Prediction of Supercritical Carbon Dioxide Solubility of Organic Compounds from Molecular Structure

Heidi L. Engelhardt and Peter C. Jurs*

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

Received June 27, 1996[⊗]

A diverse data set of 58 compounds taken from the literature was used to create models for the prediction of the solubility of organic compounds in supercritical carbon dioxide. Descriptors encoding information about the topological, geometric, and electronic properties of each compound in the data set were calculated from the molecular structures. A multiple linear regression model containing seven descriptors was generated. Several new descriptors, which were not present in the original pool, were calculated. One of the new descriptors was used to create the final seven descriptor linear model, which had a better root mean square (rms) error than the original model. The seven descriptors that appeared in the final model were used to make a neural network model which had a significantly better rms error than the linear model.

INTRODUCTION

The properties of supercritical fluids make them potentially good solvents. The density of a supercritical fluid (and thus its solvating power) can be adjusted by altering its pressure and temperature. At a higher density, a compound will dissolve in a supercritical fluid, and with a decrease in the density of the fluid, the solute can be easily recovered and the solvent can be recycled. Supercritical fluid extraction with carbon dioxide is an attractive method for purifying compounds because the solvent is nontoxic and has a relatively low critical temperature and pressure. One example of a process to which this technology is applied routinely is the decaffeination of coffee.[1] Recent investigations into other uses for supercritical carbon dioxide have been conducted in fields such as controlled drug release and environmental remediation.[2] Since solubility data are used in the evaluation of new uses for supercritical fluid technology,[2,3] methods for predicting solubilities in supercritical fluids are needed.

Several methods exist for estimating the solubility of compounds in supercritical carbon dioxide. Equations of state,[4] such as the Peng−Robinson equation,[5,6] the Redlich−Soave−Kwong equation,[7] and the augmented van der Waals equation,[8] have been used to model solubility in supercritical carbon dioxide. The predictive ability of equations of state depends upon the system being investigated.[4] Predictive methods using such equations frequently rely upon experimentally determined quantities, such as vapor pressure, which may not be available for structurally complex solutes.[4,9] In addition, they often rely upon approximations such as empirical mixing rules.[10]

Methods for estimating supercritical fluid solubility which require no experimental data for predicting solubilities have also been investigated. Politzer *et al.*[11,12] calculated solubilities with regression equations, at first for a data set of 9 compounds and later for 22 compounds, using descriptors encoding the size of the molecules and their ability to participate in electrostatic interactions. Battersby *et al.*[13] investigated the use of a neural network to predict the solubilities of 15 compounds in supercritical CO_2. Famini and Wilson[14] used calculated descriptors based on theoretical linear solvation energy relationships (TLSER)[15] to construct multiple linear regression models to predict the solubility of the same 22 compounds studied by Politzer.

In this work, we investigated the use of a quantitative structure−property relationship (QSPR) to predict the solubilities of 58 organic solutes in supercritical carbon dioxide. Numerical descriptors encoding the geometric, topological, and electronic properties of each compound were calculated from their molecular structures. The descriptors were then used to generate predictive models using multiple linear regression analysis and computational neural networks. The models were first judged by how accurately they predicted solubilities for the compounds used in creating them (the training set). As a final test, the models were used to predict the solubilities of an external prediction set of compounds not originally used to create the models.

EXPERIMENTAL SECTION

To create a QSPR, it is necessary to have a set of compounds and experimentally determined values of the property of interest for each compound. A search of the chemical literature revealed that the greatest amount of solubility data could be found for organics in supercritical CO_2 at a temperature of 308 K. Further investigation showed that the best experimental conditions for assembly of a data set would be 308 K and 140 bar pressure. Because the exact pressures at which data points were taken varied widely from source to source, a method of interpolation was needed.

Interpolation between data points taken at different pressures and at 308 K was performed using a regression of ln (solubility) vs ln (reduced density). Several papers show that an approximately linear relationship exists between ln (solubility) and ln (reduced density).[16−18] Reduced densities were calculated, using the Schmidt−Wagner equation of state,[19,20] from the temperature and pressure at which a data point was taken. Data points taken at a pressure of less than 100 bar were not used in creating the regressions, as many sources from which data were taken indicated a much greater

[⊗] Abstract published in *Advance ACS Abstracts*, March 15, 1997.

S0095-2338(96)00085-6 CCC: $14.00

uncertainty for data points taken at lower pressures. Solubilities for each compound were plotted vs the calculated reduced densities at which they were taken, and linear regressions were performed. Data points taken at other temperatures could not be used for these plots because different ln (solubility) vs ln (reduced density) lines occur at different temperatures. If there was more than one source of solubility values for a particular compound, multiple plots were made and the resulting solubility values were averaged to give the value used for this study.

To examine the validity of the assumption that the relationship between ln (solubility) and ln (reduced density) for the compounds used in this study is linear, ln (solubility) vs ln (reduced density) was plotted for each compound. For most of the compounds the assumption was satisfactory, but for a few of the compounds a second-order regression appeared to be more appropriate. From each regression, the natural log of the solubility in units of mole fraction was calculated and tabulated for the conditions 308 K and 140 bar. The solubilities of the compounds ranged from −3.55 (for diphenylmethane) to −11.83 log units (for stigmasterol). Table 1 lists the compounds used in this study, their supercritical CO_2 solubilities, and the references from which the experimental solubilities were derived. One compound, oxindole, may undergo tautomerization between keto and enol forms. A single conformation had to be adopted for this study. The enol form was chosen, as it was most consistent with our results, a decision supported by the results of Politzer.[12]

The data set contained a total of 58 compounds, which were split randomly into a training set of 52 compounds and an external prediction set of 6 compounds. The structures in the data set contained a diverse array of functional groups. Molecular structures representative of the types of compounds present in the training set are shown in Figure 1. The training set was used to create multiple linear regression models and the prediction set was used to validate the models. For development of the neural networks, the training set was further split into a training set of 47 compounds used to train the networks and a cross-validation set of 5 compounds used to prevent overtraining of the neural networks.

