	ox-66	−1.82		−3.27	−3.58[p]
168	021725−46−2	nit-168	−3.15	−3.73	−2.96		243	000097−53−0	ox-67	−1.41		−2.08	−2.18[p]
169	022212−55−1	nit-169	−4.26	−4.82	−4.82		244	000098−01−1	ox-68	−0.08		−0.08	−0.21
170	022781−23−3	nit-170	−3.75	−3.00	−3.54		245	000098−54−4	ox-69	−2.41		−2.28	−2.60
171	023103−98−2	nit-171	−1.95	−2.23	−2.79		246	000098−86−2	ox-70	−1.32		−1.68[p]	−1.52
172	023950−58−5	nit-172	−4.23	−4.19	−4.31		247	000099−04−7	ox-71	−2.14		−1.59	−1.61
173	029091−05−2	nit-173	−5.47	−4.40	−4.71		248	000099−49−0	ox-72	−2.06		−2.18	−2.19
174	031431−39−7	nit-174	−3.62	−4.40	−4.12		249	000099−71−8	ox-73	−2.19		−1.98	−1.52
175	051940−44−4	nit-175	−2.97	−3.23[p]	−3.36		250	000099−76−3	ox-74	−1.82		−1.89[p]	−1.38
176	060168−88−9	nit-176	−4.38	−1.43	−3.69[cv]		251	000099−96−7	ox-75	−1.43		−1.08	−1.03[cv]
177	000050−02−2	ox-1	−3.64		−3.21	−3.78[cv]	252	000100−52−7	ox-76	−1.26		−1.07	−1.30
178	000050−03−3	ox-2	−4.46		−3.80	−4.01	253	000101−84−8	ox-77	−3.95		−3.72[cv]	−3.79[cv]
179	000050−04−4	ox-3	−4.30		−4.63	−4.33[cv]	254	000103−36−6	ox-78	−3.00		−3.22	−2.75[cv]
180	000050−23−7	ox-4	−3.05		−3.10	−3.19	255	000103−73−1	ox-79	−2.33		−2.16	−2.20
181	000050−24−8	ox-5	−3.21		−3.13[p]	−3.23	256	000103−82−2	ox-80	−0.89		−1.30	−1.21[p]
182	000050−78−2	ox-6	−1.59		−1.46[p]	−1.54[p]	257	000104−40−5	ox-81	−4.50		−5.17	−4.55
183	000051−98−9	ox-7	−4.79		−4.55[cv]	−4.29	258	000104−46−1	ox-82	−3.13		−2.74	−3.29
184	000053−06−5	ox-8	−3.11		−3.87	−3.79	259	000104−87−0	ox-83	−1.72		−1.70	−1.59
185	000053−16−7	ox-9	−5.53		−4.81	−4.82	260	000105−30−6	ox-84	−1.23		−0.91	−0.76
186	000053−41−8	ox-10	−4.38		−4.14	−3.77[p]	261	000105−57−7	ox-85	−0.43		−1.21[p]	−0.86
187	000053−43−0	ox-11	−3.66		−4.12	−4.04	262	000105−67−9	ox-86	−1.22		−1.55	−1.45
188	000057−10−3	ox-12	−5.49		−6.12[cv]	−5.74	263	000106−44−5	ox-87	−0.70		−0.87	−0.79[p]
189	000057−83−0	ox-13	−4.53		−4.81	−4.86[p]	264	000106−51−4	ox-88	−0.88		−0.66	−0.77[p]
190	000058−18−4	ox-14	−3.95		−4.53	−4.23[cv]	265	000106−89−8	ox-89	−0.15		−0.22	0.17
191	000058−22−0	ox-15	−4.05		−4.17	−4.04[p]	266	000107−87−9	ox-90	−0.30		−0.52[p]	−0.53
192	000060−29−7	ox-16	0.41		−0.21	−0.08	267	000108−10−1	ox-91	−0.72		−0.57[p]	−0.93
193	000063−05−8	ox-17	−3.70		−4.64	−4.87[p]	268	000108−11−2	ox-92	−0.79		−0.42	−0.63
194	000065−85−0	ox-18	−1.56		−1.11[p]	−1.62	269	000108−21−4	ox-93	−0.60		−0.34	−0.52
195	000068−22−4	ox-19	−4.67		−4.57[cv]	−4.22[cv]	270	000108−39−4	ox-94	−0.68		−0.84	−0.68
196	000069−65−8	ox-20	0.07		0.41[p]	0.63	271	000108−68−9	ox-95	−1.40		−1.40[p]	−1.26
197	000069−72−7	ox-21	−1.81		−0.71[p]	−1.34	272	000108−83−8	ox-96	−1.73		−1.50[p]	−2.30
198	000070−30−4	ox-22	−3.71		−4.71	−3.88	273	000108−95−2	ox-97	−0.03		−0.34	0.06[cv]
199	000071−36−3	ox-23	−0.02		−0.13[p]	0.04	274	000109−52−4	ox-98	−0.42		−0.52[p]	−0.51
200	000071−41−0	ox-24	−0.60		−0.55[p]	−0.43	275	000110−15−6	ox-99	−0.15		0.18	0.29
201	000072−43−5	ox-25	−6.68		−5.78[cv]	−6.24	276	000110−16−7	ox-100	0.58		0.29	0.42
202	000075−84−3	ox-26	−0.40		0.23	−0.41[cv]	277	000110−19−0	ox-101	−1.27		−0.75	−1.14
203	000075−85−4	ox-27	−0.03		0.08	−0.32	278	000110−88−3	ox-102	0.29		0.47	0.65
204	000075−97−8	ox-28	−0.72		−0.71[p]	−0.70	279	000111−14−8	ox-103	−1.66		−1.52	−1.75
205	000076−03−9	ox-29	1.57		0.26	1.03	280	000111−27−3	ox-104	−1.24		−1.12[cv]	−0.99
206	000076−22−2	ox-30	−1.98		−2.44	−2.15	281	000111−55−7	ox-105	0.09		−1.17	−0.56
207	000076−84−6	ox-31	−2.26		−3.50	−2.75	282	000111−70−6	ox-106	−1.84		−1.69	−1.58
208	000076−93−7	ox-32	−2.21		−1.90	−2.33[cv]	283	000111−87−5	ox-107	−2.37		−2.36	−2.24[cv]
209	000077−74−7	ox-33	−0.38		−0.56[p]	−0.78	284	000112−72−1	ox-108	−6.05		−6.02[cv]	−6.36
210	000077−92−9	ox-34	0.90		0.01	0.55	285	000112−92−5	ox-109	−8.39		−7.90	−7.71
211	000078−70−6	ox-35	−1.99		−1.76	−2.19	286	000117−34−0	ox-110	−3.22		−2.58	−2.68
212	000078−83−1	ox-36	0.09		0.15	0.08	287	000117−80−6	ox-111	−6.36		−4.60[cv]	−5.36[cv]
213	000078−84−2	ox-37	0.09		0.06	0.05[p]	288	000118−55−8	ox-112	−3.15		−3.27	−3.57
214	000078−93−3	ox-38	0.53		−0.02[cv]	0.09[p]	289	000118−75−2	ox-113	−2.99		−4.40[p]	−3.08
215	000080−15−9	ox-39	−1.04		−1.68	−1.34	290	000118−90−1	ox-114	−2.06		−1.54	−1.46
216	000080−46−6	ox-40	−2.99		−2.68	−2.82	291	000119−53−9	ox-115	−2.85		−2.71[cv]	−2.91
217	000083−26−1	ox-41	−4.11		−4.08	−4.52	292	000119−61−9	ox-116	−3.12		−3.33	−3.42
218	000083−43−2	ox-42	−2.99		−3.01	−3.12	293	000120−47−8	ox-117	−2.27		−2.04	−1.54[p]
219	000084−61−7	ox-43	−4.92		−4.00	−5.00	294	000120−80−9	ox-118	0.62		−0.00	0.30
220	000084−65−1	ox-44	−5.19		−4.93	−5.48[cv]	295	000121−33−5	ox-119	−1.14		−1.42	−1.36
221	000084−77−5	ox-45	−6.13		−6.29[cv]	−6.22	296	000122−59−8	ox-120	−3.96		−1.50	−1.34[p]
222	000087−66−1	ox-46	0.60		−0.00	0.56	297	000122−88−3	ox-121	−2.32		−2.58	−2.44
223	000088−99−3	ox-47	−1.07		−0.84[p]	−1.41	298	000123−07−9	ox-122	−1.40		−1.47	−1.11
224	000089−78−1	ox-48	−2.53		−1.67	−1.91	299	000123−11−5	ox-123	−1.50		−1.96	−1.76
225	000090−43−7	ox-49	−2.39		−2.70	−2.85	300	000123−19−3	ox-124	−1.55		−1.58	−1.37[cv]

