# Improved Quantitative Structure Property Relationships for the Prediction of Dielectric Constants for a Set of Diverse Compounds by Subsetting of the Data Set

Robert C. Schweitzer* and Jeffrey B. Morris

Army Research Laboratory, AMSRL-WM-BD, Aberdeen Proving Grounds, Maryland 21005-5066

Received February 4, 2000

In a recent publication we explored the development of quantitative structure property relationships for the calculation of dielectric constants, which resulted in a general model for a wide range of compounds. Our current work explores the division of the set of compounds into eight more homogeneous subsets for which local models are developed. The full data set consists of 454 compounds with dielectric constants ranging from 1 to 40. A pool of up to 16 molecular descriptors is calculated for each of the eight data sets. The descriptors include dipole moment, polarizability, counts of elemental types or functional groups, charged partial surface area, and molecular connectivity. All possible 4−16 descriptor models are calculated for each of the eight data sets, and the best models are selected and compared to the results obtained from the best general model for all 454 compounds. Neural networks using the Broyden−Fletcher−Goldfarb−Shanno training algorithm are employed to build the models. The resulting combined mean test set error for the eight local models of 1.31 is significantly better than the mean test set error of 1.85 for the general model.

## 1. INTRODUCTION

The dielectric constant, $\epsilon$, is an important fundamental physical property. It is defined as the ratio of the permittivity of a substance to the permittivity of free space and is in essence a measure of the ability of a substance to solvate a charged species. The utility of the dielectric constant for the interpretation of solvent−solute behavior motivates the desire for an availability of dielectric constant data. A large body of theory has been developed for the calculation of dielectric constants from such properties as dipole moment and polarizability.[1] Unfortunately, the current state of theory does not permit the accurate calculation of dielectric constants for a wide range of compounds, especially when those compounds exert strong intermolecular forces.

Our approach to this problem is the use of quantitative structure property relationships (QSPRs).[2−3] A QSPR is a type of empirical or calibration model. Empirical model building can be divided into a series of five steps. The first step in the process is the selection of the dependent variable. This is simply the property that one is interested in calculating−the dielectric constant for instance. The second step is the selection of the independent variables. For a QSPR, the independent variables are generally molecular descriptors that are calculated directly from the structure of the molecule. It is important that the molecular descriptors chosen adequately encode the relationship between the chemical structure and the property of interest. The third step is the selection of the calibration set, which is a set of compounds for each of which the dielectric constant is known and the molecular descriptors can be calculated. It is important that the members of the calibration set cover a wide range of chemical structures and that they cover that range completely and evenly. A good way to visualize this range is to mathematically or mentally plot the compounds as points in an *n*-dimensional data space, where each coordinate axis corresponds to one of the independent variables used to build the model. The fourth step is the use of some mathematical technique to actually develop the mathematical relationship between the independent variables and the dependent variable. In this work, a multilayer, feed-forward neural network is used. The final step is the testing of the model for its ability to make accurate predictions for a set of compounds that are not included in the calibration set. It is important that each of these five steps be accomplished successfully.

In a previous paper, a QSPR was developed for a very diverse set of 497 compounds.[4] Although a useful model was developed, we concluded that the division of the set of compounds into several smaller more homogeneous sets might yield smaller, more focused models better able to make accurate predictions for the compounds in the test set. The purpose of the work presented in this paper is to test that hypothesis.

## 2. EXPERIMENTAL SECTION

In this work, a subset of 454 compounds with condensed-phase static dielectric constants was taken from the set of 497 compounds used in the previous study.[4] There were several reasons for the removal of the 43 compounds. The majority of these compounds were duplicates. Some of the compounds had very infrequently occurring functional groups such as oximes, hydrazines, pyrazines, and isocyanates. Four of the compounds had gas-phase rather than liquid-phase dielectric constants. Dielectric constant data for the set of 454 compounds were found in the *CRC Handbook of Chemistry and Physics* and the *Handbook of Organic Chemistry*.[5,6] Many of these data stemmed originally from

* To whom correspondence should be addressed at ChemIcon Inc., 7301 Penn Ave., Pittsburgh, PA 15208. Phone: (412) 241-4754. Fax: (412) 241-7311. E-mail: schweitz@chemimage.com