The compounds were sketched, and preliminary three-dimensional models were generated using HyperChem (Hypercube, Inc., Waterloo, ON) running on a 486 personal computer. The resulting structures were then transferred to a DEC 3000 AXP Model 500 workstation. The geometries of the structures were optimized using MOPAC[45] with the PM3 Hamiltonian.[46] The remainder of the work, with the exception of neural networks, was performed using the ADAPT[47] software system on the DEC workstation. ADAPT consists of a collection of programs, written by our group over the years, which perform all the necessary tasks that are part of creating a QSPR. Neural networks were run using in-house developed programs which are not part of ADAPT.

The next step in the process was descriptor generation. A total of 167 topological, electronic, geometric, and hybrid descriptors were calculated to represent each compound. Topological descriptors[48−52] include atom and bond counts, path counts,[49,50] and molecular weight. Electronic descriptors[53] encode information about the electronic aspects of molecular structure. Examples of such descriptors include the charge on the most positive and negative atoms.

Table 1. Compounds Used in Creating Models for Prediction of Solubility in Supercritical Carbon Dioxide

			ln (solubilty)		
no.	name	ln (solubility)[c]	pred by best linear model	pred by neural network	ref
1	5-aminoindole	−9.408	−9.509	−9.711	44
2	2-aminopyrazine[b]	−6.470	−8.134	−6.791	22
3	2-aminobenzoic acid[a]	−9.294	−9.101	−8.874	23
4	anthracene	−9.807	−8.463	−9.408	5
5	benzoic acid	−6.365	−6.631	−6.082	23, 24
6	2-chloropyrimidine	−4.066	−4.394	−5.123	22
7	cholesterol	−10.830	−11.203	−10.957	40
8	2,4-dichloronaphthol[b]	−7.837	−7.914	−8.711	25
9	2,3-dimethylnaphthalene	−5.460	−5.327	−5.428	26
10	2,6-dimethylnaphthalene	−5.798	−5.854	−5.922	26, 27
11	2,7-dimethylnaphthalene	−5.412	−5.841	−5.793	27
12	eicosanol	−8.643	−7.565	−8.000	28
13	fluorene	−6.442	−4.951	−5.906	8
14	hexachloroethane	−4.002	−4.849	−4.070	26
15	*o*-hydroxybenzoic acid[b]	−8.369	−9.549	−8.752	29
16	5-hydroxyindole[a]	−9.080	−9.006	−9.109	44
17	4-hydroxypyrimidine	−8.940	−7.695	−9.246	22
18	indole	−5.343	−6.433	−5.165	7
19	indole-3-aldehyde	−11.410	−10.874	−11.394	44
20	indole-3-carboxylic acid	−11.500	−10.932	−10.004	44
21	2-mercaptopyrimidine	−9.961	−8.367	−9.210	22
22	5-methoxyindole	−6.491	−8.127	−7.460	21
23	*m*-methoxyphenylacetic acid	−6.811	−7.309	−7.813	30
24	*o*-methoxyphenylacetic acid	−9.218	−7.438	−7.490	30
25	*p*-methoxyphenylacetic acid	−8.243	−7.557	−8.117	30
26	5-methoxytetralone	−5.393	−4.737	−4.666	31
27	6-methoxytetralone	−5.213	−5.023	−4.574	31
28	7-methoxytetralone	−3.653	−5.296	−4.711	31
29	methyl-*m*-nitrobenzoate	−4.616	−4.609	−4.836	31
30	methyl-*o*-nitrobenzoate	−3.804	−3.415	−4.366	31
31	methyl-*p*-nitrobenzoate	−4.983	−4.671	−4.599	31
32	dibenzofuran[a]	−5.575	−7.070	−6.003	32
33	myristic acid	−5.929	−7.333	−7.334	33
34	naphthalene	−4.271	−4.948	−4.055	7, 33, 34, 35, 36
35	a-naphthol	−6.733	−6.958	−6.957	37
36	b-naphthol	−7.728	−7.939	−7.888	23, 24, 37
37	octacosane	−9.566	−9.496	−9.457	34
38	octadecanoic acid	−9.446	−8.898	−9.007	28
39	oleic acid[a]	−6.615	−9.006	−9.256	38
40	oxindole (enol)[a]	−8.012	−9.248	−9.142	44
41	palmitic acid[b]	−7.924	−8.271	−8.241	33
42	phenylacetic acid	−5.220	−4.696	−4.631	39
43	phthalic anhydride	−6.444	−5.315	−6.321	23
44	progesterone	−8.394	−8.624	−8.839	3
45	pyrene	−8.940	−9.157	−10.027	8
46	skatole	−5.860	−5.609	−5.833	21
47	stigmasterol	−11.830	−11.902	−11.003	40
48	2,3′,4′,5-tetrachloro-biphenyl	−8.238	−5.711	−7.280	41
49	testosterone	−10.900	−10.422	−10.813	3
50	triacontane	−10.450	−10.707	−10.455	34
51	vanillin	−6.224	−7.364	−7.283	39
52	2,5-xylenol	−4.729	−5.073	−4.931	42
53	3,4-xylenol	−5.057	−4.831	−4.820	43
54	biphenyl	−4.566	−5.581	−4.692	36
55	diphenylmethane	−3.550	−3.680	−3.779	36
56	bibenzyl	−3.838	−3.663	−3.967	36
57	1-methylnaphthalene[a]	−3.628	−4.286	−4.262	36
58	2-methylnaphthalene[b]	−3.565	−5.045	−4.669	36

[a] Compounds in *p* set. [b] Compounds in cv set. All others in *t* set. [c] Solubilities are in units of mole fraction, for conditions of 308 K and 140 bar.