Table 1 (Continued)

no.	CAS no.	set no.	obsd log (mol/L)	nitrogen CNN model[a]	combined CNN model[b]	oxygen CNN model[c]	no.	CAS no.	set no.	obsd log (mol/L)	nitrogen CNN model[a]	combined CNN model[b]	oxygen CNN model[c]
301	000123−31−9	ox-125	−0.18		−0.29	0.00[p]	351	000592−84−7	ox-175	−1.13		−0.99[p]	−0.55
302	000123−86−4	ox-126	−1.18		−1.23	−1.25[cv]	352	000595−41−5	ox-176	−0.85		−0.80[p]	−1.01
303	000123−92−2	ox-127	−1.81		−1.45	−1.66	353	000597−49−9	ox-177	−0.84		−0.91	−1.05
304	000123−96−6	ox-128	−1.69		−1.95	−1.92[cv]	354	000598−53−8	ox-178	−0.06		−0.02	0.02[p]
305	000124−04−9	ox-129	−2.63		−0.83	−1.54[cv]	355	000600−36−2	ox-179	−1.22		−0.84	−1.17[cv]
306	000124−13−0	ox-130	−2.36		−2.49	−2.46[cv]	356	000601−75−2	ox-180	0.73		0.21	0.71
307	000124−19−6	ox-131	−3.17		−3.08	−3.03[cv]	357	000614−61−9	ox-181	−2.16		−1.63	−1.87[p]
308	000124−83−4	ox-132	−1.42		−1.48[cv]	−1.30	358	000616−62−6	ox-182	0.68		−0.03	0.26[p]
309	000124−94−7	ox-133	−3.69		−2.83	−3.12[p]	359	000621−82−9	ox-183	−2.41		−1.89[p]	−2.50
310	000126−07−8	ox-134	−4.61		−5.61[cv]	−4.64	360	000623−37−0	ox-184	−0.80		−0.94	−0.89[cv]
311	000133−43−7	ox-135	0.77		0.44	0.72	361	000623−42−7	ox-185	−0.83		−0.81	−0.45
312	000133−91−5	ox-136	−3.31		−2.34	−3.20	362	000626−93−7	ox-186	−0.87		−0.89[cv]	−0.91
313	000135−19−3	ox-137	−2.28		−2.16[p]	−2.46	363	000627−08−7	ox-187	−1.34		−1.28	−1.19
314	000137−32−6	ox-138	−0.47		−0.25[p]	−0.24	364	000628−63−7	ox-188	−1.88		−1.95[cv]	−1.73
315	000138−52−3	ox-139	−0.85		−0.61	−0.78	365	000821−09−0	ox-189	−0.18		−0.46	−0.27
316	000140−10−3	ox-140	−2.43		−1.89	−2.50	366	001163−19−5	ox-190	−7.58		−8.27	−7.57
317	000141−76−4	ox-141	−0.43		−0.23	−0.44	367	001185−33−7	ox-191	−1.13		−0.41	−0.81[cv]
318	000141−78−6	ox-142	−0.02		−0.17	−0.17	368	001260−17−9	ox-192	−2.58		−2.72	−2.55
319	000141−82−2	ox-143	0.87		0.65	0.86	369	001634−04−4	ox-193	−0.24		−0.07	−0.40
320	000142−62−1	ox-144	−1.05		−0.97	−0.99	370	001861−32−1	ox-194	−5.82		−5.97	−5.77
321	000142−68−7	ox-145	−0.03		0.03	−0.08	371	001918−00−9	ox-195	−1.69		−2.91[cv]	−2.68
322	000142−92−7	ox-146	−2.45		−2.48[cv]	−1.92	372	002432−90−8	ox-196	−6.56		−6.94[p]	−6.51
323	000142−96−1	ox-147	−1.99		−2.83	−2.43	373	002675−77−6	ox-197	−4.41		−4.08	−4.41[cv]
324	000143−07−7	ox-148	−4.62		−4.04	−4.46[p]	374	002976−74−1	ox-198	−2.81		−2.60	−2.47
325	000144−62−7	ox-149	0.39		0.74	1.33[p]	375	003268−87−9	ox-199	−1.28		−8.51[p]	−8.13[p]
326	000147−71−7	ox-150	0.84		0.77	1.14	376	003307−39−9	ox-200	−2.13		−2.63	−2.03
327	000262−12−4	ox-151	−5.31		−4.70	−5.12	377	003724−65−0	ox-201	0.00		0.02	−0.02
328	000334−48−5	ox-152	−3.45		−3.07[cv]	−3.56	378	003970−62−5	ox-202	−1.15		−0.81[cv]	−1.11
329	000464−07−3	ox-153	−0.62		−0.28	−0.70[cv]	379	004798−44−1	ox-203	−0.60		−0.78[p]	−0.63
330	000505−48−6	ox-154	−1.17		−1.81	−1.45[cv]	380	006032−29−7	ox-204	−0.30		−0.41	−0.49[cv]
331	000507−70−0	ox-155	−2.32		−1.56	−1.83	381	006915−15−7	ox-205	0.64		0.55	0.83
332	000512−69−6	ox-156	−0.41		−0.20	−0.64[cv]	382	014187−32−7	ox-206	−4.69		−5.38[p]	−4.77
333	000516−05−2	ox-157	0.76		0.40	0.79	383	015687−27−1	ox-207	−1.93		−3.00[cv]	−2.76
334	000526−75−0	ox-158	−1.43		−1.47	−1.41	384	017199−29−0	ox-208	−0.19		−0.62	−0.60
335	000534−59−8	ox-159	0.44		−0.50	−0.40[p]	385	017348−59−3	ox-209	−2.37		−1.09[p]	−1.68
336	000553−90−2	ox-160	−0.29		−0.26	0.29[p]	386	022204−53−1	ox-210	−4.16		−3.95[p]	−3.77
337	000557−17−5	ox-161	−0.39		−0.15	−0.08	387	029446−15−9	ox-211	−7.23		−7.17	−7.09
338	000563−80−4	ox-162	−0.15		−0.25	−0.12	388	030746−58−8	ox-212	−8.77		−7.69	−8.13
339	000565−60−6	ox-163	−0.72		−0.76[p]	−0.71	389	033857−26−0	ox-213	−7.83		−7.53	−7.73
340	000565−61−7	ox-164	−0.68		−0.86	−0.52	390	036330−85−5	ox-214	−5.06		−4.33	−4.75
341	000565−67−3	ox-165	−0.71		−0.68	−0.82	391	036653−82−4	ox-215	−7.26		−7.06	−7.14
342	000565−80−0	ox-166	−1.30		−1.04	−1.01	392	038964−22−6	ox-216	−7.18		−7.38	−7.75[p]
343	000575−89−3	ox-167	−3.01		−2.61	−2.84	393	039227−53−7	ox-217	−5.72		−5.69	−5.90
344	000576−26−1	ox-168	−1.31		−1.39[cv]	−1.61[p]	394	039227−54−8	ox-218	−5.84		−6.19	−5.97
345	000579−44−2	ox-169	−2.82		−2.73	−2.89	395	039227−58−2	ox-219	−7.53		−7.22	−7.71
346	000584−02−1	ox-170	−0.23		−0.36	−0.48	396	000445−29−4	ox-220	−1.29		−1.47	−1.52
347	000588−22−7	ox-171	−2.68		−3.14	−2.74	397	000455−38−9	ox-221	−1.97		−1.61	−1.94
348	000589−38−8	ox-172	−0.83		−1.11[p]	−0.73[p]	398	000456−22−4	ox-222	−2.07		−1.89[cv]	−2.09
349	000589−82−2	ox-173	−1.62		−1.49	−1.36	399	000611−71−2	ox-223	0.04		−1.35	−0.80
350	000590−36−3	ox-174	−0.50		−0.52	−0.79							

