Prediction of C_{60} Solubilities from Solvent Molecular Structures

Stephen M. Danauskas and Peter C. Jurs*

Department of Chemistry, The Pennsylvania State University, 152 Davey Laboratory, University Park, Pennsylvania 16801

Received October 2, 2000

Models predicting fullerene solubility in 96 solvents at 298 K were developed using multiple linear regression and feed-forward computational neural networks (CNN). The data set consisted of a diverse set of solvents with solubilities ranging from −3.00 to 2.12 log (solubility) where solubility = (1×10^4)(mole fraction of C_{60} in saturated solution). Each solvent was represented by calculated molecular structure descriptors. A pool of the best linear models, as determined by rms error, was developed, and a CNN model was developed for each of the linear models. The best CNN model was chosen based on the lowest value of a specified cost function and had an architecture of 9−3−1. The 76-compound training set for this model had a root-mean-square error of 0.255 log solubility units, while the 10-compound cross-validation set had an rms error of 0.253. The 10-compound external prediction set had an rms error of 0.346 log solubility units.

INTRODUCTION

Since the discovery of C_{60} in 1985,[1] there has been a burst of interest in identifying possible applications of this highly symmetrical molecule. This interest has led to groundbreaking research in biochemistry,[1] materials science,[2] and organic chemistry[3] among other fields. One of the drawbacks to working with C_{60} is that it is not very soluble in most organic solvents[4] and there does not appear to be a specific physical property that one can use to determine what would constitute a good solvent.[5] However, several properties can be used to determine what solvents C_{60} is not soluble in; for instance, being a nonpolar molecule C_{60} is not soluble in polar solvents.[4] In addition, due to its large volume, it is not soluble in protic solvents as they exclude C_{60}.[4]

Since there is no one property that accurately predicts C_{60} solubility, a quantitative structure−property relationship (QSPR) study would seem to be ideal in attempting to predict solubilities. In fact, this has already been done utilizing the Theoretical Linear Solvation Energy Relationship approach developed by Famini and Wilson using an equation taking into account the bulk, dipole moment, and hydrogen bonding ability of the compound.[6] However, this method, while increasing the knowledge about correlations of certain properties to solubility, did not possess predictive ability due to several outliers where the errors were several orders of magnitude greater than the correct value.[6]

A study by Murray et al.[7] also developed a model based on the electrostatic potential of a solvent and quantities related to the surface area of a solvent. This model is notable for the excellent fit to the data, showing a linear correlation coefficient of 0.954 and a standard deviation of 0.475 for the 22 organic solvents that were modeled. However, the predictive ability of this model was not tested.

A solubility study by Ruoff et al.[4] also showed that several properties can be potential indicators of solubility, but they do not always predict solubility. In that study, it was shown that polarizability, polarity, molecular size, and cohesive energy density were correlated to solubility most of the time. It was also noted that C_{60} was much more soluble in aromatic solvents, but no pattern was evident as to how the substituents affected solubility.[4,5]

One of the largest problems in achieving an accurate QSPR model is the accuracy of the data available. In one study, a 40% difference in solubilities was found for several compounds between two data sets and only an 8 °C difference in the experimental temperature. However, many of the solvents did not exhibit a change in solubility at all when the temperature was altered.[4] It was proposed that the large difference in mixing time, approximately 19 h, coupled with an abnormal temperature dependency led to the discrepancy. Another study found that C_{60} undergoes a slow reaction with some solvents, including pyridine, and this can lead to a false reading of solubility.[5] The discrepancies in the reporting of C_{60} solubility is one of the reasons that a computational neural network (CNN) might be especially well suited to the task of predicting C_{60} solubility. Due to a CNN's property of learning about the compounds they are dealing with, CNNs are remarkably fault tolerant. In many cases, they are able to overcome several faulty data points and still form a predictive model.

METHODOLOGY

Data Set. The set of organic solvents used for this study was provided by Dr. Allan Smith of Drexel University and included 96 compounds compiled from several sources.[8−11] The solubility was expressed in units of (1×10^4)(mole fraction of C_{60} in the saturated solution). To limit the range of the data, the base-ten logarithm was taken, and the new range extended from −3.00 (ethanol and propanone) to 2.12 log units (1-phenylnaphthalene). The identity of each compound and its experimental value for log solubility are presented in Table 1. The data set consisted of a mixture of 45 alkanes, 36 benzene derivatives, 7 naphthalenes, 14 oxygen-containing compounds, 10 nitrogen-containing com-

* Corresponding author phone: (814)865-3739; fax: (814)865-3314.

