Prediction of Aqueous Solubility of Organic Compounds from Molecular Structure

Brooke E. Mitchell and Peter C. Jurs*

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

Received December 18, 1997

Multiple linear regression (MLR) and computational neural networks (CNN) are utilized to develop mathematical models to relate the structures of a diverse set of 332 organic compounds to their aqueous solubilities. Topological, geometric, and electronic descriptors are used to numerically represent structural features of the data set compounds. Genetic algorithm and simulated annealing routines, in conjunction with MLR and CNN, are used to select subsets of descriptors that accurately relate to aqueous solubility. Nonlinear models with nine calculated structural descriptors are developed that have a training set root-mean-square error of 0.394 log units for compounds which span a −log(molarity) range from −2 to +12 log units.

INTRODUCTION

Aqueous solubility is a physical property that has been extensively studied. As a property involving water as the solvent, it is important in a diverse array of situations including pharmaceutical, environmental, and industrial applications. The biological activity of a drug compound is affected by the ability of the drug to be transported and absorbed. Drug design, therefore, must take into account physical property information such as aqueous solubility as well as biological activity.[1,2] The rate and extent of biodegradation is also affected by the aqueous solubility of organic compounds in the environment.[3,4] There is no question that the ability to predict the aqueous solubility of compounds is useful. Many different methods have been developed for the estimation of aqueous solubility with varying success and applicability.[1,17] Some recent developments include a neural network model relating aqueous solubility to topological descriptors,[1] the use of the mobile order solubility model,[2,5] a clustering approach,[11] group contribution approaches,[3,9,13,15] and linear and neural network models based on semiempirical quantum chemical descriptors.[6−8]

The development of a quantitative structure−property relationship (QSPR) can aid in the understanding of aqueous solubility and can provide a method of estimating the aqueous solubility value directly from structure without making an experimental measurement. In a QSPR study, a mathematical model is developed which relates the structures of a set of compounds to a physical property such as aqueous solubility. The underlying assumption in a QSPR is that there is some sort of relationship between the physical property of interest and molecular structure. Many different physical properties have been studied in this manner, including boiling point,[18,19] supercritical CO_2 solubility,[20] and autoignition temperature.[21] In addition to providing a means to predict a physical property without having to actually measure it, a QSPR may also lead to an understanding of the structural features related to the physical property.

The parametric approach employed in this QSPR study involves relating the experimental aqueous solubility values to structure-based descriptors. These descriptors are numerical representations of structural features of molecules that attempt to encode important information that causes structurally different compounds to have different physical property values. The descriptors fall into the three main categories of topological, geometric, and electronic or combinations of these categories such as charged partial surface area (cpsa)[22] and hydrogen bonding descriptors which combine geometric and electronic information. Linear models, which will be referred to as Type 1 models, and nonlinear, Type 2 and Type 3, models are developed that relate aqueous solubility to the structure-based descriptors.

Type 1 models result from feature selection experiments that use multiple linear regression (MLR) to choose the subset of descriptors that provide the relationship to aqueous solubility. Improvement of the model based on linear feature selection can be achieved with the use of computational neural networks (CNN), which can be considered as a nonlinear mathematical function. The results of this type of experiment will be referred to as Type 2 models. Improvements occur both because of the nonlinear nature of the mathematical function and because of the larger number of adjustable parameters that occur in a CNN. Although the results of neural networks are better, the computational time required is much higher than for linear regression analysis. Regression coefficients for a given subset of descriptors can be calculated in a single step, while the optimization of adjustable parameters in a neural network is an iterative process. Finally, Type 3 models, result from experiments in which the CNN is used as a fitness function during feature selection to choose a good subset of descriptors. The subset of descriptors chosen as the best model based on linear criteria will not necessarily be the best subset of descriptors when considering a nonlinear relationship between the descriptors and the aqueous solubility. Type 3 models are developed that use the nonlinear CNN as the determining factor of the quality of a subset of descriptors.

S0095-2338(97)00117-0 CCC: $15.00 © 1998 American Chemical Society
Published on Web 04/03/1998

EXPERIMENTAL SECTION

This QSPR study was performed using the Automated Data Analysis and Pattern Recognition Toolkit (ADAPT)[23,24] as well as genetic algorithm,[25] simulated annealing,[26,27] and computational neural network[28] routines developed at Penn State. The computations were performed on DEC 3000AXP Model 500 and DEC Alpha station 500/500 workstations.

The general steps involved in performing a QSPR study using the ADAPT methodology included compilation of a data set, calculation of descriptors to numerically encode the structural features of the data set compounds, feature selection to choose a small subset of descriptors that relate molecular structure to aqueous solubility, and validation of the models that were developed.

Selection of the data set involved choosing a large set of compounds with accurately measured experimental aqueous solubility values. In the case of this study, the set of compounds used was inspired by the results of an aqueous solubility study that was presented by Lee et al. and Verlin et al. at the 210th American Chemical Society National Meeting.[29,30] The data set that was used in the study by Lee et al. consisted of 353 compounds from the AQUASOL database of Yalkowsky.[31] The 332 compounds compatible with ADAPT were chosen as the basis of this QSPR study. The ADAPT software system allows compounds with between 2 and 46 non-hydrogen atoms, up to 100 hydrogen atoms, and up to 51 bonds. The compounds used in this study contained carbon, hydrogen, oxygen, nitrogen, sulfur, chlorine, fluorine, bromine, and iodine and are listed with their experimental and calculated aqueous solubility values in Table 1. The molecular weights of the compounds ranged from 30 to 547. Aqueous solubility values were expressed as −log(molarity) values in order to compress the range covered by the data which was from −1.57 to 12.8 log units. This data set was much larger and more diverse than previous aqueous solubility studies performed using the ADAPT methodology.[16,17]

The 332 compound data set was divided into a training set (tset) of 300 compounds and an external prediction set (pset) of 32 compounds. The tset was used for the development of Type 1 linear regression models and then further divided into a tset of 265 compounds and a cross-validation set (cvset) of 30 compounds for use with CNN. Five compounds were flagged as outliers and removed from the 300-compound tset during the linear regression experiments. The cvset was used to determine when to stop training the neural network so that the network would have good, general predictive ability. The same pset was used for MLR and CNN for comparison purposes. The specific compounds in the pset and cvset were chosen randomly, but they were chosen to be representative of the types of compounds in the data set and to span the entire range of the data. These compounds are denoted by superscripts in Table 1.