Geometric descriptors[54−56] encode information that depends upon the three-dimensional structure of a compound. An example of such a descriptor is the surface area of a

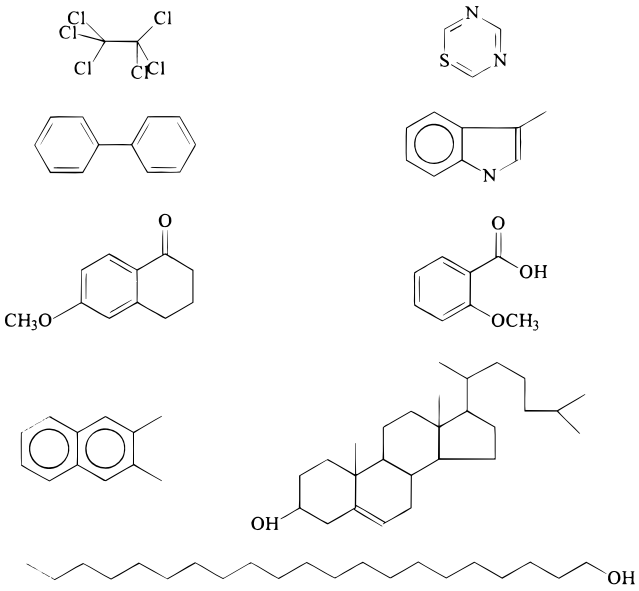

Figure 1. A sample of the compounds present in the training set, demonstrating the structural diversity of the data set used for this study.

molecule.[55] Finally, some hybrid descriptors encode information that represents both geometric and electronic information, such as charged partial surface area descriptors.[57]

To reduce the original pool of descriptors to an appropriate size, objective descriptor reduction was performed using various criteria. Reducing the pool of descriptors eliminates those descriptors which contribute either no information or whose information content is redundant with that of other descriptors present in the pool. Any descriptor that had identical or zero values for greater than 90% of the compounds was eliminated. Next, the collinearity of the descriptors was checked and one of each pair of descriptors with pairwise correlation coefficients exceeding 0.9 was eliminated. These methods reduced the pool of descriptors to 52.

Using the reduced pool of descriptors, multiple linear regression models were found using both a genetic algorithm[58,59] and a simulated annealing algorithm.[60] The overall root mean square (rms) error was used by the algorithms as the measure of quality of each model in their search for (nearly) global minima. The two algorithms found similar models. To determine the number of descriptors that should be used in creating models, the algorithms were used to generate several models with four descriptors, several models with five descriptors, and so on, with nine descriptors being the maximum subset size investigated. The maximum size subset investigated was dictated by the fact that the ratio of the number of compounds being used in a study to the number of descriptors in a model should be no less than 5. The average rms error for each subset size was calculated. The rms error of the models with fewer descriptors was compared to that of the models with more descriptors. The subset size chosen was the point at which the rms error did not increase significantly with the addition of another descriptor. A model size of seven descriptors was chosen, as this had an acceptable ratio of number of observations to number of descriptors and also gave the best marginal change in error for the addition of a descriptor.

Many models containing seven descriptors were generated. The models with the lowest rms errors were checked for statistical validity. The *p*-statistic was required to be no higher than 1×10^{-5}, the *t*-statistic was required to be no lower than 4, and the correlation coefficient was checked to ensure that is was greater than 0.90. Models with satisfactory statistical properties were kept for further consideration, while unsatisfactory models were eliminated. The models were next checked for multiple correlations. Any model with a multiple correlation coefficient among the descriptors greater than or equal to 0.95 was eliminated.[61]

Finally, the models were checked for the presence of outliers. The procedure for determining outliers is to run standard statistical checks[62] on the models. Statistics examined include the residuals, standardized residuals, studentized residuals, leverage points, DFITS values, Cook's distances, and Mahalanobis distances. Any observations flagged as outliers by four or more of the statistical tests were considered suspect points. Outliers were removed from the models and the coefficients were recalculated. If the models did not change substantially when the outliers were removed, they were readded to the model.

The model with the lowest rms error that passed all the statistical tests was chosen to be the final model. This final model was then used to predict the solubilities of an external prediction set of six compounds not used to generate the original models as a final means of model validation.

The first model found by the procedure outlined above had an rms error higher than desired. In an attempt to create a better model, several new descriptors were calculated. It was thought that perhaps one or more of the TLSER descriptors used by Famini[14] might capture information important for predicting supercritical CO_2 solubility, which had been absent from our original pool of descriptors. Famini used six descriptors; one, the molecular van der Waals volume, was not computed by his method; instead, we used a molecular volume descriptor already calculated by an ADAPT program. The five new descriptors were calculated from MOPAC's output, and the process described above for finding linear regression models was repeated.