10.1021/ci000140s CCC: $20.00
Published on Web 01/16/2001

Table 1. Data Set Compounds and Model Prediction

no.	compound	log(solubility)[a]	Type II	no.	compound	log(solubility)[a]	Type II
1	1,1,1-trichloroethane	−0.678	−0.277	49	bromocyclohexane	0.565	0.540
2	1,1,2,2-tetrachloroethane	0.886	0.418	50	bromocyclopentane	−0.215	0.107
3	1,1,2-trichlorotrifluoroethane	−1.770	−1.76	51	chlorobenzene[p]	0.996	0.557
4	1,2,3,4-tetramethylbenzene	1.053	1.00	52	chlorocylohexane	−0.086	−0.165
5	1,2,3,5-tetramethylbenzene	1.640	1.734	53	cisdecalin	0.663	0.858
6	1,2,3-tribromopropane	1.127	1.20	54	cyclohexane	−1.268	−1.45
7	1,2,3-trichloropropane	−0.027	−0.368	55	cyclohexene	0.220	0.175
8	1,2,3-trimethylbenzene	0.943	1.115	56	cyclopentane	−2.523	−2.357
9	1,2,4-trichlorobenzene[p]	1.176	2.033	57	decane	−0.721	−0.737
10	1,2,4-trimethylbenzene	1.533	1.446	58	dibromomethane	−0.456	−0.214
11	1,2-dibromobenzene	1.358	1.705	59	dichlormethane	−0.699	−1.169
12	1,2-dibromoethane	−0.201	−0.342	60	diiodomethane	−0.824	−0.071
13	1,2-dibromoethene[cv]	0.334	0.135	61	dodecane	−0.541	−0.593
14	1,2-dibromopropane	−0.292	−0.189	62	ethanol	−3.000	−3.256
15	1,2-dichlorobenzene[cv]	1.623	1.107	63	ethylbenzene	0.565	0.489
16	1,2-dichlorodifluoroethane	−1.377	−0.887	64	fluorobenzene	−0.108	0.281
17	1,2-dichlorethane[cv]	−1.046	−1.142	65	hexane[cv]	−1.137	−1.417
18	1,2-dichloropropane[p]	−0.854	−0.819	66	iodobenzene[p]	0.512	0.844
19	1,2-dimethylbenzene	1.095	0.863	67	iodocyclohexane	1.168	1.009
20	1,2-dimethylnaphthalene	1.892	1.869	68	iodomethane	−0.174	−0.558
21	1,3,4 trimethylbenzene	0.516	0.671	69	methoxybenzene	0.926	0.348
22	1,3-dibromobenzene	1.364	1.217	70	nitrobenzene	0.041	0.363
23	1,3-dichlorobenzene	0.580	0.853	71	nitroethane	−2.699	−2.653
24	1,3-dichloropropane	−0.824	−0.664	72	*N*-methyl-2-pyrrolidone	0.079	0.135
25	1,3-dimethylbenzene[p]	0.682	0.591	73	*N*,*N*-dimethylformamide	−1.538	−1.536
26	1,4-dimethylbenzene	0.728	0.65	74	*N*-propylbenzene	0.462	0.564
27	1,4-dioxane[cv]	−1.310	−1.355	75	octane	−1.252	−0.937
28	1-bromo-2-methylnaphthalene	1.876	1.878	76	pentane	−2.097	−1.875
29	1-bromopropane[cv]	−1.155	−1.116	77	phenylbromomethane	0.912	0.869
30	1-chloro-2-methylpropane[p]	−1.398	−1.15	78	phenylchloromethane	0.584	0.594
31	1-chloronaphthalene	1.987	1.753	79	phenylisocyanate	0.566	0.365
32	1-iodo-2-methylpropane[cv]	−0.268	−0.592	80	phenylmethanal	−0.229	−0.085
33	1-iodopropane	−0.638	−0.907	81	phenyltrichloromethane[cv]	0.975	1.103
34	1-methylnaphthalene	1.833	1.731	82	propanone[cv]	−3.000	−2.689
35	1-phenylnaphthalene[p]	2.117	2.12	83	pyridine	−0.004	−0.091
36	2,2,4-trimethylpentane	−1.222	−1.392	84	quinoline	1.072	1.067
37	2-butylbenzene	0.377	0.488	85	*tert*-butylbenzene[cv]	0.288	0.139
38	2-iodopropane	−0.824	−1.08	86	tetrachloroethene	0.230	0.498
39	2-methylphenol	−1.538	−1.516	87	tetrachloromethane	−0.398	0.039
40	2-methylthiophene	0.959	0.469	88	tetradecane	−0.342	−0.423
41	2-nitrotoluene[p]	0.600	0.586	89	tetrahydrothiophene	−1.444	−1.042
42	2-phenylbutane	0.377	0.699	90	tetralin	1.491	1.368
43	2-phenylpropane	0.365	0.565	91	thiophene	−0.357	−0.619
44	3-nitrotoluene	0.589	0.576	92	toluene	0.602	0.389
45	benzene	0.322	0.335	93	transdecalin[p]	0.462	0.883
46	benzonitrile	−0.149	−0.115	94	tribromomethane	0.780	0.100
47	bis (methoxyethyl) ether	−1.194	−1.262	95	trichloroethene	0.230	0.127
48	bromobenzene	0.681	0.769	96	trichloromethane[p]	−0.658	0.100