The compounds were sketched into HyperChem as two-dimensional structures. Preliminary molecular modeling was performed in HyperChem to generate three-dimensional representations of the compounds. The three-dimensional structures were refined using MOPAC,[32] a semiempirical molecular orbital modeling routine, with the PM3 Hamiltonian to put the compounds into their energy-minimized conformations. Accurate three-dimensional representations of the structures were necessary for the generation of descriptors dependent on geometry.

ADAPT routines were used to calculate descriptors to numerically represent the topological, geometric, and electronic structural features of the data set compounds. The descriptors were calculated directly from the energy-minimized three-dimensional conformations, with the goal of encoding the structural features that give rise to differences in aqueous solubility for different compounds. The topological descriptors that were calculated included counts of atom types, bond types, and functional group types, molecular distance-edge indices,[33] and molecular connectivity indices[34] to encode the size and degree of branching in the compounds. Geometric descriptors included solvent-accessible molecular surface area and volume,[35] moments of inertia, shadow area projections,[36,37] and gravitational indices.[38] The electronic environment of compounds was described through the use of partial atomic charges,[39,40] providing descriptors such as the charge on the most positive or negative atom. Some descriptors also combined geometric and electronic information, such as the cpsa descriptors[22] and hydrogen bonding descriptors. Because aqueous solubility was being studied, the hydrogen bonding descriptors were calculated assuming a mixed solution of solute and water to take into account the effects that water would have on the hydrogen bonding of the system. In addition to the descriptors calculated in ADAPT routines, the heat of formation, electronic energy, and ionization potentials that were calculated during the MOPAC energy minimization and the theoretical linear solvation energy relationship (TLSER) descriptors reported by Lowrey et al.[41] were included in the pool of descriptors used for development of QSPR models. A total of 210 descriptors were calculated for each compound. The intent is to encode the compounds' structures as completely as possible by generating a spectrum of descriptor types that capture different features of the structure. The types of interactions these compounds can have with water in aqueous solution is captured, and this is demonstrated by our development of excellent models using the descriptors.

The next step was to use various forms of feature selection to choose a small subset of descriptors that accurately relates the molecular structure to aqueous solubility. The first form of feature selection was objective feature selection, which was based only on the information content of the descriptors. The dependent variable, aqueous solubility, was not used. Subjective feature selection, or model development, used the dependent variable to find subsets of descriptors that provide a good relationship with the aqueous solubility.

All of the descriptors were subjected to objective feature selection to remove those that did not contribute useful information to the pool. Pairwise correlations between descriptors were examined so that only one descriptor was retained from a pair contributing similar information (correlation coefficients ≥ 0.95), and descriptors with greater than 90% identical values were dropped since those descriptors were not encoding the structural differences between compounds that accounts for their different aqueous solubility values. Objective feature selection left a reduced pool of 122 descriptors for the 300 compound tset, well below the cutoff of 0.6 for the ratio of descriptors to observations used

Table 1. Compounds Used in the Aqueous Solubility Study with Their Experimental and Calculated −log(molarity) Values[d]