[a] Nitrogen model is type III with 7−6−1 architecture. [b] Combined model is type III with 11−5−1 architecture. [c] Oxygen model is type III with 11−5−1 architecture. [d] This compound was removed from the nitrogen set because it was >1.5 log units lower than the next highest log *S* value. It was kept in the combined set. [cv] Denotes compound was in cross-validation set. [p] Denotes compound was in external prediction set.

prediction set. Once a model was chosen, it was validated using the external prediction set compounds to show the general predictive ability of aqueous solubility (log S), then plotted for visual inspection.

Linear Feature Selection /Nonlinear Model Formation. The descriptors from the best type I models were used to build a three-layer, fully connected, feed-forward CNN, or type II model. Each CNN consisted of an input layer, a hidden layer, and an output layer. The input layer contained a number of neurons equal to the number of descriptors from the best linear model. The output layer contained one neuron representing the predicted log S. The hidden layer neurons varied in number. Two hidden layer neurons were used initially in calculating the rms error of the training and the cross-validation sets. More hidden layer neurons were added until no marked improvement was found in the training set and cross-validation set rms errors. One restriction placed on this process was to keep the ratio of observations in the training set to CNN adjustable parameters greater than two, keeping the possibility of chance correlations low.[51] This architecture was then defined as optimal, and the fully trained network was applied to the external prediction set to demonstrate its ability to generalize.

Network training was optimized using the BFGS (Broyden-Fletcher-Goldfarb-Shanno) quasi-Newton method.[27,52] Overtraining of the network was prevented by periodically observing the cross-validation rms error. When the cross-validation set rms error reached a minimum and started to increase, network training was halted. At this point, despite the continuing decrease in training set rms error, the network

Table 2. Descriptors of the Linear Type I Model for the Nitrogen Data Set

descriptor	type	coefficient	error	range	explanation[a]
constant		8.29	1.91		
WTPT-2	topo	−5.02	0.884	1.67−2.11	molecular ID/number of atoms
MDE-14	topo	8.13×10^{-2}	1.39×10^{-2}	0.00−41.5	distance-edge between all 1° and 4° carbons
EAVE-2	geom	−0.244	4.68×10^{-2}	2.24−12.7	av E-state value over all heteroatoms
GEOM-3	geom	0.908	0.208	0.00−1.57	third geometric moment
PPSA-1	comb	-1.77×10^{-2}	1.71×10^{-3}	68.1−498	partial positive surface area
FPSA-1	comb	7.40	0.893	0.18−0.96	partial positive surface area over all surface area
SCDH-2	comb	0.147	2.58×10^{-2}	0.00−6.83	average surface area times charge on donatable hydrogen

[a] WTPT-2, the sum of unique weighted paths divided by the total number of atoms in the molecule;[36] MDE-14, molecular distance-edge between all primary and quaternary carbons;[39] EAVE-2, the average electrotopological state value over all heteroatoms; GEOM-3, magnitude of the third geometric moment;[41] PPSA-1, $\Sigma(+SA_i)$, summation of positive molecular surface area;[47] FPSA-1, $\Sigma(+SA_i)/SA_{total}$, PPSA-1/total surface area;[47] SCDH-2, $\Sigma(+SA_i \times Q_{don})/H_{tot}$, average surface area times charge on donatable hydrogens.