[a] $(1 \cdot 10^4)$(mole fraction of C_{60} in saturated solution). [cv] Cross-validation set members. [p] Prediction set members.

pounds, 21 chlorine-containing compounds, and 15 bromine-containing compounds.[6] The molecular weight ranged from 46.1 daltons for ethanol to 280.8 daltons for 1,2,3-tribromopropane. There were no obvious groupings of compounds involving size, molecular weight, degree of branching, or polarity in the data set.

For the generation of Type I linear regression models the data are split into a training set (tset) composed of 90% of the data set and a disjoint prediction set consisting of the remaining solvents. The data set is split into three disjoint sets for Type II and Type III CNN model building—a training set, a cross-validation set (cvset), and a prediction set (pset). The cross-validation and prediction sets each contain 10% of the data, and the remaining 80% belong to the training set. The prediction set members are the same for all three model types, while the training set is different between Type I and Types II and III since the cross-validation set is removed in the latter cases. Table 1 indicates which compounds are members of each data set.

Molecular Modeling and Optimization. The structure of each compound is sketched using Hyperchem and is placed in a preliminary conformation using a conjugate gradient descent algorithm. The compounds are then optimized again using the semiempirical MOPAC software program.[12] The PM3 Hamiltonian[13] is used for each optimization. To ensure the optimizations did not begin in a local minimum, Peter's test for optimization had to be satisfied. If it is not satisfied, the intermolecular distances and angles are altered slightly, and the optimization is repeated until convergence is attained.

Descriptor Generation and Feature Selection. Molecular structure descriptors are then generated for each compound based solely on their structures. For each compound, four types of descriptors are calculated: topological, electronic, geometric, and geometric/electronic hybrids. The topological descriptors calculated include the following: κ indices,[14−16]

molecular connectivity descriptors,[17−20] refractive index, atom types, molecular distance edge descriptors,[21,22] and weighted path routines.[23,24] Geometric descriptors include computation of the geometric axes, moments of inertia,[25] surface area,[26] shadow descriptors,[27,28] molecular polarizability, and the gravitational index[29] of each molecule. The electronic descriptors[30,31] calculated include the atomic charges, σ charges, HOMO, LUMO, electronegativity, and hardness (defined as $^1/_2$(HOMO − LUMO)). The electronic/geometric hybrid descriptors encompass charged partial surface area descriptors[32] and a set of related hydrogen-bonding descriptors.

After descriptor generation, the descriptors are submitted to a feature selection routine in order to limit the amount of potentially irrelevant or redundant information. Feature selection was performed using a routine that tests for identical values (85% of descriptor values are identical) and descriptors showing a pairwise correlation greater than 0.85. These descriptors are then eliminated from the original pool due to the information they possess being superfluous. If this reduced pool is not less than the theoretical cutoff limit of 60% of the data set,[33] a vector space descriptor analysis routine is performed to further reduce the pool to reach the theoretical cutoff.

Model Generation. Each of the models presented here were generated using the ADAPT (Automated Data and Pattern Recognition Toolkit) software[34,35] developed at Penn State. Three types of models were generated: Type I models are multiple linear regression models, Type II models are computational neural network (CNN) models built with the linearly selected descriptors coming from the Type I model, and Type III models are CNN models found with nonlinear feature selection from the reduced descriptor pool.