10.1021/ci0000070 CCC: $19.00
Published on Web 08/24/2000

**Table 1.** Division of the Full Data Set into Subsets

| subset | functional groups | no. of members | dielectric constant range |
|---|---|---|---|
| 1 | hydrocarbons | 66 | 1.84−3.00 |
| 2 | halogenated hydrocarbons (Cl and Br) | 81 | 2.14−11.37 |
| 3 | ethers | 27 | 2.24−22.60 |
| 4 | esters and carboxylics | 85 | 2.29−20.00 |
| 5 | amines | 37 | 2.40−18.30 |
| 6 | ketones, aldehydes, acid halides, anhydrides, and cyclic esters | 43 | 3.50−39.10 |
| 7 | alcohols | 71 | 2.46−37.72 |
| 8 | nitros, nitriles, and amides | 44 | 2.90−38.30 |

the NBS Circular 514.[7] Some of the descriptors require 3-dimensional coordinates, and these were obtained from the Center for Intelligent Instrumentation located at Ohio University.[8] The 3-dimensional coordinates had been calculated using the molecular mechanics package, MM2, developed by Allinger at the University of Georgia.[9] Gaussian 94 was used to calculate dipole moments, polarizabilities, and partial atomic charges.[10] The Hartree−Fock self-consistent field calculation with the 6-31G basis set was used to carry out these calculations.

The computational work for this project was performed on a cluster of five Silicon Graphics Origin 2000’s running under IRIX and located at the Army Research Laboratory at the Aberdeen Proving Grounds, MD. The neural network software and the molecular descriptor software used for this work were developed using FORTRAN 77. The database software was supplied by the Center for Intelligent Chemical Instrumentation at Ohio University.

## 3. RESULTS AND DISCUSSION

**3.1. Data Subsets.** An important step in the research is the division of the full set of 454 compounds into subsets of similar compounds for which more focused and more accurate models can be built. Each of the 454 compounds was categorized using 20 different functional groups, with many of the compounds containing multiple functional groups. The importance of the effect of each functional group on the dielectric constant was subjectively rated by an examination of sets of compounds that have the same functional groups. This procedure allowed groupings of compounds such that the functional groups of the compounds in a given group appeared to have similar effects on the dielectric constants. For example, hydrocarbons with no functional groups have the smallest dielectric constants, while compounds with nitros, amides, and nitriles have the largest dielectric constants. The resulting eight subsets of compounds are displayed in Table 1 in order of increasing priority. Several of the functional groups occur in such a small number of compounds that it was necessary to combine multiple functional groups into one subset so that reliable models could be built. The ether subset has the smallest number of entries−27. This was deemed acceptable, although the models developed for this subset will not be as robust as models developed for the hydrocarbon subset, for example. If a compound has multiple functional groups that belong to different subsets, that compound is assigned to the subset that corresponds to its highest priority functional group. Subsets 1 and 2 are the most restrictive and have no compounds that overlap with other subsets. Subset 3 contains some chlorinated compounds. All of the compounds in subset 3, except for three cyclic ethers and one dichlorinated compound, have dielectric constants of less than 8.0. Subset 4 includes some chlorinated and brominated compounds, as well as some compounds with ether groups. All of the compounds in subset 4, except for chloroacetic acid ($\epsilon = 20.0$), have dielectric constants of less than 13.0, and the majority of the compounds have dielectric constants of less than 9.0. Subset 5 contains some chlorinated and brominated compounds and one compound with an ether group. Ethyleneimine, which is a three-membered ring, has the highest dielectric constant in this subset. While the majority of the compounds that comprise subset 5 have dielectric constants in the range of 2.0−7.0, subset 5 also contains six compounds with dielectric constants greater than 10.0−primarily diamines and halogenated anilines. Subset 6 includes some chlorinated and brominated compounds, as well as some compounds with ester groups. Subset 7 contains some compounds with ester, ketone, carboxylic, aldehyde, ether, amine, and chlorine functional groups. Subset 8 contains some compounds with chlorine, bromine, carboxylic, alcohol, amine, and ester functional groups. The dielectric constants for subsets 6−8 are fairly evenly spaced across their respective ranges.

**3.2. Molecular Descriptors.** Hundreds of descriptors have been used by various researchers in the development of QSPRs. In our previous work, a total of 65 molecular descriptors which are very representative of the commonly used descriptors were considered for use as independent variables.[4] The descriptors used were divided into three groups. The first group included counts of the number of occurrences of a particular element (H, C, N, O, Cl, and Br) in the compound, the dipole moment, polarizability, and ability or inability of the compound to form a hydrogen bond. The second group of descriptors were charged partial surface area (CPSA) descriptors. The CSPA class of descriptors was introduced by Jurs[11] and combines surface areas with partial atomic charges. The third group of descriptors were molecular connectivity descriptors, which were introduced by Randic[12] and improved and expanded by Kier and Hall.[13−16] The molecular connectivity descriptors describe the degree of branching in a compound.

The set of 65 descriptors is a substantial pool of descriptors. Some of the descriptors, such as molecular connectivity and counts of atom types, are commonly used in many of the QSPRs reported in the literature. Other descriptors were selected because of their particular relevance to the prediction of dielectric constants. Dipole moment, polarizability, CPSA, and hydrogen bond descriptors fall into this category.