Table 3. Descriptors of the Linear Type I Model for the Oxygen Data Set

descriptor	type	coefficient	error	range	explanation[a]
constant		6.49	0.397		
ALLP-4	topo	−2.76	0.261	1.09−2.60	total weighted paths/number of atoms
MDE-11	topo	−0.537	0.119	0.00−3.50	distance edge between 1° and 1° carbons
MDE-12	topo	9.36×10^{-2}	1.99×10^{-2}	0.00−23.5	distance edge between 1° and 2° carbons
MDE-14	topo	0.165	3.01×10^{-2}	0.00−22.1	distance edge between 1° and 4° carbons
MDE-34	topo	-3.62×10^{-2}	7.17×10^{-3}	0.00−77.7	distance edge between 3° and 4° carbons
ESUM-2	topo	-1.69×10^{-2}	3.14×10^{-3}	4.69−135	sum of E-state values over all heteroatoms
SHDW-3	geom	8.99×10^{-2}	8.25×10^{-3}	15.2−84.0	shadow area on YZ plane
PNSA-1	comb	-3.15×10^{-2}	2.48×10^{-3}	22.7−285	partial negative surface area
DPSA-1	comb	-1.96×10^{-2}	9.84×10^{-4}	−95.8−845	difference in partial surface areas
WPSA-2	comb	1.25×10^{-3}	2.67×10^{-4}	14.1−4220	surface weighted CPSA
CHDH	comb	2.46	0.211	0.00−2.32	sum of charges on all donatable hydrogens

[a] ALLP-4, the total weighted number of paths in the molecule divided by the total number of atoms;[35] MDE-11, the molecular distance edge between all primary and primary carbons, MDE-12, the molecular distance edge between all primary and secondary carbons, MDE-14, the molecular distance edge between all primary and quaternary carbons, MDE-34, the molecular distance edge between all tertiary and quaternary carbons;[39] ESUM-2, the sum of electrotopological state values over all heteroatoms; SHDW-3, the shadow area projected onto the YZ plane;[42,43] PNSA-1, $\Sigma(-SA_i)$; DPSA-1, $[\Sigma(+SA_i)] - [\Sigma(-SA_i)]$, difference in partial positive and partial negative surface areas;[47] WPSA-2, $[(\Sigma(+SA_i))Q^+_T][SA]/1000$, surface weighted CPSA;[47] CHDH-1, $\Sigma(Q_{\text{donatable H}})$ charge sum on donatable hydrogens.

was starting to memorize the specific characteristics of the training set and losing general predictive ability.

Nonlinear Feature Selection/Nonlinear Model Formation. The final, fully nonlinear model, type III, used a CNN for both descriptor selection and modeling. The best subsets of descriptors found using linear means were not always the best subset when considering a nonlinear relationship between structure and physical property. Therefore, both simulated annealing and genetic algorithm routines were used along with a CNN fitness evaluator to determine the optimal set of descriptors from the reduced pools. Once the best subsets of descriptors were found, they were trained by the same procedures outlined above. After testing several models, the best one was evaluated by the external prediction set compounds to show its ability to generalize.

RESULTS AND DISCUSSION

Type I Models. Generating a linear model is the starting point for a quantitative structure−property relationship development, and its success shows that the descriptors being used are adequately encoding the structures in the data set. In addition, the linear model provides information with which to proceed in nonlinear model training.

The best type I model for data set 1 contained seven descriptors. Pairwise correlations ranged from −0.295 to 0.755 (mean $|r| = 0.278$) and descriptor T-values were above |4.0|. The training set rms error was 0.715 log units ($r^2 = 0.75$), and the prediction set rms error was 0.661 log units ($r^2 = 0.80$). A list of descriptors with their corresponding coefficients is shown in Table 2.

The type I model for data set 2 contained 11 descriptors. Pairwise correlations ranged from −0.820 to 0.887 (mean $|r| = 0.504$) and descriptor T-values were above |4.0|. Training set rms error was 0.588 log units ($r^2 = 0.92$), and the prediction set rms error was 1.555 log units ($r^2 = 0.57$). A list of descriptors and corresponding coefficients is shown in Table 3. The prediction set contained one compound, octachlorodibenzo-p-dioxin, which was a severe outlier. Without this one compound, the prediction set rms error was reduced to 0.679 log units ($r^2 = 0.89$). At first glance, it was thought that some of the descriptor values for this compound fell outside the descriptor range of the training set compounds. The descriptor values did not exceed these ranges. However, several of the more highly weighted descriptors resulted in large negative values. This had the effect of grossly under-predicting the solubility value. Several other linear models, which had been tested previously, also showed this trend.

The type I model for data set 3 also contained 11 descriptors. Pairwise correlations ranged from −0.761 to 0.728 (mean $|r| = 0.365$), and descriptor T-values were above |4.0|. The training set rms error was 0.691 log units ($r^2 = 0.86$). In the case of data set 3, the total number of data set compounds limited the number of compounds available to the training, cross-validation, and external prediction sets; therefore, the compounds that made up the

Table 4. Descriptors of the Linear Type I Model for the Combined Data Set

descriptor	type	coefficient	error	range	explanation[a]
constant		2.58	0.441		
NSB	topo	−0.133	1.87×10^{-2}	0.00−36.0	number of single bonds
NDB	topo	0.356	4.10×10^{-2}	0.00−6.00	number of double bonds
MDE	topo	6.46×10^{-2}	1.41×10^{-2}	0.00−41.5	distance edge between 1° and 4° carbons
MDE−24	topo	3.83×10^{-2}	7.20×10^{-3}	0.00−66.0	distance edge between 2° and 4° carbons
MDE−34	topo	-3.37×10^{-2}	4.10×10^{-3}	0.00−131	distance edge between 3° and 4° carbons
EAVE−2	topo	−0.186	2.39×10^{-2}	2.24−12.7	av E-state value over all heteroatoms
GEOM-1	geom	-8.47×10^{-2}	1.19×10^{-2}	0.66−48.4	first geometric moment
GEOM-3	geom	1.10	0.160	0.00−2.41	third geometric moment
GRAV-6	geom	−0.306	2.95×10^{-2}	6.99−28.3	cube root gravitation index over atom pairs
RNCG	comb	3.00	0.656	0.07−0.93	relative negative charge
CTDH	comb	0.787	4.10×10^{-2}	0.00−11.0	number of donatable hydrogens