Neural networks may yield better results than linear models if the relationship between the descriptors and the property of interest is nonlinear. In an attempt to create an improved model, the descriptors in the final linear model were used as input for a fully-connected, feed-forward computational neural network. The neural networks employ a Broyden−Fletcher−Goldfarb−Shanno[63] algorithm for optimization. A more detailed explanation of the neural networks used in this study has been given in a previous paper.[64] A 7−2−1 architecture (with 19 adjustable parameters) was chosen for the network as this was the only available option since any architecture with more hidden layer neurons would have had too many adjustable parameters. To avoid chance correlations, it is important that the ratio of observations to adjustable parameters be no less than 2.[65]

The training set of 47 compounds was used to create the neural network model. As the training of a neural network proceeds, it reaches a point at which it becomes more efficient at predicting the data used to train it, but its ability to generalize decreases. The cross-validation set is used to monitor the training of the network, and at the point at which the network begins to lose its ability to generalize, training is terminated. The neural network with the lowest rms error was chosen to be the final model. As with the multiple linear

Table 2. The Best Linear Regression Model before Calculation of New Descriptors[a]

descriptor	explanation	coefficient	std error
constant	intercept	8.72	1.5
DPOL1	calculated dipole moment	−0.514	0.11
PPSA3	sum of surface area of postively charged atoms times their charge	−0.224	0.020
S3C8	simple cluster molecular connectivity for paths of length 3	−1.82	0.25
SCAA1	sum of surface area times charge of hydrogen-bonding acceptors	−0.148	0.035
CTAA0	count of acceptor atoms	−2.11	0.32
SHDW5	area of molecule projected onto *XZ* plane divided by box of area defined by maximum *X* and *Z* dimensions	−9.32	1.6
NAB15	count of aromatic bonds	−0.284	0.051

[a] rms (linear) = 1.00, $R = 0.907$, $N = 52$, and $F = 29.1$.

Table 3. Best Model for Prediction of Solubility in Supercritical Carbon Dioxide[a]

descriptor	explanation	coefficient	std error
constant	intercept	22.9	2.8
PPSA3	sum of surface area of postively charged atoms times their charge	−0.254	0.022
RNCG1	charge on the most negative atom divided by the total negative charge	−8.09	1.9
S3C8	simple cluster molecular connectivity for paths of length 3	−1.68	0.23
CTAA0	count of acceptor atoms	−1.69	0.20
MCHG0	maximum charge difference between a donor and an acceptor	4.35	0.86
SHDW5	area of molecule projected onto *XZ* plane divided by box of area defined by maximum *X* and *Z* dimensions	−6.91	1.4
PVOL	polarization volume divided by molecular volume	−131	18

[a] rms (neural network) = 0.65, rms (linear) = 0.90, $R = 0.924$, $N = 52$, $F = 36.8$.

regression model, the final test of the model was its ability to predict the solubilities of the prediction set.

The rms error of a model is one of the primary features used in its evaluation. As it is meaningless to create a model with an rms error lower than the experimental error of the data used in creating it, an approximation of the error needed to be made. Many authors estimated the errors in their experimental supercritical CO_2 solubility values to be ~5%. As systematic errors may occur in the data,[38,66] and as an interpolation was performed to obtain the final data set, 10% was taken to be a realistic estimate of the error in the final solubility values obtained by interpolation. Since an rms error of 0.66 log unit corresponds to 10% of the average solubility value for the data set, that was chosen as the lower limit for rms error in the calculated solubilities. The rms errors for the neural network model are approximately equal to this experimental error estimate, while the rms errors for the linear models are higher. Thus, the models do not overfit the data.

RESULTS AND DISCUSSION

Two somewhat different multiple linear regression models were generated for the prediction of solubility in supercritical CO_2. The best linear regression model found using the descriptors available at the start of this study is shown in Table 2. This model had an rms errors of 1.00 log unit for the training set and 1.09 log units for the prediction set. While these results were encouraging, it was hoped that a more accurate model could be created, and the model building process was repeated using the original pool of descriptors plus five additional descriptors, as mentioned in the Experimental Section.

The best linear regression model found using the original pool of descriptors plus the five new descriptors is found in Table 3. It had rms errors of 0.90 log unit for the training set and 1.28 log units for the prediction set. One compound in the prediction set, oleic acid, had an unusually large residual of 2.39 log units, the largest residual of any compound in the data set. If the prediction is repeated excluding oleic acid, the rms error becomes a more reasonable 0.92 log unit. Justification for the exclusion of oleic acid from the prediction set will be made later with the discussion of the neural network model.

The first model and the second model are very similar, having four descriptors in common. The three descriptors that are different in each model encode similar types of information—the final model has a charge descriptor, a hydrogen-bonding descriptor, and a descriptor encoding polarizability, while the first model had a descriptor that is a count of the number of aromatic bonds in the molecule (which could be considered to be carrying information about the electronic structure of the molecule), a hydrogen-bonding descriptor, and a descriptor that is the calculated dipole moment of the molecule.

While no causal relationship exists between the descriptors in our best model and the solubility of a compound in supercritical CO_2, an examination of the descriptors may provide some insight about the factors influencing supercritical CO_2 solubility. The vapor pressure of the pure compound is highly correlated with the compound's solubility in supercritical CO_2. Solvent−solute interactions have a lesser effect upon the compound's solubility.[4] The importance of solute−solute interactions in the supercritical phase has been overlooked until recently; investigations into possible solute−solute interactions have found them to be more influential than previously thought.[67−69]

Molecular size has been found to influence supercritical solubility.[11] Two of our descriptors, S3C8 and SHDW5, are likely to be encoding size information. S3C8 is the simple cluster molecular connectivity descriptor for paths of length 3. SHDW5 is the area of the molecule projected onto the *XZ* plane divided by a box with dimensions corresponding to the maximum dimension of the molecule in the *X* and *Z* planes.

Hydrogen bonding may also affect solubility. Solutes may interact with one another, and weak hydrogen-bonding interactions between solutes and carbon dioxide may occur.[70,71] MCHG0 is a hydrogen-bonding descriptor encoding the maximum charge difference between a hydrogen-bonding donor and a hydrogen-bonding acceptor atom. The third descriptor in each model is CTAA0, the number of hydrogen-bonding acceptor atoms in each molecule.