compound name	expt AqSol	calc AqSol	compound name	expt AqSol	calc AqSol
ethane	2.73	2.82	9-methylanthracene	5.87	5.48
propane	2.84	2.65	9,10-dimethylanthracene	6.57	5.98
butane	2.95	2.91	perylene	8.80	8.60
2-methylbutane	3.18	3.29	benz[*g*,*h*,*i*]perylene	9.03	8.92
pentane	3.25	3.52	1-butanol	0.0209	−0.111
2,2-dimethylbutane	3.62	3.59	1-pentanol	0.609	0.422
2,3-dimethylbutane[a]	3.63	3.66	1-hexanol	1.23	1.12
2-methylpentane[b]	3.79	3.88	D-mannitol	−0.0569	0.0136
3-methylpentane	3.78	3.79	1-heptanol	1.84	1.78
hexane	3.91	4.06	1-octanol	2.37	2.52
2,4-dimethylpentane	4.36	4.27	isobutanol	−0.0864	−0.466
3-methylhexane[b]	4.47	4.42	2-methyl-1-butanol[a]	0.470	−0.0512
heptane	4.55	4.70	3-pentanol	0.236	0.270
2,3,4-trimethylpentane[a]	4.83	4.18	*tert*-pentanol	0.0273	0.0985
2,2,4-trimethylpentane	4.77	4.55	4-methyl-2-pentanol	0.785	0.942
octane	5.26	5.39	2,4-dimethyl-3-pentanol	0.889	1.35
nonane	5.92	5.98	2-octanol	1.69	1.92
decane[a]	6.55	6.65	2-butanone[b]	−0.525	−0.549
isobutane[b]	3.08	2.78	4-methyl-2-butanone[a]	0.731	0.849
2,2-dimethylpentane	4.36	4.32	3,3-dimethyl-2-butanone	0.731	0.810
2,3-dimethylpentane	4.28	4.17	oxalic acid[a]	−0.389	−0.885
2-methylhexane	4.60	4.54	malonic acid	−0.762	−0.780
3,3-dimethylpentane	4.23	4.16	succinic acid	0.188	0.265
3-methylheptane	5.16	5.15	pentanoic acid	0.510	0.00511
2,2,5-trimethylhexane	5.18	5.28	suberic acid[b]	1.29	2.22
4-methyloctane[b]	6.05	5.79	dodecanoic acid	4.79	4.65
1-propene	2.20	1.66	ethyl acetate	0.0227	−0.0369
1-octene	4.52	3.93	isopropyl acetate	0.602	0.125
1-pentyne	1.76	1.80	isobutyl acetate	1.24	1.04
1-hexyne	2.20	2.36	butyl acetate[a]	1.18	0.743
benzene	1.64	1.85	pentyl acetate[a]	1.88	1.31
toluene[b]	2.22	2.20	diethyl ether	−0.408	0.339
ethylbenzene	2.78	2.54	methyl *tert*-butyl ether	0.235	0.556
m-xylene	2.83	2.97	propyl isopropyl ether[b]	1.34	1.32
o-xylene	2.77	2.80	isopropyl *tert*-butyl ether	2.37	2.06
p-xylene	2.80	2.85	dibutyl ether	1.99	2.01
1,2,3-trimethylbenzene	3.22	3.50	styrene	2.58	2.14
(1-methylethyl)benzene	3.30	3.24	phenol	0.0334	0.471
1,3,5-trimethylbenzene	3.34	3.82	hydroquinone	0.168	0.555
propylbenzene	3.25	3.06	*m*-cresol	0.721	0.704
1,2,4-trimethylbenzene	3.33	3.53	*o*-cresol	0.635	0.645
naphthalene[b]	3.58	3.48	2,4-dimethylphenol[a]	1.22	1.37
1-naphthol	2.39	1.62	*o*-phenylphenol[a]	2.39	2.07
2-naphthol	2.49	1.36	furfural	0.0757	0.736
butylbenzene[b]	3.78	3.82	benzaldehyde	1.26	1.12
sec-butylbenzene[a]	3.93	3.70	acetophenone[b]	1.32	1.64
tert-butylbenzene	3.69	3.71	benzoic acid	1.57	1.07
1-methylnaphthalene	3.69	3.60	phthalic acid[b]	1.15	1.27
2-methylnaphthalene	3.75	3.75	phenylacetic acid	0.886	1.34
dibenzofuran	4.46	4.65	*m*-toluic acid	2.14	1.19
dibenzo-*p*-dioxin	5.33	4.74	diphenyl ether	3.95	3.77
diphenyl	4.34	4.74	D-tartaric acid	−0.837	−0.833
1,5-dimethylnaphthalene	4.69	4.38	DL-tartaric acid	−0.767	−0.432
2,3-dimethylnaphthalene[b]	4.77	4.47	citric acid	−0.896	−0.833
phenylhexane	5.22	5.35	salicylic acid[b]	1.81	0.968
fluorene	4.91	5.37	methylparaben	1.82	1.71
diphenylmethane	4.17	4.87	propylparaben	2.62	2.62
anthracene	6.55	5.95	butylparaben[a]	2.86	3.14
phenanthrene	5.21	5.79	ethylparaben	2.31	1.97
1-methylphenanthrene	5.88	5.51	*p*-hydroxybenzoic acid	1.43	1.19
fluoranthrene	5.96	6.47	D-mandelic acid	−0.0414	0.998
pyrene	6.19	6.62	L-mandelic acid	0.185	1.00
2,3-benzofluorene	8.91	7.56	salicylic acid acetate	1.65	1.42
1,2-benzanthracene	7.19	7.57	cinnamic acid	2.46	1.66
chrysene	8.14	7.73	propionyl-r-mandelic acid[a]	1.60	1.28
naphthacene	8.19	8.02	benzoyl-r-mandelic acid	1.51	2.48
triphenylene[a]	6.88	7.07	1-methylfluorene	5.22	5.22
benzo[*a*]pyrene[b]	7.97	8.26	1,2-benzofluorene[b]	6.68	7.47
9,10-dimethyl-1,2-benzanthracene	6.83	7.56	cyclopentane	2.52	2.69
3-methylcholanthrene	7.94	8.02	cyclohexane	3.14	3.03
1,2:5,6-dibenzanthracene	8.23	8.86	methylcyclopentane[a]	3.30	3.15
coronene	9.38	9.32	methylcyclohexane	3.81	3.59
thiophene	1.45	1.42	1,1,3-trimethylcyclopentane[a]	4.48	4.60
2-ethylthiophene[b]	2.59	2.16	propylcyclopentane	4.74	4.43
1,2,4,5-tetramethylbenzene	4.59	4.04	1,1,3-trimethylcyclohexane	4.85	4.93
pentylbenzene[a]	4.61	4.52	cyclopentene	1.73	2.02
1,3-dimethylnaphthalene	4.29	4.47	1,4-cyclohexadiene[b]	1.97	1.83
1,4-dimethylnaphthalene	4.14	4.38	cyclohexene	2.48	2.48
1-ethylnaphthalene	4.16	3.91	cycloheptatriene	2.15	1.82
2,6-dimethylnaphthalene	4.89	4.65	D-limonene	4.00	3.98
1,4,5-trimethylnaphthalene	4.92	4.91	pentylcyclopentane[b]	6.09	6.26
2-methylanthracene	6.80	5.83	acenaphthene	4.54	4.78