[a] NSB-12, number of single bonds in the molecule; NDB-13, number of double bonds in the molecule; MDE-14, molecular distance edge between all primary and quaternary carbons, MDE-24, molecular distance edge between all secondary and quaternary carbons, MDE-34, molecular distance edge between all tertiary and quaternary carbons;[39] EAVE-2, electrotopological state value over all heteroatoms; GEOM-1, magnitude of the first geometric moment, GEOM-3, magnitude of the third geometric moment;[41] GRAV-6, the cube root of the gravitation index of all heavy atoms over all pairs of atoms;[53] RNCG-1, $Q_{most\ negative}/\Sigma(Q^-_i)$;[47] CTDH-0, count of donatable hydrogens.

cross-validation set were used as a second external prediction set. These cross-validation and prediction sets had rms errors of 0.846 ($r^2 = 0.75$) and 1.233 ($r^2 = 0.62$) log units, respectively. Again, the octachlorodibenzo-p-dioxin was a large outlier and yielded a rms error of 0.834 log units ($r^2 = 0.79$) in the prediction set when removed. A list of the descriptors with corresponding coefficients is shown in Table 4.

Comparing these three models showed that some descriptors from data set 3 model were also in the subset models. The nitrogen and combined models shared the descriptors EAVE-2 and GEOM-3. The oxygen and combined models shared MDE-34, and all three models shared MDE-14. All three contained CPSA and hydrogen bonding descriptors. Training set and prediction set RMS errors of the combined set fell between the RMS errors of the nitrogen and oxygen subsets.

Type II Models. The descriptors from the successful linear models generated above can be provided to nonlinear CNN modeling to generate hybrid linear/nonlinear models for each data set. The seven descriptors used in the linear model for data set 1 were passed to a three-layer, fully connected, feed-forward CNN. Architectures of 7−2−1 to 7−6−1 were tested using a committee of neural network trainings. The output values from five trainings were averaged to decrease the dependence on initial random weights and biases. The average results of these models for each architecture were reviewed and plotted. The best set 1 model contained a 7−5−1 architecture having rms errors of 0.503 ($r^2 = 0.88$), 0.654 ($r^2 = 0.74$), and 0.700 ($r^2 = 0.76$) log units for the training set, cross-validation set, and prediction set, respectively. Training set error was improved 30% compared to the linear model; however, the prediction set error increased only slightly.

The 11 descriptors from the linear model for data set 2 followed the approach above. After testing architectures of 11−2−1 to 11−7−1, an optimal model found yielded a 11−4−1 architecture with training set, cross-validation set, and prediction set rms errors of 0.428 ($r^2 = 0.96$), 0.510 ($r^2 = 0.92$), and 1.490 ($r^2 = 0.56$) log units, respectively. Again, the octachlorodibenzo-p-dioxin proved to be a problematic compound. When removed from the prediction set, the prediction set rms error dropped to 0.675 ($r^2 = 0.89$) log units. Training set error improved by 27%, and the prediction set error improved by 4% over the linear model.

Table 5. Descriptors of the Nonlinear Type III Model for the Nitrogen Data Set

descriptor	type	range	explanation[a]
NN	topo	0−6	number of nitrogens
NSB	topo	0−26	number of single bonds
WTPT-2	topo	1.67−2.11	molecular ID/number of atoms
EAVE−2	geom	2.39−12.7	av E-state values over all heteroatoms
GEOM-1	geom	0.66−16.6	first geometric moment
FPSA-2	comb	0.12−3.00	fractional charged partial surface area
CTDH	comb	0−6	number of donatable hydrogens

[a] NN, the number of nitrogens in the molecule; NSB, the number of single bonds in the molecule; WTPT-2, molecular ID/number of atoms;[36] EAVE−2, the average electro-topological state values over all heteroatoms; GEOM-1, the first major geometric moment;[41] FPSA-2, $[(\Sigma(+SA_i))Q^+_T]/SA_{tot}$;[47] CTDH, count of all donatable hydrogens.

The 11 descriptors from the linear model for data set 3 were tested using architectures from 11 to 2-1 to 11−7−1, and the optimal architecture was 11−4−1. This model gave rms errors of 0.635 ($r^2 = 0.88$), 0.713 ($r^2 = 0.81$), and 1.234 ($r^2 = 0.56$) log units for the training set, cross-validation set, and prediction set, respectively. By removing the octachlorodibenzo-p-dioxin, the prediction set rms error dropped to 0.733 ($r^2 = 0.80$) log units. The training set error improved by 8%, and the full prediction set remained unchanged but improved by 12% over the linear model without octachlorodibenzo-p-dioxin.

Type III Models. The best models in QSPR studies are usually found using fully nonlinear feature selection and model building, type III models. However, the search for good type III models is much more costly in terms of computation, and this is done only after more economic linear and hybrid methods have been tried.

A genetic algorithm routine using a CNN fitness evaluator was applied to a 7−5−1 architecture for data set 1. An extensive search of the reduced descriptor pool yielded several seven-descriptor models. These fully nonlinear models were then trained and tested as above. This gave rms errors of 0.507 ($r^2 = 0.88$), 0.542 ($r^2 = 0.83$), and 0.644 ($r^2 = 0.79$) log units for the training, cross-validation, and prediction sets, respectively. Training set and prediction set rms errors improved by 30% and 3% respectively over the

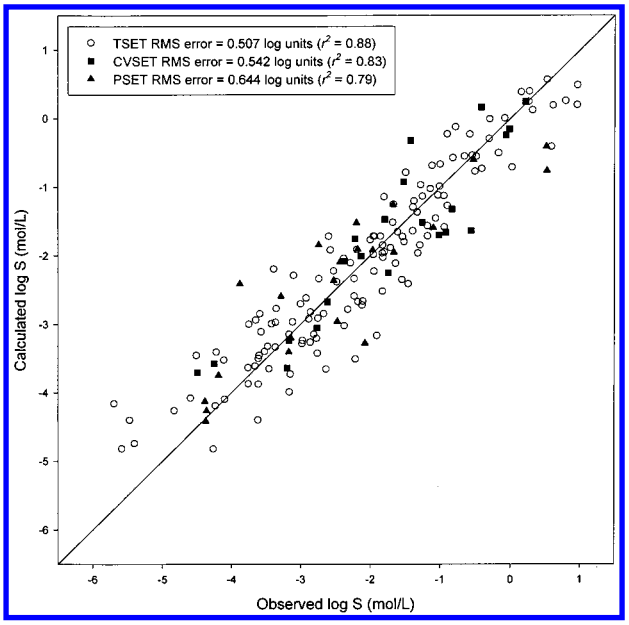

Figure 1. Plot of calculated versus observed log S (mol/L) of the type III model for data set 1, nitrogen compounds: TSET (n = 131), CVSET (n = 22), and PSET (n = 22).