Type I: Linear Regression Models. Linear regression models are generated using a simulated annealing technique[36] that converges based on root-mean-square (rms) errors for each model generated and terminates based on the number of steps involved in generating a better model. Another criterion may be specified to ensure the validity of the model, T-values. The T-value for a descriptor in a model equals the descriptor's coefficient divided by the standard error of that coefficient. A set of coefficients with T-values greater than four is the normal cutoff criterion for a good model. The selected model is then subjected to a stepwise linear regression analysis routine to determine if multicollinearities greater than 0.95 are present in the selected model. Multicollinearities may cause difficulties in certain aspects of forming a linear model, and if a model contains these multicollinearities, it is generally discarded.

The coefficients of the model are then computed using a multiple linear regression algorithm that computes the standard deviation, F-, and P-values. It has been observed that models with high F-values and low P-values tend to form better models. If two models have similar rms errors, F- and P-values can help to determine which model to choose. The chosen model is then stored for further analysis.

The next step in the formation of a linear model is to determine if outliers are present in the model. This is done by performing eight statistical tests of fitness for each compound in the training set, including the residual, Cook's distance, and Studentized Residual. If a compound fails four of the eight tests, that compound is considered a possible outlier. Based upon the overall effect of removal on the rms error of the model, the compound may be removed from the study entirely.

Finally, the external prediction set is used to test the predictive properties of the model. Regardless of the outcome of this test, the chosen model remains the best model based on the statistical tests that were performed, and the descriptors of this model are used to develop a nonlinear model.

Type II: Nonlinear Models using Linear Feature Selection. The Type I model descriptors are used as the input to a three-layer, feed-forward, fully connected computational neural network. The descriptor values are transformed to restrict their range to the interval [0,1] and they are then sent to a hidden layer that employs a sigmoidal activation function. These neurons serve as a summing junction and adjust the original weights to a more optimal value. The number of observations in the data restricts the number of neurons in the hidden layer. The number of observations must exceed twice the number of adjustable parameters in the CNN. Training of the CNN model is performed using the Broyden-Fletcher-Goldfarb-Shanno quasi-Newton algorithm.[37] The training is monitored by watching the rms error of the cross-validation set compounds. In a general case, initially the errors of the tset and cvset each decrease. However, after a certain point the cvset error begins to rise as the tset error continues to decrease.[36] It is believed that the network is beginning to overtrain at this point. Overtraining leads to a degradation of the model and more precisely a degradation of its predictive ability.

To find a good set of starting weights and biases, after initial CNN runs, a simulated annealing algorithm is run on the CNN, and the weights and biases determined are used as the starting weights and biases for the standard CNN algorithm. This technique, while a better way to avoid local minima than random chance, will not always find the local minimum each time and needs to be run several times.

Once the best set of weights and biases has been determined, the network is then tested using the external prediction set. Even if the pset values are not within an acceptable range of error, the model developed is considered the best that can be generated with the tools at hand.

Type III: Nonlinear Modeling and Nonlinear Feature Selection. The Type III CNN model combines nonlinear feature selection, utilizing a genetic algorithm, with a CNN having the same architecture as the Type II and the same number of descriptors as the Type I model. This approach creates a nonlinear model, as a CNN is used as the fitness evaluator. Just as in Type II model development, the training of a Type III model utilizes a cross-validation set to determine when to stop training. The practical difference between this type of model and the other two model types is the much greater amount of computational time used. This limitation necessitates that one use the descriptor subset selected by Type I and the CNN architecture deemed optimal by Type II.

This study was performed on a DEC 3000 AXP Model 500 workstation and a DEC Alpha 500 Workstation, both of which run under the UNIX operating system. Hyperchem is a product of Chemsoft, and MOPAC was developed by J. P. P. Stewart.[13] All other routines were developed at The Pennsylvania State University.