Table 1. Continued

compound name	expt AqSol	calc AqSol
raffinose	0.410	0.328
sucrose[c]	−0.0719	
norethindrone[b]	4.67	4.17
norethindrone acetate	4.79	4.85
quinone	0.879	0.986
testosterone	4.05	3.84
progesterone	4.53	4.55
hydrocortisone acetate	4.60	4.58
hydrocortisone	3.08	3.58
methylprednisolone	2.99	2.72
acrylonitrile[a]	−0.146	−0.344
trichloromethane	1.19	1.20
dichloromethane	0.740	0.745
tetrachloromethane	2.26	2.24
tetrafluoromethane[c]	3.68	
bromomethane	0.851	1.02
dichlorodifluoromethane[a]	2.64	2.18
1,1,2,2-tetrachloroethane	1.76	1.78
methoxychlor	6.68	6.71
trichloroethene	2.04	1.61
1,2-dichloroethene	1.07	1.22
tetrachloroethene	2.57	2.81
1,2,3,5-tetrachlorobenzene	4.73	4.56
1,2,4-trichlorobenzene	3.64	3.84
m-dichlorobenzene	3.04	3.10
o-dichlorobenzene	3.02	2.98
p-dichlorobenzene	3.31	3.16
chlorobenzene	2.42	2.30
2,2′,3,3′,4,4′,5,5′,6-nonachlorobiphenyl[b]	9.93	10.1
2,2′,3,3′,5,5′,6,6′-octachlorobiphenyl	9.30	9.67
2,2′,3,3′,4,4′-hexachlorobiphenyl	9.00	8.28
2,2′,3,3′,6,6′-hexachlorobiphenyl	7.86	7.90
2,2′,4,4′,5,5′-hexachlorobiphenyl	8.57	8.16
2,2′,4,4′,6,6′-hexachlorobiphenyl[a]	8.48	8.11
2,2′,3,4,5-pentachlorobiphenyl	7.10	7.20
2,2′,4,5,5′-pentachlorobiphenyl	7.44	7.38
2,3,4,5,6-pentachlorobiphenyl	7.78	7.69
2,2′,5,5′-tetrachlorobiphenyl	6.44	6.59
2,3,4,5-tetrachlorobiphenyl	7.26	6.93
3,3′,4,4′-tetrachlorobiphenyl	8.68	8.29
2,4,5-trichlorobiphenyl	6.27	6.21
2,4,6-trichlorobiphenyl	6.07	5.99
2,2′-dichlorobiphenyl	5.36	5.11
2,4-dichlorobiphenyl	5.29	5.46
2,5-dichlorobiphenyl	5.27	5.36
4,4′-dichlorobiphenyl[b]	6.63	6.50
2-chlorobiphenyl	4.63	4.57
4-chlorobiphenyl	5.25	5.36
decachlorobiphenyl	1.08	10.6
p,*p*′-DDE	6.85	7.14
p,*p*′-DDT	8.18	7.43
pentachlorobenzene	5.37	5.87
1,2,3,4-tetrachlorobenzene	4.38	4.79
1,2,4,5-tetrachlorobenzene	5.19	5.18
1,2,3-trichlorobenzene	4.08	3.62
1,3,5-trichlorobenzene	4.55	3.76
iodobenzene	3.03	2.57
2,2′,3,3′,4,5-hexachlorobiphenyl[b]	8.04	8.01
2,3,3′,4′,5,6-hexachlorobiphenyl	7.83	8.06
2,2′,5-trichlorobiphenyl	5.65	5.87
2,4,4′-trichlorobiphenyl	5.14	6.34
2,4′-dichlorobiphenyl	5.60	5.40
2,6-dichlorobiphenyl[a]	5.07	5.18
3-chlorobiphenyl	4.88	5.05
hexachlorophene	3.71	4.07
dichlorophen	3.95	3.34
ioxynil	3.61	3.31
m-fluorobenzoic acid	1.97	1.56
o-fluorobenzoic acid	1.29	1.44
4-chlorophenoxyacetic acid	2.32	2.14
(2,4,5-trichlorophenoxy)acetic acid	2.97	3.34
(2,4-dichlorophenoxy)acetic acid[a]	2.51	2.88
2-(2,4,5-trichlorophenoxy)propionic acid	3.31	3.25
3,6-dichloro-2-methoxybenzoic acid	1.69	2.47
4-(2,4-dichlorophenoxy)propionic acid	3.64	3.84
1,2,3,4-tetrachlorodibenzo-p-dioxin	8.77	8.72
2-chlorodibenzo-*p*-dioxin	5.86	6.03
1,2,4-trichlorodibenzo-*p*-dioxin	7.53	7.76
2,3-dichlorodibenzo-*p*-dioxin	7.23	7.59
2,7-dichlorodibenzo-*p*-dioxin	7.83	7.40
1-chlorodibenzo-*p*-dioxin[a]	5.72	5.70
octachlorodibenzo-*p*-dioxin[c]	1.28	
griseofulvin	4.60	4.68
o,*p*′-DDT	6.80	7.05
DDD	6.76	6.65
trichloroacetic acid	−1.57	−1.07
1,2,3,4,5,6-hexachlorocyclohexane	4.60	4.96
nitromethane	−0.255	−0.777
octylamine	2.81	2.94
nitrobenzene	1.80	1.66
aniline	0.416	0.593
benzidine	2.63	2.28
acridine	3.60	3.18
amitrole	−0.522	0.0822
m-dinitrobenzene	2.37	2.53
indole	1.21	1.64
o-phenanthroline	1.82	3.49
2,2′-biquinoline[a]	5.40	5.66
13*H*-dibenzo(*a*,*i*)carbazole	7.41	7.76
picric acid[b]	1.26	1.43
p-nitrophenol	0.900	1.75
styphnic acid	1.66	1.52
m-nitrophenol	1.01	1.68
o-nitrophenol	1.75	1.54
metronidazole	1.22	1.56
L-tyrosine	2.57	1.84
salicylamide	1.79	0.828
o-nitrobenzoic acid[b]	1.35	1.82
L-phenylalanine	0.804	1.76
cinchomeronic acid	1.86	1.34
isocinchomeronic acid	2.14	1.75
lutidinic acid	1.83	1.51
quinolinic acid	1.19	1.36
L-tryptophan	1.23	1.84
hippuric acid	1.68	1.89
acetanilide[a]	1.35	1.46
N,*N*-dimethyl-*N*-phenylurea	1.67	2.15
urea	−0.946	−0.532
glycine	−0.493	−0.462
DL-alanine[b]	−0.243	−0.475
L-alanine[b]	−0.250	−0.480
α-aminobutyric acid	−0.305	−0.189
DL-leucine	1.10	1.25
norleucine[a]	1.06	2.01
L-leucine	0.750	1.26
L-aspartic acid	1.41	0.697
α-aminoisobutyric acid	−0.210	−0.423
L-norleucine	0.975	1.38
meprobamate	1.71	1.65
diphenamid[a]	2.98	3.39
mebendazole[a]	3.88	3.89
carboxin	3.14	3.41
N-(2-methylcyclohexyl)-*N*′-phenylurea	4.11	3.56
triallate	4.88	4.26
tetracycline[c]	3.12	
oxytetracycline[c]	3.14	
caffeine	0.947	1.42
barbital[a]	1.42	1.35
metharbital	1.98	1.74
butabarbital	2.18	1.70
butethal	1.83	2.64
thioopental	3.36	3.03
amobarbital	2.55	2.48
pentobarbital	2.54	2.33
thiamylal[b]	3.68	3.33
theophylline	1.39	0.953
alloxantin	2.23	2.01
vinbarbital	2.43	2.22
allobarbital	2.07	2.10
secobarbital	2.23	2.57
phenobarbital	2.29	1.90
riboflvain	3.68	3.84
lenacil	4.59	4.25
terbacil[b]	2.48	2.15
dicyanodiamide	0.311	−0.0295
CDEC	3.37	3.78
picloram	2.75	2.93
chloramben	2.47	2.22
diuron[a]	3.76	3.35
linuron	3.52	3.61
monuron	2.92	2.93
fluometuron	3.42	3.30
neburon[b]	4.76	4.48
barban[a]	4.37	4.33
cyanazine	3.15	3.37
indomethacin	4.62	4.33
bromacil	2.52	2.57
diallate[a]	4.08	4.07