The number of descriptors in the pool was reduced to avoid the possibility of building models with chance correlations with the dielectric constant. The first step in this reduction process was to remove descriptors that did not span a wide range of values or that were highly correlated with other descriptors. The remaining descriptors were further reduced to a set of 16 descriptors by selecting the descriptors that best correlated with the dielectric constant and collectively covered the widest range of variance. These 16 descriptors are listed in Table 2. Six additional descriptors deemed useful for particular data subsets are listed in Table 2 as well. Table

**Table 2.** Set of 22 Molecular Descriptors

| descriptor | label | description |
|---|---|---|
| 1 | dipole | dipole moment |
| 2 | polar | polarizability |
| 3 | C | count of the number of carbons |
| 4 | O | count of the number of oxygens |
| 5 | N | count of the number of nitrogens |
| 6 | H-bond | presence of a hydroxyl group or a primary or secondary amine |
| 7 | PNSA-2[a] | sum of the surface area for negatively charged atoms × sum of negative charges |
| 8 | FPSA-1[a] | sum of the surface area for positively charged atoms/total surface area |
| 9 | FNSA-2[a] | PNSA-2/total surface area |
| 10 | FNSA-3[a] | sum of each (surface area for a negatively charged atom × negative charge)/total surface area |
| 11 | WNSA-3[a] | sum of each (surface area for a negatively charged atom × negative charge) × total surface area/1000 |
| 12 | $^{0}\chi^{v}$ [b] | connectivity for atoms |
| 13 | $^{2}\chi^{v}$ [b] | connectivity for two-bond paths (paths with three atoms) |
| 14 | $^{3}\chi^{v}$ [b] | connectivity for three-bond paths (paths with four atoms) |
| 15 | $^{7}\chi^{v}_{c}$ [b] | connectivity for clusters with six-bond paths |
| 16 | $^{7}\chi^{v}$ [b] | connectivity for seven-bond paths |
| 17 | Cl | count of the number of chlorines |
| 18 | Br | count of the number of bromines |
| 19 | OH-1 | count of the number of primary hydroxyl groups |
| 20 | OH-2 | count of the number of secondary hydroxyl groups |
| 21 | NH-1 | count of the number of primary amine groups |
| 22 | NH-2 | count of the number of secondary amine groups |

[a] Symbols are taken from Jurs.[11] [b] Symbols are taken from Kier and Hall.[13−14]

**Table 3.** Pool of Molecular Descriptors Used for Each Data Subset

| data sets | | descriptors |
|---|---|---|
| 1 | hydrocarbons | 1−3, 7−16 |
| 2 | hydrocarbons with Cl and Br | 1−3, 7−18 |
| 3 | ethers | 1−4, 7−16 |
| 4 | esters and carboxylics | 1−4, 7−16 |
| 5 | amines | 1−3, 5−16, 21−22 |
| 6 | ketones, etc. | 1−4, 7−16 |
| 7 | alcohols | 1−4, 7−16, 19−20 |
| 8 | nitros, etc. | 1−16 |
| 9 | full data set | 1−16 |

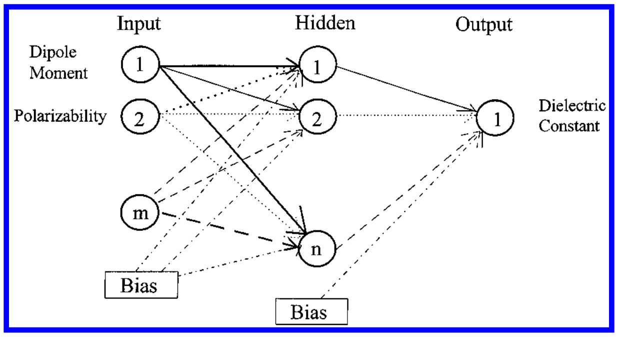

**Figure 1.** Representation of a three-layer neural network. Circles represent nodes, rectangles represent bias nodes, and arrows represent connections.

3 lists the pool of descriptors used for the exploratory experiments for each of the data sets listed in Table 1, as well as those for the full data set.