Solutes may experience dispersion interactions, dipole-induced−dipole interactions and electrostatic interactions.[72] Electronic charge descriptors and charged partial surface area descriptors that appear in the best model are important for

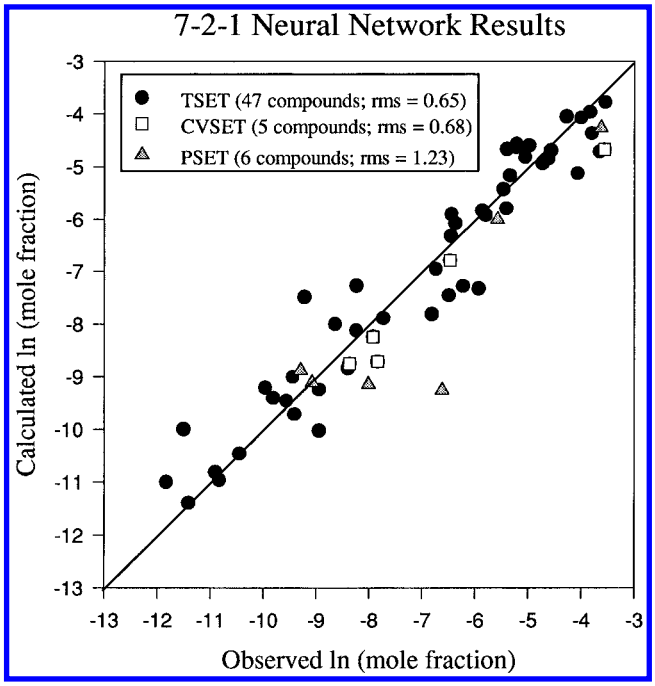

Figure 2. Plot of the solubilties calculated using the best neural network model vs the experimentally observed solubilities (at 308 K and 140 bar) for the tset and cvset compounds.

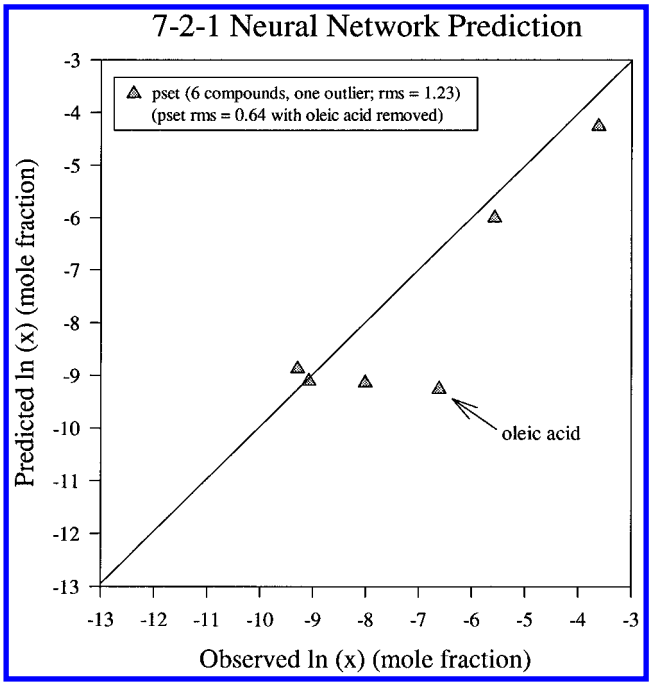

Figure 3. Plot of the solubilties calculated using the best neural network model vs the experimentally observed solubilities (at 308 K and 140 bar) for the pset compounds. The compound farthest from the line is oleic acid.

encoding these interactions. PPSA3 is the sum of the surface area of the positively charged atoms in a structure times their respective charges. RNCG1 is the charge on the most negative atom divided by the total negative charge of the molecule. PVOL is the polarization volume of the molecule (polarizability) divided by its molecular volume. This final descriptor was one of the TLSER descriptors used by Famini[14] and calculated by us in an attempt to improve our models.

The final linear model passed all statistical tests performed upon it and sufficiently predicted the solubilities of compounds in the prediction set. The first model and the final model have most of their descriptors in common. This lends confidence to the conclusion that our model is sound and not merely the result of chance correlations. The model performs suprisingly well despite the diverse array of compounds used to create and validate it.

When the descriptors from the best linear model were used to create a computational neural network model, the rms error of the training set fell to 0.65 log unit, while the cross-validation set error was 0.68 log unit. A plot of the calculated vs observed ln (solubility) for the training set and cross-validation set compounds as produced by the neural network model is shown in Figure 2. The line shown is the ideal 1:1 correlation line.

A plot of the predicted vs observed ln (solubility) for the prediction set compounds is shown in Figure 3. The prediction set rms error is high (1.23 log units) due to the presence of oleic acid, the only long-chain olefin present in the data set. If oleic acid is removed from the prediction set, the rms error for the remaining five compounds falls to 0.64 log unit, nearly identical to the rms errors for the training set and cross-validation set.

A possible contributing factor to the unusually high residual of oleic acid is the failure of the descriptors in the model to account for the structural features contributing to its solubility. Oleic acid is the only long-chain olefin in the data set. The descriptors chosen for the model are not encoding adequately information about the presence of a double bond in oleic acid, which results in its value being predicted as if it were the alkane octadecanoic acid,[38] a compound incidentally contained in the training set used for this study. Oleic acid is similar in structure to octadecanoic acid, differing only by a trans double bond in the middle of the hydrocarbon chain. The values of the descriptors calculated for both compounds are identical in three cases and very close in the other four cases.