Table 2. Type I/II Model Descriptors

descriptor	range	type	description[a]
S6CH	0 to $5.55 \cdot 10^{-2}$	top.	χ index of chains of length six
NCl	0 to 4	top.	number of chlorines
NBr	0 to 3	top.	number of bromines
MDE 14	0 to 9.16	top.	distance edge for 1° to 4° carbons
MDE 44	0 to 2.03	top.	distance edge for 4° to 4° carbons
EMIN	−6.06 to 2	top.	minimum atomic e-state value
MOMH 7	1.02 to 6.5	geom.	radius of gyration
MPOL	5.11 to 27.73	top.	molecular polarizability
LUMO	−1.24 to 3.82	elec.	lowest unoccupied molecular orbital

[a] S6CH—simple molecular connectivity index descriptor of paths of length six;[17,19,20] MDE 14—molecular distance edge value based on the distances between all 1° and 4° carbons;[38] MDE 44—molecular distance edge value based on the distances between all 4° and 4° carbons;[38] EMIN—encodes information equivalent to free valence;[39] MOMH 7—radius of gyration which includes hydrogens in the calculations;[25] MPOL—molecular polarizability calculated using a sum of atomic contributions;[40] LUMO—lowest unoccupied molecular orbital calculated by MOPAC using an AM1 Hamiltonian.[13]

RESULTS AND DISCUSSION

The best Type I multiple linear regression model was developed using the simulated annealing method, and it consisted of nine descriptors. These nine descriptors were chosen from a reduced pool of 85 descriptors, which had previously been identified using the identical value test and testing for pairwise correlations greater than 0.85. The descriptors in this model are presented in Table 2. The training set rms error was 0.416 log solubility units, and the prediction set error was 0.50 log solubility units. A large portion of the prediction set error was due to the predicted value of 1-phenylnaphthalene; the pset rms error decreased to 0.40 log units when this compound was omitted from the calculation. For the nine descriptors in this model, the largest pairwise correlation was 0.84, and the mean over all 36 pairwise correlations was 0.26.

The best Type II CNN model had a 9−3−1 architecture. The rms error for the training set decreased to 0.30 log units. The cross-validation set error for this model was 0.45 log units. However, the prediction set rms error increased to 0.52 log units although 1-phenylnaphthalene was brought within the range of the training and cross-validation sets. The compound contributing most to the rms error in this model is trichloromethane, and the rms error decreases to 0.34 log units when this compound is not taken into consideration.

No viable Type III CNN model was found. Although the Type II model shows reasonable predictive ability over a wide range of data, the rms errors were large and did not represent a good model, especially considering the difference in the errors between the training set and the cross-validation set.

A variation of the standard Type II model development procedure was developed to try to create a more accurate model, as the Type III model generation failed to yield a better model. For convenience, it will be referred to as Type IIA. In this variation, the architecture of the CNN was determined based on the merits of the CNN model itself. The simulated annealing and linear regression routine was used to create a pool of linear regression models, containing five to nine descriptors each, with T values greater than 1.0. As the model being sought is nonlinear, the tests of linear fitness will be less important in determining the model chosen. The models with the lowest rms errors were selected and sent to the computational neural network routines. All of the possible CNN architectures allowed were tested for each model. The CNN routines were run multiple times as the error values can vary greatly depending on the initial random seed. The number of architectures being considered was then pared down by ranking each run of each architecture by its associated cost function. The cost function was defined as

$$\text{cost} = \text{tset}_{\text{rms}} + {}^1/_2 |\, \text{tset}_{\text{rms}} - \text{cvset}_{\text{rms}} \,|$$

where tset_{rms} is the rms error for the training set compounds and $\text{cvset}_{\text{rms}}$ is the rms error for the cross-validation set compounds.

For each architecture, the cost of the best run was added to the sum of the cost of the next four best runs divided by 7.5. The five best values were used to somewhat discourage the selection of a model that may have only one very low rms error value for one of the runs and much higher rms errors for the remaining runs. This procedure was intended to find a model that was stable over several random seeds and not just a particular seed. This value was then assigned to each architecture as the fitness evaluator.

Utilizing this method, a second CNN model was developed consisting of a 9−3−1 architecture, and the descriptors that compose this model are seen in Table 3. The rms error was 0.255 log units for the training set, 0.253 log units for the cross-validation set, and 0.346 log units for the external prediction set. The values calculated or predicted for the 96 individual compounds are provided in Table 1. Figure 1 shows that this model gives a good fit to the one-to-one correlation line. The prediction set rms error is thrown off slightly by trichloromethane, which has a residual of 0.75. With the exception of trichloromethane, the remaining prediction set compounds fall within the range of the training and cross-validation sets, indicating the formation of a predictive model. The Type IIA model represents a 15.3% improvement in the training set rms error, a 43.5% improvement in the cross-validation error, and a 33.7% improvement