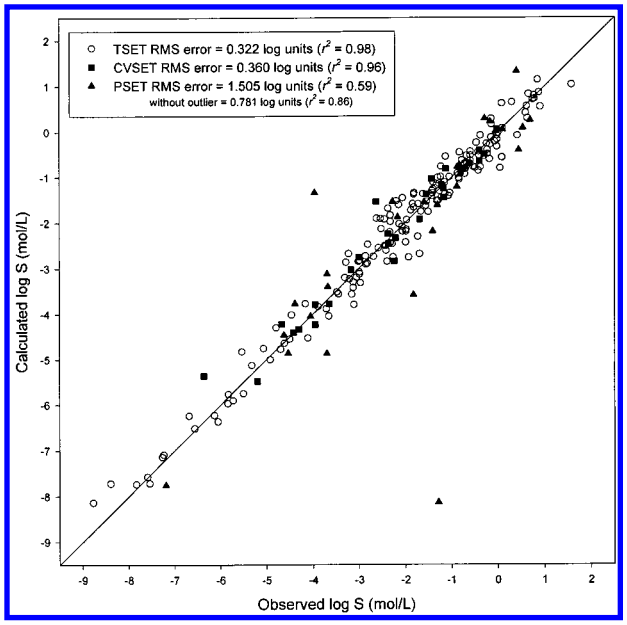

Figure 2. Plot of calculated versus observed log S (mol/L) of the type III model for data set 2, oxygen compounds: TSET (n = 167), CVSET (n = 28), and PSET (n = 28).

type I model. A list of the nitrogen model descriptors is shown in Table 5. A calculated versus observed plot of this model is shown in Figure 1, and the calculated log S (mol/L) values are listed found in Table 1. The fully nonlinear model gave the best overall results for modeling the 175 nitrogen-containing compounds and could therefore predict the aqueous solubility values of other structurally similar compounds with similar accuracy.

For data set 2, an architecture of 11−4−1 was passed to the genetic algorithm/CNN evaluator described above. Several 11-descriptor models were generated, then tested, and trained. The best model, optimally trained using a 11−5−1 architecture yielded a training set error of 0.322 (r^2 = 0.98) log units, a 45% improvement over the type I model. Cross-validation set error was 0.360 (r^2 = 0.96) log units. The prediction set error was 1.505 (r^2 = 0.59) log units or 0.781 (r^2 = 0.86) log units without octachlorodibenzo-p-dioxin. This was a 3% improvement over type I results. For a list of the descriptors for the oxygen model, see Table 6. A calculated versus observed plot of the model is shown in Figure 2, and the calculated log S (mol/L) values are listed in Table 1. Again, the type III model gave the best results for the 223 oxygen-containing compounds and will be able to predict aqueous solubility values of structurally similar compounds.

For data set 3, the 11−4−1 architecture was used with the genetic algorithm/CNN routine. Several 11-descriptor models were generated, tested, and trained. The best model, optimally trained using a 11−5−1 architecture gave a training set error was 0.576 (r^2 = 0.90) log units, representing a 17% improvement over the type I model. Cross-validation set error was 0.587 (r^2 = 0.88) log units. The prediction set error decreased to 1.223 (r^2 = 0.53) log units. This is only a 1% improvement, but when removing the octachlorodibenzo-p-dioxin the error drops to 0.692 (r^2 = 0.82) log units—a 17% improvement over type I results. A list of the descriptors can be seen in Table 7. A calculated versus observed plot of the model is shown in Figure 3, and the calculated log S (mol/L) values are listed in Table 1. These results show that a combined type III model is effective at predicting aqueous solubility values for all 399 compounds and could be used for both oxygen- and nitrogen-containing

Table 6. Descriptors of the Nonlinear Type III Model for the Oxygen Data Set

descriptor	type	range	explanation[a]
KAPA-3	topo	0.67−18.8	Kappa index
NC	topo	2−32	number of carbons
NO	topo	1−16	number of oxygens
MDE−44	topo	0.0−77.8	distance edge between 4° and 4° carbons
ELOW-1	topo	1.21−9.86	through-space distance between max. and min. atomic E-state values
SHDW-3	geom	15.2−84.0	shadow area projected to YZ plane
SHDW-6	geom	0.42−0.82	standardized SHDW-3
GRAV-3	geom	6.56−15.7	cube root gravitation index heavy atoms
DPSA-3	comb	6.81−141	difference in charges partial surface areas
CHAA-2	comb	−0.35 to −0.08	charges on acceptor atoms
RDTA	comb	0−1	ratio of donatable hydrogens to acceptor atoms

[a] KAPA-3, Kappa index of three-bond counts;[34] NC, number of carbons in the molecule; NO, the number of oxygens in the molecule; MDE-44, distance-edge between all quaternary carbons;[39] ELOW-1, the through-space distance between maximum and minimum electrotopological state values; SHDW-3, shadow area of the molecule projected onto the YZ plane, SHDW-6, standardized shadow area of the molecule projected onto the YZ plane;[42,43] GRAV-3, cube root of the gravitational index for all heavy atoms;[53] DPSA-3, $[\Sigma(+SA_i)(Q^+_i)] - [\Sigma(-SA_i)(Q^-_i)]$;[47] CHAA-2, $\Sigma(Q_i)_{acc}/N_{acc}$; RDTA, N_{don-H}/N_{acc}.

Table 7. Descriptors of the Nonlinear Type III Model for the Combined Data Set

descriptor	type	range	explanation[a]
KAPA-6	topo	0.0−18.1	Kappa index − atom corrected
NO	topo	0−16	number of oxygens
NN	topo	0−6	number of nitrogens
NDB	topo	0−6	number of double bonds
WTPT-3	topo	2.35−43.2	sum of path wts from heteroatoms
MDE−44	topo	0.0−83.3	distance edge between 4° and 4° carbons
SYMM-25	topo	0.05−1.0	geometrical symmetry
GEOM-1	geom	0.66−48.4	first geometric moment
DPSA-2	comb	64.3−4334	difference in partial surface areas
CHDH-1	comb	0.0−2.32	charge on donatable hydrogens
SAAA-3	comb	0.10−62.1	surface area of acceptor atoms

[a] KAPA-6, Kappa index of three-bond counts, corrected for atom type;[34] NO, the number of oxygens in the molecule; NN, the number of nitrogens in the molecule; NDB, number of double bonds in the molecule; WTPT-3, sum of all path weights starting from heteroatoms;[36] MDE-44, distance-edge between all quaternary carbons;[39] SYMM-25, geometric consideration of (number of unique atoms/total atoms); GEOM-1, first major geometric moment;[41] DPSA-1, $[\Sigma(+SA_i)] - [\Sigma(-SA_i)]$;[47] CHDH-1, sum of charges on donatable hydrogens; SAAA-3, $\Sigma(SA_{acc})/SA_{tot}$.