A limited number of descriptors can be used to create models, and if a descriptor is useful for describing only one or a few compounds, it may be rejected in favor of descriptors useful for describing the majority of compounds, possibly leaving important information about a minority of compounds unaccounted for. Also, if, as was the case for oleic acid, a compound is present in the prediction set, but no compounds of adequate structural similarity are present in the training set, the prediction for the compound may be poor. Usually this difficulty is solved by assembling a homogeneous data set, or by subsetting the data into structurally similar classes. In the present case, the scarcity of data precluded the creation of a more homogeneous data set. It is important to note outliers (such as oleic acid), but due to the diversity of the data set, they cannot be discounted solely on the basis of insufficient structural commonalities with other compounds, as many compounds in the data set are unique in certain ways.

There is a second explanation for the large residual of oleic acid. It is possible that the experimental values we interpolated from were not very accurate. In the paper from which the data were derived,[38] the author showed graphs comparing his data and data published earlier by other authors. The graphs show significant variation among the data taken by different authors. Data taken at different temperatures is plotted in several graphs. A graph for 35

°C shows the solubility measured by another author to be lower than that reported by the author of the paper, indicating that the true solubility value may be lower than the value we used and that the lower solubility predicted by our model may not itself be as inaccurate as the residual would lead us to believe.

CONCLUSION

Multiple linear regression and neural network models satisfactory for predicting the solubility of a set of organic compounds in supercritical CO_2 were constructed. The neural network model had a significantly better rms error than the linear model. Five descriptors not present in the original pool of descriptors were calculated; one of the descriptors was chosen by an optimization routine in what turned out to be the best model.

The data set used was small and encompassed a wide variety of molecular structures, due to the nature of the data available in the literature. Considering the diversity and small size of the data set, it is encouraging that good models could be found. A more homogeneous data set may have produced excellent models, judging by the current results.

REFERENCES AND NOTES

(1) Bartle, K. D.; Clifford, A. A.; Jafar, S. A.; Shilstone, G. F. Solubilities of Solids and Liquids of Low Volatility in Supercritical Carbon Dioxide. *J. Phys. Chem. Ref. Data* **1991**, *20*, 713−756.

(2) *Supercritical Fluid Engineering Science: Fundamentals and Applications*; Kiran, E., Brennecke, J. F., Eds.; ACS Symposium Series 514; American Chemical Society: Washington, DC, 1993; Chapter 1.

(3) Kosal, K.; Lee, C. H.; Holder, G. D. Solubility of Progesterone, Testosterone, and Cholesterol in Supercritical Fluids. *J. Supercrit. Fluids* **1992**, *5*, 169−179.

(4) Johnston, K. P.; Peck, D. G.; Kim, S. Modeling Supercritical Mixtures: How Predictive Is It? *Ind. Eng. Chem. Res.* **1989**, *28*, 1115−1125.

(5) Kosal, E.; Holder, G. D. Solubility of Anthracene and Phenanthrene Mixtures in Supercritical Carbon Dioxide. *J. Chem. Eng. Data* **1987**, *32*, 148−150.

(6) Peng, D.-Y.; Robinson, D. B. A New Two-Constant Equation of State. *Ind. Eng. Chem. Fundam.* **1976**, *15*, 59−64.

(7) Sako, S.; Ohgaki, K.; Katayama, T. Solubilities of Naphthalene and Indole in Supercritical Fluids. *J. Supercrit. Fluids* **1988**, *1*, 1−6.

(8) Johnston, K. P.; Ziger, D. H.; Eckert, C. A. Solubilities of Hydrocarbon Solids in Supercritical Fluids. The Augmented van der Waals Treatment. *Ind. Eng. Chem. Fundam.* **1982**, *21*, 191−197.

(9) King, J. W.; Friedrich, J. P. Quantitative Correlations Between Solute Molecular Structure and Solubility in Supercritical Fluids. *J. Chromatogr.* **1990**, *517*, 449−458.

(10) Mitra, S.; Wilson, N. K. An Empirical Method to Predict Solubility in Supercritical Fluids. *J. Chromatogr. Sci.* **1991**, *29*, 305−309.

(11) Politzer, P.; Murray, J. S.; Lane, P.; Brinck, T. Relationships between Solute Molecular Properties and Solubility in Supercritical CO_2. *J. Phys. Chem.* **1993**, *97*, 729−732.

(12) Politzer, P.; Lane, P.; Murray, J. S; Brinck, T. Investigation of Relationships between Solute Molecule Surface Electrostatic Potentials and Solubilities in Supercritical Fluids. *J. Phys. Chem.* **1992**, *96*, 7938−7943.

(13) Battersby, P.; Dean, J. R.; Tomlinson, W. R.; Hitchen, S. M.; Myers, P. Predicting Solubility in Supercritical Fluid Extraction Using a Neural Network. *Analyst* **1994** *119*, 925−928.

(14) Famini, G. R.; Wilson, L. Y. Using Theoretical Descriptors in Structure−Activity Relationships: Solubility in Supercritical CO_2. *J. Phys. Org. Chem.* **1993**, *6*, 539−544.

(15) Lowrey, A. H.; Cramer, C. J.; Urban, J. J.; Famini, G. R. Quantum Chemical Descriptors for Linear Solvation Energy Relationships. *Comput. Chem.* **1995**, *19* (3), 209−215.

(16) Kumar, S. K.; Johnston, K. P. Modelling the Solubility of Solids in Supercritical Fluids with Density as the Independent Variable. *J. Supercrit. Fluids* **1988**, *1*, 15-22.