Table 3. Type IIA Descriptors

descriptor	range	type	description[a]
S6CH	0 to $5.55 \cdot 10^{-2}$	top.	χ index of chains of length six
N5C	0 to 9	top.	number of fifth-order paths
MOLC 5	0 to 4.72	top.	path three molecular connectivity
NN	0 to 1	top.	number of nitrogens
MDE 14	0 to 9.16	top.	distance edge for 1° to 4° carbons
QNEG	-0.55 to $-6.24 \cdot 10^{-2}$	elec.	charge on the most negative atom
PNSA 3	−55.41 to −0.8	hybrid	partial negative surface area
MPOL	5.11 to 27.73	top.	molecular polarizability
HARD	4.15 to 7.48	elec.	hardness coefficient

[a] S6CH—simple molecular connectivity index descriptor of paths of length six;[17,19,20] N5C—number of fifth-order paths;[17,19,20] MOLC 5—path 3 molecular connectivity: A−B−C-D;[17,18,41] MDE 14—molecular distance edge value based on the distances between all 1° and 4° carbons;[38] QNEG—charge on the most negative atom, computed by taking the sum of σ (based on charge transfer) and π (based on simple Huckel molecular orbital calculations on conjugated π systems) charges;[31] PNSA—atomic charge weighted partial negative surface area with terms of atomic partial charges and partial solvent accessible surface areas;[32] MPOL—molecular polarizability calculated using a sum of atomic contributions;[40] HARD = 0.5·(HOMO − LUMO) where the HOMO and LUMO are calculated by MOPAC using an AM1 Hamiltonian.[13]

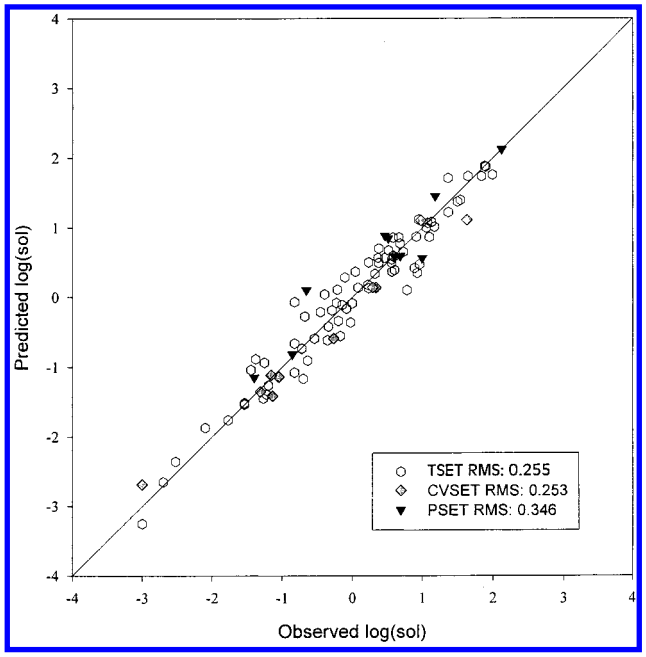

Figure 1. Plot of calculated/predicted vs observed log(solubility) values for the training, cross-validation, and prediction set compounds used to develop the Type IIA computational neural network model. The descriptors used are listed in Table 3. The data set members are listed in Table 1.

in the prediction set rms error over the Type II model. For the nine descriptors in this model, the largest pairwise correlation was 0.83, and the mean over all 36 pairwise correlations was 0.26. The set of nine descriptors was also tested for multicolllinearities using forward selection regression; no multicollinearities greater than 0.95 were found.

The nine-descriptor model consists of six topological descriptors, two electronic descriptors, and one electronic/geometric hybrid descriptor. S6CH is a simple χ descriptor, which does not take into account the valence of each carbon, and the descriptor encodes the number of six-membered chains. The choice of the S6CH descriptor agrees with the generally observed increased solubility in aromatic solvents, since this descriptor can indirectly encode for benzene. N5C encodes the number of fifth-order paths and highlights the molecules in the data set that are the most branched. MOLC 5 encodes molecular connectivity of path three, encoding branching and size, and NN counts the number of nitrogens present. Polarizability is another topological descriptor, and it was expected that this descriptor would appear in most models, as this is a measure of a solvent's ability to distort about another molecule; it is one of the basic properties that effects the solvation of any molecule. The electronic descriptors included are QNEG and HARD. HARD is half of the energy difference between the highest occupied molecular orbital and the lowest unoccupied molecular orbital, and QNEG is a measure of the charge on the most negative atom. The appearance of QNEG in the final model agrees with what is known about C_{60} solvation, since it will encode the presence of heteroatoms, which are an indicator of solvent polarity. The types of descriptors forming this model are not directly comparable to those used by Famini[6] or by Politzer[7] which are based on quite different approaches to the problem of C_{60} solubility, although descriptors encoding the potential for polar interactions play an important role in all models.