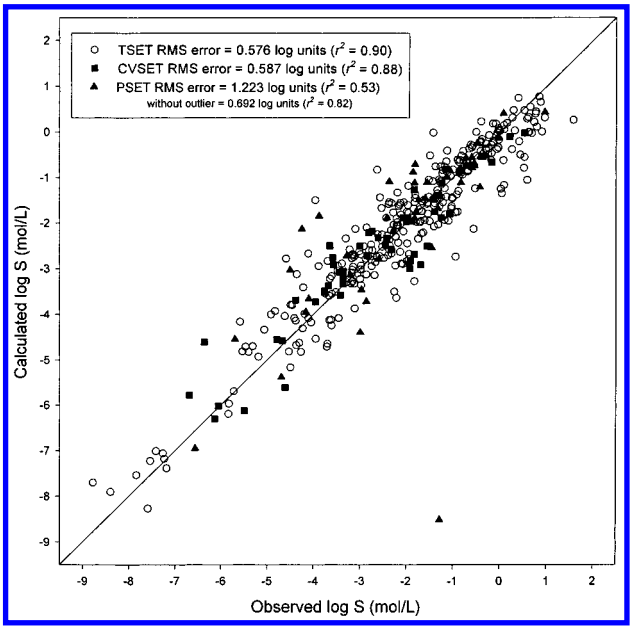

Figure 3. Plot of calculated versus observed log S (mol/L) of the type III model for data set 3, all compounds: TSET (n = 298), CVSET (n = 50), and PSET (n = 51).

compounds of similar structure. Depending on the structural diversity of a data set of similar compounds, one could use the combined model or a subset model for solubility prediction.

Randomizing Experiments. To guarantee that the model results in linear and nonlinear training were not due to chance correlations, a randomizing experiment was performed. The dependent variables for all three data sets were randomly scrambled. Using the same procedures outlined above for model training and testing, models of the identical sizes and architectures of those deemed best were trained and tested. The results showed that the rms errors for training set, cross-validation set, and prediction set were much higher than those of the reported models above.

Results of these experiments for the best type III models were as follows: data set 1 gave rms errors of 1.029 (r^2 = 0.61), 1.165 (r^2 = 0.46), and 1.783 (r^2 = 0.03) log units for the training set, cross-validation set, and prediction sets respectively; data set 2 rms errors were 1.323 (r^2 = 0.61), 1.628 (r^2 = 0.38), and 2.221 (r^2 = 0.03) log units for the training set, cross-validation set, and prediction set respectively; data set 3 rms errors were 1.486 (r^2 = 0.38), 1.604 (r^2 = 0.36), and 2.031 (r^2 = 0.007) log units for the training set, cross-validation set, and prediction set, respectively. For all three data sets, rms errors were considerably worse with randomizing experiments for the training, cross-validation, and prediction sets. A visual inspection of the calculated versus observed plots revealed a scattered distribution over the network averages. The r^2 values for the prediction sets of all three randomized models were extremely low, which proved that the predictive ability of the scrambled models was almost completely random. From these scrambling experiments, it was concluded that the best models presented above were valid in their predictive ability and not due to chance effects.

Comparison of Models. Many of the descriptors in all three models give some insight into the solute−solvent interaction when placing organic compounds in water. The concept that larger molecules are less soluble than smaller molecules, or "like dissolves like", are generalizations first learned in freshman chemistry. More complete consideration is needed to help describe these interactions. Branching information held in κ indices (KAPA-3, KAPA-6), weighted paths (WTPT), and distance-edge descriptors allow for more topological detail. Molecular shape, which affects packing and solvent interactions, can be described through geometry dependent descriptors (SHDW, GEOM, GRAV). The charged partial surface areas of the molecule, which play an important role in solvent−solute interactions, are described with CPSA and hydrogen bonding descriptors (DPSA, FPSA, CDTH, SAAA, CHAA). One descriptor by itself may not adequately explain these interactions but combined in linear and nonlinear models as a whole can create valid and robust predictive models.

One of the goals of this study was to discover if any information could be gained by modeling compounds based on heteroatom content. Clearly, the models from above show that there are different optimal architectures and descriptors used to model the subsets compared to the larger set. Identical descriptors can be found in the combined set model and in the subset models. For example, both the data set 1 model and data set 3 model share NN, the number of nitrogens, and GEOM-1, the first geometric moment. Similar descriptor families, such as weighted paths and bond counts, are found in both models.

The set 2 and set 3 models share NO, the number of oxygens, and MDE-44, distance-edge between quaternary carbons. The also have information from the κ index. All three models share classes of descriptors, such as atom counts, CPSA descriptors, and hydrogen bonding descriptors.

These models included no factors to account for the physical states of the molecules, such as melting point. The inclusion of an indicator variable for each molecule (i.e. liquid = 1; solid = 2) in the descriptor pools did not aid in training and validation. Although this descriptor was not highly correlated with any other in the reduced pools, none of the linear or nonlinear modeling algorithms chose this descriptor as being useful for predicting aqueous solubility.

Another concern in modeling was the simplified assumption that molecules were in their neutral state, which in actuality was not so. Molecules interacting with water can be protonated or deprotonated, changing the electronic and geometric environment. To study this possible effect on solubility prediction, compounds that contained either $-NH_2$ or −COOH groups, or both, were singled out. By calculating the rms errors and r^2 values for these compounds alone, it was determined that these groups did not contribute negatively to the overall model results. Furthermore, having these functional groups did not lead to any systematic over- or underprediction of aqueous solubility. Visual inspection of the calculated versus observed plots showed a similar distribution for these particular compounds as that of the whole model, and rms errors were only slightly higher than the best models shown above.

CONCLUSIONS

Two subsets of organic compounds, separated on the basis of heteroatom types, were modeled using linear and nonlinear processes and compared to a combined set model. Throughout the process, several individual descriptors and descriptor families were shared between the subset and combined models, while greater differences were seen in descriptor choice between the subset models. Training and cross-validation set rms errors were improved over type I using type III models, and in two cases the prediction set rms error was improved. Overall, nonlinear modeling gave the best results. This study indicated that aqueous solubilities for a wide range of compounds could be predicted accurately based solely on molecular structure, with no corrective factor for physical state or the use of other data. Likewise, assuming a neutral molecular state does not degrade the predictive ability of the models.