(17) Lee, C.; Ellington, R. T. Density-Based Correlation for Solid Solubility in Supercritical Solvents. *Sep. Sci. Technol.* **1987**, *22*, 1557−1576.

(18) Chrastil, J. Solubility of Solids and Liquids in Supercritical Gases. *J. Phys. Chem.* **1982**, *86*, 3016−3021.

(19) Ely, J. F. Proceedings of the 65th Annual Convention of the Gas Processors Association, San Antonio, TX, 1986; pp 185−192.

(20) Schmidt, R.; Wagner, W. A New Form of the Equation of State for Pure Substances and Its Application to Oxygen. *Fluid Phase Equilibr.* **1985**, *19*, 175−200.

(21) Sako, S.; Shibata, K.; Ohgaki, K.; Katayama, T. Solubilities of Indole, Skatole, and 5-Methoxyindole in Supercritical Fluids. *J. Supercrit. Fluids* **1989**, *2*, 3−8.

(22) Nakatani, T.; Tohdo, T.; Ohgaki, K.; Katayama, T. Solubilities of Pyrimidine and Pyrazine Derivatives in Supercritical Fluids. *J. Chem. Eng. Data* **1991**, *36*, 314-316.

(23) Dobbs, J. M.; Wong, J. M.; Lahiere, R. J.; Johnston, K. P. Modification of Supercritical Fluid Phase Behavior Using Polar Cosolvents. *Ind. Eng. Chem. Res.* **1987**, *26*, 56−65.

(24) Schmitt, W. J.; Reid, R. C. Solubility of Monofunctional Organic Solids in Chemically Diverse Supercritical Fluids. *J. Chem. Eng. Data* **1986**, *31*, 204−212.

(25) Yoon, J.-H.; Lee, H.-S.; Lee, H. Solubilities of 2,4-Dichloro-1-naphthol in Supercritical Carbon Dioxide. *J. Chem. Thermodyn.* **1993**, *25*, 193−196.

(26) Kurnik, R. T.; Holla, S. J.; Reid, R. C. Solubility of Solids in Supercritical Carbon Dioxide and Ethylene. *J. Chem. Eng. Data* **1981**, *26*, 47−51.

(27) Iwai, Y.; Mori, Y.; Hosotani, N.; Higashi, H.; Furuya, T.; Arai, Y.; Yamamoto, K.; Mito, Y. Solubilities of 2,6- and 2,7-Dimethylnaphthalenes in Supercritical Carbon Dioxide. *J. Chem. Eng. Data* **1993**, *38*, 509−511.

(28) Iwai, Y.; Koga, Y.; Maruyama, H.; Arai, Y. Solubilities of Stearic Acid, Stearyl Alcohol, and Arachidyl Alcohol in Supercritical Carbon Dioxide at 35 °C. *J. Chem. Eng. Data* **1993**, *38*, 506−508.

(29) Gurdial, G. S.; Foster, N. R. Solubility of o-Hydroxybenzoic Acid in Supercritical Carbon Dioxide. *Ind. Eng. Chem. Res.* **1991**, *30*, 575−580.

(30) Lee, H.-K.; Kim, C.-H. Solid Solubilities of Methoxyphenylacetic Acid Isomer Compounds in Supercritical Carbon Dioxide. *J. Chem. Eng. Data* **1994**, *39*, 163-165.

(31) Chang, H.; Morrell, D. G. Solubilities of Methoxy-1-tetralone and Methyl Nitrobenzoate Isomers and Their Mixtures in Supercritical Carbon Dioxide. *J. Chem. Eng. Data* **1985**, *30*, 74−78.

(32) Hansen, P. C. Binary Supercritical Fluid Enhancement Factors for Separation Processes. Ph.D. Thesis, University of Illinois, Normal, IL, 1985; pp 19−32.

(33) Iwai, Y.; Fukuda, T.; Koga, Y.; Arai, Y. Solubilities of Myristic Acid, Palmitic Acid, and Cetyl Alcohol in Supercritical Carbon Dioxide at 35 °C. *J. Chem. Eng. Data* **1991**, *36*, 430−432.

(34) Reverchon, E.; Russo, P.; Stassi, A. Solubilities of Solid Octacosane and Triacontane in Supercritical Carbon Dioxide. *J. Chem. Eng. Data* **1993**, *38*, 458−460.

(35) McHugh, M.; Paulaitis, M. E. Solid Solubilities of Naphthalene and Biphenyl in Supercritical Carbon Dioxide. *J. Chem. Eng. Data* **1980**, *25*, 326−329.

(36) Chung, S. T.; Shing, K. S. Multiphase Behavior of Binary and Ternary Systems of Heavy Aromatic Hydrocarbons With Supercritical Carbon Dioxide. *Fluid Phase Equil.* **1992**, *81*, 321−341.

(37) Tan, C.-S.; Weng, J.-Y. Solubility Measurements of Naphthol Isomers in Supercritical CO_2 By a Recycle Technique. *Fluid Phase Equilib.* **1987**, *34*, 37−47.

(38) Foster, N. R.; Yun, S. L. J.; Ting, S. S. T. Solubility of Oleic Acid in Supercritical Carbon Dioxide. *J. Supercrit. Fluids* **1991**, *4*, 127−130.

(39) Wells, P. A.; Chaplin, R. P.; Foster, N. R. Solubility of Phenylacetic Acid and Vanillin in Supercritical Carbon Dioxide. *J. Supercrit. Fluids* **1990**, *3*, 8−14.

(40) Wong, J. M.; Johnston, K. P. Solubilization of Biomolecules in Carbon Dioxide Based Supercritical Fluids. *Biotechnol. Prog.* **1986**, *2*, 29−39.