[a] Prediction set compound. [b] Cross-validation set compound. [c] Outlier compound. [d] The calculated values are from the Type 3 model.

to limit the possibility of finding a relationship based purely on chance effects.[42]

The reduced pool of information-rich descriptors was then examined to find subsets of descriptors that accurately represented the relationship between molecular structure and aqueous solubility, the subjective feature selection portion of the experiment. Simulated annealing and genetic algorithm routines, along with an interactive regression routine, were used to generate statistically valid linear models. The genetic algorithm[25] and simulated annealing[26] routines each investigated a large number of descriptor subsets, with the quality of the model based on the root-mean-square (rms) error and statistical integrity. A variety of subset sizes was investigated to determine the optimum number of descriptors in a model. When adding another descriptor did not improve the statistics of a model, it was determined that the optimum subset size had been achieved. The optimum model size in this study was nine descriptors. Other measures of the quality of a model were the correlation coefficient, the overall *F*-value of the model, and the *T*-values of the individual descriptors in the model. The presence of outliers was detected by looking at regression diagnostics that measured the effect of individual data points on the model, such as standardized and studentized residuals, leverage, and Cook's distance.[43] Compounds were often flagged as outliers if they were not well represented in the data set. Five outliers were flagged and removed from the tset based on the linear model chosen.

The descriptors chosen for the Type 1 models were submitted to CNN for improvement as Type 2 models. A fully connected, feed-forward, three-layer neural network was used. The input layer consisted of as many neurons as there were descriptors in the Type 1 model. The number of neurons in the hidden layer was considered to be optimized when the tset error did not significantly decrease with the addition of another neuron, six hidden-layer neurons in this case. A single neuron in the output layer provided the estimated aqueous solubility. A quasi-Newton BFGS (Broyden−Fletcher−Goldfarb-Shanno)[44−48] algorithm[28] was used to train the network. As mentioned previously, a cvset was used to prevent overtraining. The cvset was a small subset of compounds randomly drawn from the tset which was not included during training but was tested periodically during training. When the cvset rms error was minimized, training was stopped since beyond this point the network was fitting characteristics specific to individual tset compounds rather than general characteristics of the entire data set. The use of a cvset increased the confidence with which external predictions could be made using the trained network.

The CNN results were very dependent upon the starting weights and biases, which were randomly selected. To sample a variety of starting points an automated CNN program was run that trained the neural network from a user-determined number of random starting weights and biases. In this study, 250 random sets of starting weights and biases were studied. The results from these trials were then examined to find a good starting point for the neural network training.

CNN were also used as the basis for feature selection to generate Type 3 models. A genetic algorithm routine[25] was used to investigate subsets of descriptors. The use of the tset/cvset and the method of evaluating the quality of the neural network model was slightly different from the CNN experiments involving the Type 2 model. In the Type 2 experiments a single cvset with approximately 10% of the tset compounds was chosen and used throughout, while in the Type 3 experiments each subset of descriptors was trained multiple times with different compounds in the cvset for each training session. The number of training sessions was determined by the user-determined percentage of compounds placed in the cvset. Each compound in the linear regression tset was placed in the cvset for one and only one network training session. In the aqueous solubility study, 25% of the compounds were placed in the cvset, so four training sessions were performed on each subset of descriptors. The same cvsets are used when investigating different descriptor subsets, and the quality of each model was evaluated by comparing the overall rms error between the predicted and experimental values of all of the tset compounds when they were predicted as cvset members.