ACKNOWLEDGMENT

We thank DIPPR (Design Institute for Physical Property Data) Project 931 for funding of this study.

REFERENCES AND NOTES

(1) 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.

(2) 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.

(3) Bakken, G. A.; Jurs, P. C. Prediction of Hydroxyl Radical Rate Constants from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1064−1075.

(4) Bakken, G. A.; Jurs, P. C. Prediction of Methyl Radical Addition Rate Constants from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 508−514.

(5) Goll, E. S.; Jurs, P. C. Prediction of the Normal Boiling Points of Organic Compounds form Molecular Structures with a Computational Neural Network Model. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 974−983.

(6) 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.

(7) Johnson, S. R.; Jurs, P. C. Prediction of the Clearing Temperatures of a Series of Liquid Crystals from Molecular Structure. *Chem. Mater.* **1999**, *11*, 1007−1023.

(8) McClelland, H. E.; Jurs, P. C. Quantitative Structure−Property Relationships for the Prediction of Vapor Pressure of Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 967−975.

(9) Kauffman, G. W.; Jurs, P. C. Prediction of Inhibition of the Sodium Ion-Proton Anitporter by Benzoylguanidine Derivatives from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 753−761.

(10) Patankar, S. J.; Jurs, P. C. Prediction of IC50 Values for ACAT Inhibitors from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 706−723.

(11) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of Aqueous Solubility of Organic Molecules by the Group Contribution Approach. Application to the Study of Biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474−482.

(12) Kühne, R.; Ebert, R.-U.; Kleint, F.; Schmidt, G.; Schüürmann, G. Group Contribution Methods to Estimate Water Solubility of Organic Chemicals. *Chemosphere* **1995**, *30*, 2061−2077.

(13) Ruelle, P.; Kesserling, U. W. Prediction of the aqueous solubility of proton-acceptor oxygen-containing compounds by the mobile order solubility model. *J. Chem. Soc., Faraday Trans.* **1997**, *93*, 2049−2052.

(14) Ruelle, P.; Kesserling, U. W. Solubility Prediction of n-fatty Alcohols and Sterols in Complexing and Noncomplexing Solvents According to the Mobile Order Theory. *Can. J. Chem.* **1998**, *76*, 553−565.

(15) Makino, M. Predictions of Aqueous Solubility Coefficients of Polychlorinated Biphenyls by Use of Computer-Calculated Molecular Properties. *Environ. Int.* **1998**, *24*, 653−663.

(16) Huibers, P. D. T.; Katritzky, A. R. Correlation of the Aqueous Solubility of Hydrocarbons and Halogenated Hydrocarbons with Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 283−292.

(17) Abraham, M. H.; Le, J. The Correlation and Prediction of the Solubility of Compounds in Water Using an Amended Solvation Energy Relationship. *J. Pharm. Sci.* **1999**, *88*, 868−880.

(18) Bodor, N.; Huang, M.-J. A New Method for the Estimation of the Aqueous Solubility of Organic Compounds. *J. Pharm. Sci.* **1992**, *81*, 954−960.

(19) Huuskonen, J.; Salo, M.; Taskinen, J. Neural Network Modeling for Estimation of the Aqueous Solubility of Structurally Related Drugs. *J. Pharm. Sci.* **1997**, *86*, 450−454.

(20) Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous Solubility Prediction of Drugs Based on Molecular Topology and Neural Network Modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450−456.

(21) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.

(22) *Handbook of Physical Properties of Organic Chemicals*; Howard, P. H., Meylan, W. M., Eds.; CRC Press: 1997.

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

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

(25) 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.

(26) 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.

(27) Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. Quantitative Structure−Activity Relationships For Toxicity Of Phenols Using Regression Analysis And Computational Neural Networks. *Environ. Toxicol. Chem.* **1994**, *13*, 841−851.

(28) Stewart, J. P. P. *MOPAC 6.0, Quantum Chemistry Program Exchange*; Program 455; Indiana University: Bloomington, IN, 1990.

(29) Stewart, J. J. P. MOPAC: A semiempirical molecular orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−105.

(30) Dewar, M. J. S.; Zoebisch, E. G.; Healey, E. F.; Stewart, J. P. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(31) Alemán, C.; Luque, F. J.; Orozco, M. Suitability of the PM3-Derived Molecular Electrostatic Potentials. *J. Comput. Chem.* **1993**, *14*, 799−808.

(32) Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109−116.

(33) Kier, L. B. Distinguishing Atom Differences in a Molecular Graph Shape Index. *Quant. Struct.-Act. Relat.* **1986**, *5*, 7−12.

(34) Kier, L. B. Shape Indexes of Orders One to Three from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1986**, *5*, 1−7.

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

(36) Randic, M. On Molecular Identification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164−175.
(37) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(38) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; John Wiley & Sons: New York, 1986; pp 18−20.
(39) Cao, C. Distance-Edge Topological Index: Research on Structure−Property Relationship of Alkanes. *Huaxue Tongao* **1996**, *54*, 533−538.
(40) Pearlman, R. S. In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker: New York, 1980; Chapter 10.
(41) Goldstein, H. *Classical Mechanics*; Addison-Wesley: Reading, MA, 1950; pp 144−156.
(42) 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.
(43) 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*, 1048−1054.
(44) Dixon, S. L.; Jurs, P. C. Atomic Charge Calculations for Quantitative Structure−Property Relationships. *J. Comput. Chem.* **1992**, *13*, 492−504.
(45) Miller, K. J.; Savchik, J. A. A New Empirical Method To Calculate Average Molecular Polarizabilities. *J. Am. Chem. Soc.* **1979**, *101*, 7206−7213.
(46) Abraham, R. J.; Smith, P. E. Charge Calculations in Molecular Mechanics IV: A General Method for Conjugated Systems. *J. Comput. Chem.* **1988**, *9*, 288−297.
(47) 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.
(48) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative-Structure Property Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.
(49) Belsley, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics*; John Wiley & Sons: New York, 1980.
(50) Draper, N. R.; Smith, H. *Applied Regression Analysis*; John Wiley & Sons: New York, 1981.
(51) Livingstone, D. J.; Manallack, D. T. Statistics Using Neural Networks: Chance Effects. *J. Med. Chem.* **1993**, *36*, 1295−1297.
(52) 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.
(53) 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.

CI010035Y