(41) Yu, E.; Richter, M.; Chen, P.; Wang, X.; Zhang, Z.; Tavlarides, L. L. Solubilitites of Polychlorinated Biphenyls in Supercritical Carbon Dioxide. *Ind. Eng. Chem. Res.* **1995**, *34*, 340−346.

(42) Iwai, Y.; Yamamoto, H.; Tanaka, Y.; Arai, Y. Solubilities of 2,5- and 2,6-Xylenols in Supercritical Carbon Dioxide. *J. Chem. Eng. Data* **1990**, *35*, 174−176.

(43) Mori, Y.; Shimizu, T.; Iwai, Y.; Arai, Y. Solubilities of 3,4-Xylenol and Naphthalene + 2,5-Xylenol in Supercritical Carbon Dioxide at 35 °C. *J. Chem. Eng. Data*, **1992**, *37*, 317−319.

(44) Nakatani, T.; Ohgaki, K.; Katayama, T. Solubilities of Indole Derivatives in Supercritical Fluids. *J. Supercrit. Fluids* **1989**, *2*, 9−14.

(45) Stewart, J. P. P. Mopac 6.0, Quantum Chemistry Program Exchange; Indiana University, Bloomington, IN, Program 455.

(46) Stewart, J. P. P. Mopac: A semi-empirical molecular orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1.

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

(48) Kier, L. B. Distinguishing Atom Differences in a Molecular Graph Shape Index. *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.* **1986**, *5*, 7−12.
(49) Randic, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L. Search for all Self-Avoiding Paths for Molecular Graphs. *Comput. Chem.* **1979**, *3*, 5.
(50) Wiener, H. J. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17.
(51) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Research Studies Press, Ltd., John Wiley and Sons: New York, 1986; pp. 18−20.
(52) Randic, M. On Molecular Identification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164−175.
(53) Dixon, S. L. Ph.D. Thesis, The Pennsylvania State University, University Park, PA, 1994.
(54) Goldstein, H. *Classical Mechanics*; Addison-Wesley: Reading, MA, 1950; pp 144−156.
(55) Perlman, R. S. Molecular Surface Area and Volumes and Their Use in Structure/Activity Relationships In *Physical Chemical Properties of Drugs*; Yalkowsky, S. H., Sinkula, A. A., Valvani, S. C., Eds.; Marcel Dekker Inc.: New York, 1980; Chapter 10.
(56) 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.
(57) 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.
(58) Lucasius, C. B.; Kateman, G. Understanding and Using Genetic Algorithms Part 1. Concepts, Properties and Context. *Chemom. Intell. Lab. Sys.* **1993**, *19*, 1.
(59) Hibbert, D. B. Genetic Algorithms in Chemistry. *Chemom. Intell. Lab. Sys.* **1993**, *19*, 227.
(60) Sutter, J. M.; Jurs, P. C. in *Data Handling in Science and Technology. Adaptation of Simulated Annealing to Chemical Optimization Problems*; Kalivas, J. H., Ed., Elsevier: Amsterdam, 1995; Vol. 15, Chapter 5.
(61) Stanton, D. T.; Jurs, P. C.; Hicks, M. G. Computer-Assisted Prediction of Normal Boiling Points of Furans, Tetrahydrofurans, and Thiophenes. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 301.
(62) Belsey, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics*; Wiley: New York, 1980.
(63) Xu, L.; Ball, J. W.; Dixon, S. L; Jurs, P. C. Quantitative Structure−Property Relationships for Toxicity of Phenols Using Regression Analysis and Computational Neural Networks. *Environ. Toxicol. Chem.* **1994**, *13*, 841.
(64) 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.
(65) Livingstone, D. J.; Manallack, D. T. Statistics Using Neural Networks: Chance Effects. *J. Med. Chem.* **1993**, *36*, 1295−1297.
(66) Yau, J.-S.; Tsai, F.-N. Solubilities of Heavy *n*-Paraffins in Subcritical and Supercritical Carbon Dioxide. *J. Chem. Eng. Data* **1993**, *38*, 171−174.
(67) Tsugane, H.; Yagi, Y.; Inomata, H.; Saito, S. Dimerization of Benzoic Acid in Saturated Solution of Supercritical Carbon Dioxide. *J. Chem. Eng. Jpn.* **1992**, *25*, 351−353.
(68) Chialvo, A. A.; Debenedetti, P. G. Molecular Dynamics Study of Solute−Solute Microstructure in Attractive and Repulsive Supercritical Mixtures. *Ind. Eng. Chem. Res.* **1992**, *31*, 1391−1397.
(69) Brennecke, J. F.; Tomasko, D. L.; Peshkin, J.; Eckert, C. A. Fluorescence Spectroscopy Studies of Dilute Supercritical Solutions. *Ind. Eng. Chem. Res.* **1990**, *29*, 1682−1690.
(70) Gupta, R. B.; Combes, J. R.; Johnston, K. P. Solvent Effect on Hydrogen Bonding in Supercritical Fluids. *J. Phys. Chem.* **1993**, *97*, 707−715.
(71) Fulton, J. L.; Yee, G. G.; Smith, R. D. Hydrogen Bonding of Methyl Alcohol-*d* in Supercritical Carbon Dioxide and Supercritical Ethane Solutions. *J. Am. Chem. Soc.* **1991**, *113*, 8327−8334.
(72) Nakatani, T.; Ohgaki, K.; Katayama, T. Substituent Effect on Solubilities of Solids in Supercritical Fluids. Naphthalene Derivatives. *Ind. Eng. Chem. Res.* **1991**, *30*, 1362−1366.

CI960085G