Unlike the development of Type 1 models, it was extremely computationally intensive to investigate a large number of different Type 3 models. The architecture for the CNN was retained from the Type 2 models, providing a basis for comparison and to save the computational expense involved in determining the best network architecture. Another method that was used to save computational time was the use of a generalized simulated annealing algorithm[27] to try to optimize the starting weights and biases as opposed to running a large number of training sessions from different starting points. Type 3 models were found that had lower rms errors than the corresponding Type 2 models, proving that the descriptors chosen based on linear criteria were not the best subset that could be found for use with CNN.

All of the models that were developed in the QSPR study were validated with an external prediction set. The models were able to accurately predict the aqueous solubilities of these compounds, which had not been used in the development of the models, with pset rms errors of the same magnitude as the tset and cvset errors. An additional method of validation for all three model types was the visual inspection of both calculated/predicted versus observed plots of aqueous solubility and plots of residuals.

Principal components analysis (PCA)[49] was also performed on the descriptors that were calculated. Experiments were performed using both the full set of 210 calculated descriptors and the reduced pool of 122 descriptors. The principal components that were calculated were then used as descriptors and regression models were developed that related the principal component descriptors to the aqueous solubility.

RESULTS AND DISCUSSION

The best Type 1 model that was generated from the combination of the multiple linear regression routines contained nine descriptors which are defined in Table 2. There were two geometry-dependent descriptors (SHDW 2 and SHDW 5), one topological descriptor (MOLC 3), three cpsa descriptors (PPSA 1, DPSA 3, and WNSA 1), and three descriptors that encoded hydrogen bonding (SAAA 3, CHAA 2, and EHBB). The geometric descriptors were the shadow area projection of the molecules and the normalized shadow area projection of the molecules. The topological index was a molecular connectivity index to describe the size and shape

Table 2. Descriptors Chosen for the Type 1 Model for the Prediction of Aqueous Solubility[a]

descriptor	coefficient		std dev of coef
shadow area in the XZ plane with molecule oriented with moments of inertia	−0.0364	±	0.0083
shadow area in the XZ plane/area of box defined by X and Z dimensions	3.21	±	0.43
path 1 valence and ring-corrected molecular connectivity	0.747	±	0.110
partial positive surface area	0.0189	±	0.0014
difference in total charge weighted partial positive and negative surface areas	−0.00439	±	0.00025
surface weighted partial negative surface area	0.0391	±	0.0033
sum of surface areas of hydrogen bonding acceptor atoms/total molecular surface area	4.11	±	0.52
sum of charges on hydrogen bonding acceptor atoms/no. of hydrogen bonding acceptor atoms	4.18	±	0.45
electrostatic hydrogen bonding basicity	−4.60	±	0.60
constant	−3.25	±	0.37

[a] rms error = 0.638 log units; $R = 0.965$; $n = 295$.

of the molecules. The cpsa descriptors encoded a combination of the partial positive and negative surface areas of the molecules, including both partial atomic charge and solvent-accessible surface area information. Two of the hydrogen-bonding descriptors (SAAA 3 and CHAA 2) were from an ADAPT program and included information about hydrogen-bonding acceptor atoms, while the third hydrogen-bonding descriptor (EHBB) was the electrostatic hydrogen bonding basicity described by Lowrey et al.[41] in the TLSER studies.

This Type 1 model had a tset rms error of 0.638 log units after five compounds were flagged as outliers and removed from the tset. Two hundred ninety-five compounds were used in the development of the linear regression model. The outliers were sucrose, tetrafluoromethane, octachlorodibenzo-*p*-dioxin, tetracycline, and oxytetracycline. The main reason for these compounds being outliers probably was their structural difference from the remainder of the set of compounds. In the case of the two halogenated compounds, the very high degree of halogenation may have contributed to their outlier status. The external pset of 32 compounds had an rms error of 0.556 log units. The aqueous solubility of sucrose was the highest value in the data set.

The descriptors that were chosen based on linear feature selection were then used as input for a CNN to develop a Type 2 model. The architecture of the network was nine input neurons, six hidden neurons, and one output neuron, giving 67 adjustable parameters. The 295 compounds in the tset from the linear regression experiments were split into a 265-compound tset and a 30-compound cvset for the neural network experiments. Training of the network resulted in tset, cvset, and pset rms errors of 0.460, 0.455, and 0.446 log units, respectively. The correlation coefficient, *R*, between the experimental values and the calculated values was 0.982 for this Type 2 model. The tset rms error improved by 28% and the pset improved by 20% over the Type 1 model results.

Nonlinear feature selection was also used to choose a subset of descriptors to relate molecular structure to aqueous solubility. The 9:6:1 architecture of the neural network was retained from the Type 2 experiments. A new set of nine descriptors, defined in Table 3, was selected as the best nonlinear model. A variety of descriptor types was represented in the model. A shadow projection descriptor (SHDW 3) was again present as well as a gravitation index (GRAV 3) which was another geometry-based descriptor. There were three topological descriptors (ALLP 3, WTPT 4, and 2SP3) that contained information about weighted paths and carbon types, one pure electronic descriptor (QNEG), and three cpsa descriptors (PPSA 1, FPSA 3, and WPSA 3). The partial positive surface area was the only descriptor that appeared in both the Type 1 and Type 3 models. Quite surprisingly, there were no descriptors that directly encoded hydrogen-bonding effects. Seemingly the combination of information represented by the nine descriptors that were chosen was enough to describe the physical interactions present in controlling aqueous solubility.