To ensure that the model obtained was not due to chance, a Monte Carlo study was performed by randomizing the dependent variable and rerunning the Type IIA models development process. If the model developed were due to chance, then the reasoning employed is that another predictive model should also be found using the same method even though the relationship between the structural descriptors and the log solubility values has been broken. The best Type IIA model developed consisted of seven descriptors and had a CNN architecture of 7−4−1. The training set rms error was 0.747 log units, and the cross-validation set rms error was 0.775 log units. There was no discernible pattern given by either of these sets. The prediction set rms error was 1.71 log units, which is approximately five times the error of the unscrambled dependent variable model and illustrates a lack of predictive ability. From the results obtained, it is highly unlikely that a predictive model could be developed for the scrambled dependent variables using the methods and cutoffs employed.

CONCLUSIONS

The Type I and Type II models were developed using standard tests of linear fitness and T-values greater than 4. The Type II model had a 9−3−1 architecture and consisted of topological, geometric, and electronic descriptors that tend to agree with basic solvation principles. The rms errors of 0.30 for the training set, 0.45 for the cross-validation set, and 0.52 for the prediction set indicated a poor predictive model in that the external prediction set rms error was much higher than the training set. In this case, both the prediction and the cross-validation set rms errors were higher than expected for this type of study. There was no Type III model found.

The Type IIA CNN model was developed using a cost function of the neural network itself as a fitness evaluator, as opposed to the linear regression model being used as the fitness evaluator. The model developed had a 9−3−1 architecture, and the rms errors for this model were as follows: 0.255 log units for the training set, 0.253 log units for the cross-validation set, and 0.346 log units for the prediction set. Thus, a predictive model was generated. While this method is a computationally more intense way to develop a nonlinear Type II CNN model, it yielded a 15.3% decrease in the rms error of the training set, 43.5% decrease in rms error of the cross-validation set, and a 33.7% decrease in the rms error of the external prediction set from the Type II model. The descriptors chosen by this routine were a mixture of topological, electronic, and hybrids. Several of the descriptors agreed with the known properties associated with solubility of C_{60}.

REFERENCES AND NOTES

(1) Koruga, D.; Hameroff, S.; Withers, J.; Lourty, R.; Sundareshan, M. *Fullerene C_{60}: History, Physics, Nanobiology, Nanotechnology*; North-Holland: New York, 1993.

(2) Rudolf, P. C_{60} Adsorption on Metal Surfaces, International Winterschool on Electronic Properties of Novel Materials, Kirchberg, Tyrol, Austria, March 2−9; World Scientific: 1996.

(3) Bianco, A.; et al. Synthesis and Chemical Properties of Fullerene Derivatives, International Winterschool on Electronic Properties of Novel Materials, Kirchberg, Tyrol, Austria, March 2−9; World Scientific: 1996.