Table 3. Descriptors Chosen for the Type 3 Model for Prediction of Aqueous Solubility[a]

label	descriptor
SHDW 3	shadow area in the XY plane with molecule oriented with moments of inertia
GRAV 3	cube root of gravitation index
ALLP 3	total weighted no. of paths
WTPT 4	sum of path weights starting from oxygens
2SP3	no. of secondary sp^3 carbons
QNEG	charge on the most negative atom
PPSA 1	partial positive surface area
FPSA 3	fractional atomic charge weighted partial positive surface area
WPSA 3	surface weighted atomic charge weighted partial positive surface area

[a] tset rms error = 0.394 log units; cvset = 0.358 log units; pset = 0.343 log units.

The Type 3 model that was developed provided the most accurate tool for the estimation of aqueous solubility. The model had tset, cvset, and pset rms errors of 0.394, 0.358, and 0.343 log units. The correlation coefficient, *R*, between the experimental values and the calculated values was 0.987 for this Type 3 model. The rms errors represented a tset improvement of 14%, a cvset improvement of 21%, and a pset improvement of 23% compared to the previous neural network results. It also represented a 38% improvement in both the tset and pset over linear regression results. The results of the highest quality model that was developed is shown in Figure 1, a plot of calculated and predicted versus observed aqueous solubility values.

At attempt was also made to develop linear models using PCA. Models that were comparable to, but not better than, the Type 1 model reported here were able to be developed only by using the top 70 principal components that were calculated from the entire pool of 210 descriptors that were generated as the pool from which the model descriptors were chosen. Many of these descriptors were highly correlated among themselves or contained very little information, with less than 10% of the values of a descriptor varying for the compounds in the data set. The nine-descriptor principal component model had a tset rms error of 0.600 log units, a pset rms error of 0.557 log units, and *R* value of 0.969. Taking into consideration the fact that the principal components that appeared in the model had a contribution from all 210 descriptors as opposed to the nine pure descriptors in the Type 1 model, the Type 1 model had equal predictive power with many fewer adjustable parameters. Attempts were also made to improve the results of the linear principal

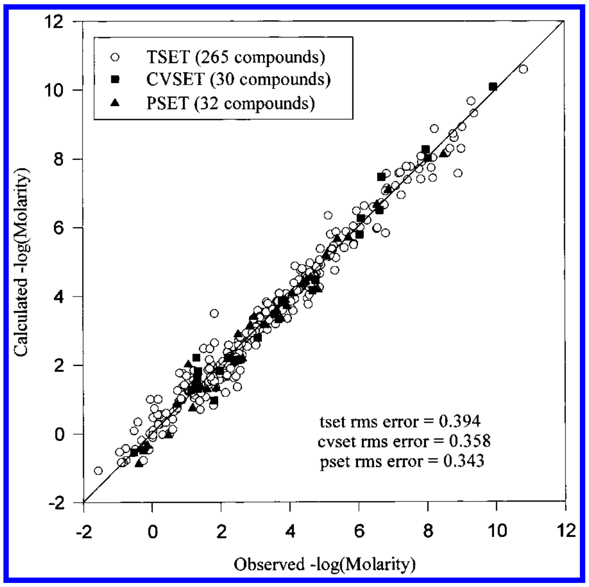

Figure 1. A plot of calculated and predicted versus observed aqueous solubility values for the Type 3 model.

component model using neural networks, but again the results were not better than the Type 2 model results already described, with $R = 0.983$.

CONCLUSIONS

In this study, linear and nonlinear models have been reported that relate molecular structure to aqueous solubility. Features of the molecular structures of a large, diverse set of organic compounds were numerically encoded as structure-based descriptors to represent the topology, geometry, and electronic nature of the compounds as well as to try and encode hydrogen-bonding effects. Multiple linear regression and computational neural networks were used in conjunction with genetic algorithm and simulated annealing routines to develop Type 1, Type 2, and Type 3 models that can be used for the estimation of aqueous solubility. The Type 3 model had the lowest rms errors of the models that were generated with rms errors ranging from 0.343 to 0.394 log units for a set of compounds which covered an −log(molarity) range from about −2 to 12 log units. These models provide the ability to predict aqueous solubility for compounds that were not used in their development. This predictive ability will be useful in cases where it is difficult or impossible to make experimental measurements, such as in the case of proposed drug candidates which have not been synthesized.

ACKNOWLEDGMENT

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

REFERENCES AND NOTES

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

(2) Ruelle, P.; Kesselring, U. W. The Hydrophobic Propensity of Water toward Amphiprotic Solutes: Prediction and Molecular Origin of the Aqueous Solubility of Aliphatic Alcohols. *J. Pharm. Sci.* **1997**, *86*, 179−186.

(3) 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.

(4) Patil, G. S. Prediction of Aqueous Solubility and Octanol−water Partition Coefficient for Pesticides Based on their Molecular Structure. *J. Hazard. Mater.* **1994**, *36*, 35−43.

(5) Ruelle, P.; Kesselring, 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.

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

(7) Bodor, N.; Harget, A.; Huang, M. Neural Network Studies. 1. Estimation of the Aqueous Solubility of Organic Compounds. *J. Am. Chem. Soc.* **1991**, *113*, 9480−9483.

(8) Bodor, N.; Huang, M.; Harget, A. Neural Network Studies. 4. An Extended Study of the Aqueous Solubility of Organic Compounds. *International Journal of Quantum Chemistry: Quantum Chemistry Symposium Vol. 26*; John Wiley & Sons: New York, 1992; 853−867.

(9) Suzuki, T. Development of an Automatic Estimation System for Both the Partition Coefficient and Aqueous Solubility. *J. Comput.-Aided Mol. Design* **1991**, *5*, 149−166.

(10) Silla, E.; Tuñón, I.; Villar, F.; Pascual-Ahuir, J. L. Molecular Surface Calculations on Organic Compounds. Molecular area-Aqueous Solubility Relationships. *J. Mol. Struct. (THEOCHEM)* **1992**, *254*, 369−377.