(4) Ruoff, R. S.; et al. Solubility of C_{60} in a Variety of Solvents. *J. Phys. Chem.* **1993**, *97*, 3379−3383.
(5) Taylor, R. *The Chemisty of Fullerenes*; World Scientific: NJ, 1995; pp 35−52.
(6) Smith, A L.; Wilson, L. Y.; Famini, G. R. A Quantitative Structure−Property Relationship Study of C_{60} Solubility, 189th Electrochemical Society Meeting, Los Angeles, CA, May 1996; paper 546.
(7) Murray, J.; Gagarin, S.; Politzer, P. Representations of C_{60} Solubilities in Terms of Compured Molecular Surface Electrostatic Potentials and Areas. *J. Phys. Chem.* **1995**, *99*, 12081−3.
(8) Beck, M. T.; Mandi, G.; Keki, S. Solubility and Molecular Structure State of C_{60} Organic Solvents, 187th Electrochemical Society Meeting, Reno, NV, May 1995; paper 956.
(9) Ruoff, R. S.; Tse, R.; Malhoutra, R.; Lorents, D. C. Solubility of C_{60} in a Variety of Solvents. *J. Phys. Chem.* **1993**, *97*, 3379−3383.
(10) Scrivens, W. A.; Tour, J. M. Potent Solvents for C_{60} and their Utility for the Rapid Acquisition of ^{13}C NMR Data for Fullerenes. *J. Chem. Soc., Chem. Commun.* **1993**, *15*, 1207−1209.
(11) Sivaraman, N.; Dhamodaran, R.; Kaliappan, I.; Srinivasan, T. G.; Vasudeva Rao, P. R.; Mathews, C. K. Solubility of C_{60} in Organic Solvents. *J. Org. Chem.* **1992**, *57*, 6077−6079.
(12) Stewart, J. P. P. *MOPAC 6.0, Quantum Chemistry Program Exchange*; Indiana University: Bloomington, IN, Program 455,
(13) Stewart, J. P. P. MOPAC: A Semiempirical Molecular Orbital Program. *J. Comput.-Aided Molecular Design* **1990**, *4*, 1−105.
(14) Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct.-Act Relat. Pharmacol., Chem. Biol.* **1985**, *4*, 109−116.
(15) Kier, L. B. Shape Indices for Orders One and Three from Molecular Graphs. *Quant. Struct.-Act Relat. Pharmacol., Chem. Biol.* **1986**, *5*, 1−7.
(16) Kier, L. B. Distinguishing Atom Differences in Molecular Graph Shape Index. *Quant. Struct.-Act Relat. Pharmacol., Chem. Biol.* **1986**, *5*, 7−12.
(17) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Research Press Ltd., John Wiley and Sons: 1986.
(18) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(19) Kier, L. B.; Hall, L. H. Molecular Connectivity VII: Specific Treatment to Heteroatoms. *J. Pharm. Sci.* **1976**, *65*, 1806−1809.
(20) Kier, L. B.; Hall, L. H.; Murray, W. J.; Randic, M. Molecular Connectivity I: Relationship to Nonspecific Local Anesthesia. *J. Pharm. Sci.* **1975**, *64*, 1971−4.
(21) Liu, S.; Cao, C.; Li, Z. Approach to Normal Boiling Point of Alkanes Based on a Novel Molecular Distance-Edge Vector. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387−394.
(22) Mitchell, B. Ph.D. Dissertation, Penn State University, State College, PA, 1997.
(23) Weiner, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(24) Randic, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L. Search for All Self-Avoiding Paths for Molecular Graphs. *Computational Chem.* **1979**, *3*, 5−13.
(25) Goldstein, H. *Classical Mechanics*; Addison-Wesley: Reading, MA, 1950.
(26) Pearlman, R. S. In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H. S., A. A., Valvani, S. C., Eds.; Marcel Dekker: New York, 1980; Chapter 10.
(27) Rohrbaug, R. H.; Jurs, P. C. Molecular Surface Area and the Prediction of High-Performance Liquid Chromatographic Retention Indexes of Polycyclic Aromatic Hydrocarbons. *Anal. Chem.* **1987**, *59*, 1048.
(28) 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*, 12−14.
(29) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Organics. *J. Phys. Chem.* **1996**, *100*, 10400−7.
(30) Dixon, S. L.; Jurs, P. C. Atomic Charge Calculations for Quantitative Structure−Property Relationships. *J. Computational Chem.* **1992**, *13*, 492−504.
(31) Abraham, R. H.; Smith, P. E. Charge Calculations in Molecular Mechanics IV: A General Method for Conjugated Systems. *J. Computational Chem.* **1987**, *9*, 288−297.
(32) 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.
(33) Topliss, J. G.; Edwards, R. P. Chance Gactors in Studies of Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238.
(34) Jurs, P. C.; Chou, J. T.; Yuan, M. in *Computer-Assisted Drug Design*; Olson, E. C., Christofferson, R. E., Eds.; The American Chemical Society: Washington, DC, 1979; pp 103−129.
(35) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley: New York, 1979.
(36) 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.
(37) Luke, B. T. Evolutionary Programming Applied to the Development of Quantitative-Structure Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279−1287.
(38) Shushen, L.; Chenzhong, C.; Zhiliang, L. Approach to Estimation and Prediction for Normal Boiling Point of Alkanes Based on a Novel Molecular Distance-Edge Vector. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387−394.
(39) Kier, L. B.; Hall, J. The E-State as an Extended Free Valence. *Comput. Sci. Chem. Inf.* **1997**, *37*, 548−552.
(40) Miller, K. J.; Savchik, J. A. A New Empirical Method to Calculate Average Molecular Polarizabilities. *J. Am. Chem. Soc.* **1979**, *101*, 7206−7213.
(41) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

CI000140S