(11) Nouwen, J.; Hansen, B. Correlation Analysis Between Watersolubility, Octanol−water Partition Coefficient and Melting Point Based on Clustering. *Quant. Struct.-Act. Relat.* **1996**, *15*, 17−30.

(12) Zhang, X.; Gobas, F. A. P. C. A Thermodynamic Analysis of the Relationships Between Molecular Size, Hydrophobicity, Aqueous Solubility and Octanol−water Partitioning of Organic Chemicals. *Chemosphere* **1995**, *31*, 3501−3521.

(13) Li. A.; Doucette, W. J.; Andren, A. W. Estimation of Aqueous Solubility, Octanol/Water Partition Coefficient, and Henry's Law Constant for Polychlorinated Biphenyls Using UNIFAC. *Chemosphere* **1994**, *29*, 657−669.

(14) Yalkowsky, S. H. Estimation of the Aqueous Solubility of Complex Organic Compounds. *Chemosphere* **1993**, *26*, 1239−1261.

(15) Myrdal, P. B.; Manka, A. M.; Yalkowsky, S. H. AQUAFAC 3: Aqueous Functional Group Activity Coefficients; Application to the Estimation of Aqueous Solubility. *Chemosphere* **1995**, *30*, 1619−1637.

(16) 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.

(17) Nelson, T. M.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 601−609.

(18) Egolf, L. M.; 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.

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

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

(21) Mitchell, B. E.; Jurs, P. C. Prediction of Autoignition Temperatures of Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 538−547.

(22) 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.

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

(24) Jurs, P. C.; Chou, J. T.; Yuan, M. Studies of Chemical Structure-Biological Activity Relations Using Pattern Recognition. In *Computer-Assisted Drug Design*; Olson, E. C., Christoffersen, R. E., Eds.; The American Chemical Society: Washington, D. C., 1979; Chapter 4.

(25) Wessel, M. D. *Computer-Assisted Development of Quantitative Structure−Property Relationships And Design of Feature Selection Routines*; Ph.D. Dissertation, Pennsylvania State University, University Park, PA, 1996.

(26) Sutter, J. M.; Jurs, P. C. Selection of Molecular Descriptors for Quantitative Structure−Activity Relationships. In *Adaption of Simulated Annealing to Chemical Problems*; Kalivas, J. H., Ed.; Elsevier Science Publishers B. V.: Amsterdam, 1995; Chapter 5.

(27) 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.

(28) 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. *Environmental Toxicol. Chem.* **1994**, *13*, 841−851.

(29) Lee, R.; Dung, M.; Verlin, J.; Carreira, L.; Famini, G.; Hilderbrandt, R.; Frazer, J.; Heckler, C.; Mishra, R. S. *Comparison of Five Computational Methods for the Calculation of Aqueous Solubility-I: Construction of Data Set and Comparison of Results from SPARC, TLSER, and UNIFAC Solvation Models*; Presented at the 210th American Chemical Society National Meeting; Chicago, 1995.

(30) Verlin, J.; Dung, M.; Lee, R.; Hilderbrandt, R.; Heckler, C.; Frazer, J.; Carreira, L.; Famini, G.; Mishra, R. S. *Comparison of Five Computational Methods for the Calculation of Aqueous Solubility-II: QSPR Predictions Based on SIMS−Calculated Descriptors Using All Possible Subsets Regression (APSREG) and Partial Least-Squares (PLS) Analysis*; Presented at the 210th American Chemical Society National Meeting; Chicago, 1995.

(31) Yalkowsky, S. H. *Adb, the ARIZONA dATAbASE of Aqueous Solubility;* College of Pharmacy, University of Arizona, Tucson, AZ.

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

(33) Cao, C. Distance-Edge Topological Index − Research on Structure−Property Relationship of Alkanes. *Huaxue Tongbao* **1996**, *54*, 533−538.

(34) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Studies: Hertfordshire, England, 1986.

(35) Pearlman, R. S. In *Physical Chemistry Properties of Drugs*; Sinkula, A. A., Valvani, S. C., Eds.; Quantum Chemistry Program Exchange No. 413; Marcel Dekker: New York, 1980; Chapter 10.

(36) 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.

(37) 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.

(38) 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 Inorganics. *J. Phys. Chem.* **1996**, *100*, 10400−10407.

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

(40) Dixon, S. L. *Development of Computational Tools for use in Quantitative Structure−Activity and Structure−Property Relationships*; Ph.D. Dissertation, Pennsylvania State University: University Park, PA, 1994.

(41) Lowrey, A. H.; Cramer, C. J.; Urban, J. J.; Famini, G. R. Quantum Chemical Descriptors for Linear Solvation Energy Relationships. *Computers Chem.* **1995**, *19*, 209−215.

(42) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.

(43) Neter, J.; Wasserman, W.; Kutner, M. H. *Applied Linear Statistical Models*; Irwin: Homewood, IL, 1990.

(44) Broyden, C. G. The Convergence of a Class of Double-Rank Minimization Algorithms. *J. Inst. Maths. Appl.* **1970**, *6*, 76−90.

(45) Fletcher, R. A New Approach to Variable Metric Algorithms. *Comput. J.* **1970**, *13*, 317−322.

(46) Goldfarb, D. A Family of Variable-Metric Methods Derived by Variational Means. *Math. Comput.* **1970**, *24*, 23−26.

(47) Shanno, D. F. Conditioning of quasi-Newton Methods for Function Minimizations. *Math. Comput.* **1970**, *24*, 647−656.

(48) Fletcher, R. *Practical Methods of Optimization − Volume 1*; Wiley-Interscience: New York, 1980.

(49) Massart, D. L.; Vandeginste, B. G. M.; Deming, S. N.; Michotte, Y.; Kaufman, L. *Chemometrics: A Textbook*; Elsevier: Amsterdam, 1988; Chapter 21.

CI970117F