**3.3. Neural Networks.** Neural networks are a computational technique, that among other purposes, can be used to build calibration models. Neural networks have been extensively reviewed.[17−28] They are loosely patterned after the connection of neurons in the brain and consist of one input layer, one or more hidden layers, and one output layer as illustrated in Figure 1. Each pair of neurons in adjacent layers is connected, and each connection has a weight value, $W^{k}_{ij}$ , associated with it, where $k$ represents the second layer in the connection, $i$ represents the node in the first layer, and $j$ represents the node in the second layer. In addition, each layer has one bias node that is connected to each of the nodes in the following layer and is assigned a value of 1.0. The bias is analogous to an offset in regression analysis and gives added flexibility to the model. The weight values are analogous to the regression coefficients of regression analysis, and the process of training a neural network is an iterative process in which the collective weight values that produce the minimum error for the calibration set are determined. The training process requires many training cycles in which each training cycle consists of the presentation of each compound in the calibration set to the neural network. For a given compound, the molecular descriptors are calculated, and these calculated values are propagated forward through the neural network. The calculated value for the dielectric constant in the output layer of the neural network is compared to the actual value of the dielectric constant, and an error is calculated. The error is propagated backward through the neural network so that some fraction of the error is associated with each connection in the neural network. A training algorithm then uses these distributed errors to make a small adjustment in each of the weight values. There are many algorithms which can be used. The most widely used method is the steepest descent method introduced by Rummelhart[29] and popularly known as the “back-propagation” method. The method used in this work is the Broyden−Fletcher−Goldfarb−Shanno (BFGS) algorithm.[30−37] The BFGS method is a quasi-Newton method employing estimated second derivatives and a line search routine and is a much more efficient training algorithm than the steepest descent approach. The process of training a neural network is an optimization process, and as such a local minimum can be obtained rather than a global minimum. To maximize the chance of finding a global minimum, the training process is repeated 200 times with randomly assigned initial weight values, and the best neural network of the 200 is selected. The training process also requires the determination of the point at which the neural network has been sufficiently trained. To avoid the danger of overtraining, a monitoring set is used. Each of the steps in the neural network training

**Table 4.** Division of Compounds into Training, Monitoring, and Test Sets

| data set | subset 1 | subset 2 | subset 3 | subset 4 | subset 5 | subset 6 | subset 7 | subset 8 | full data set |
|---|---|---|---|---|---|---|---|---|---|
| training | 42 | 53 | 19 | 57 | 26 | 30 | 46 | 31 | 304 |
| monitoring | 14 | 16 | 4 | 17 | 7 | 7 | 15 | 7 | 87 |
| test | 10 | 12 | 4 | 11 | 4 | 6 | 10 | 6 | 63 |
| total | 66 | 81 | 27 | 85 | 37 | 43 | 71 | 44 | 454 |

**Table 5.** Best Model for Each Data Subset

| data set | | descriptors used |
|---|---|---|
| 1 | hydrocarbons | dipole, polar, C, FPSA-1, FNSA-3, WNSA-3, ${}^0\chi^\nu$, ${}^2\chi^\nu$, ${}^7\chi^\nu{}_c$ |
| 2 | halogenated hydrocarbons | dipole, polar, Cl, Br, FPSA-1, FNSA-3, WNSA-3, ${}^0\chi^\nu$, ${}^3\chi^\nu$, ${}^7\chi^\nu{}_c$ |
| 3 | ethers | dipole, O, PNSA-2, FPSA-1, FNSA-2, FNSA-3 |
| 4 | esters and carboxylics | dipole, polar, O, FPSA-1, FNSA-2, FNSA-3, WNSA-3, ${}^7\chi^\nu{}_c$ |
| 5 | amines | dipole, polar, C, NH-1, FPSA-1, FNSA-3, WNSA-3, ${}^2\chi^\nu$ |
| 6 | ketones, etc. | dipole, polar, C, PNSA-2, FNSA-2, ${}^2\chi^\nu$, ${}^7\chi^\nu{}_c$, ${}^7\chi^\nu$ |
| 7 | alcohols | dipole, C, O, OH-1, OH-2, PNSA-2, FPSA-1, FNSA-3, WNSA-3, ${}^3\chi^\nu$, ${}^7\chi^\nu{}_c$ |
| 8 | nitros, etc. | dipole, C, O, N, FNSA-2, WNSA-3, ${}^3\chi^\nu$, ${}^7\chi^\nu$ |
| 9 | full data set | dipole, polar, O, N, H-bond, FPSA-1, FNSA-3, WNSA-3, ${}^0\chi^\nu$ |

process is more fully explained in our previous paper[4] and in the references cited above.

**3.4. Model Building Parameters.** The neural networks developed for this work consist of an input layer, one hidden layer with six nodes, which includes one node for the bias, and an output layer with a single node. The number of nodes in the input layer is the number of molecular descriptors plus one node for the bias. The number of nodes in the hidden layer was varied in some preliminary studies, and it was determined that six nodes were sufficient to describe the nonlinear relationship between the inputs and the output. It is important that the number of nodes in the hidden layer is not so large that the models developed are overparametrized. The input values for a given molecular descriptor for all the compounds in the training set are scaled to fit a range of −1.0 to +1.0. The hyperbolic tangent is the activation function used for both the hidden layer and the output layer. A mean squared error is used for the calculation of the error at the output layer. Most of the reported error values, however, are calculated as a mean of the absolute values of the errors in the training, monitoring, or test set.

The compounds for the training, monitoring, and test set are selected so that the widest possible range of compounds is present in the training set, the next widest range of compounds makes up the monitoring set, and the remaining compounds form the test set. The selection process is based on some work by Small and Carpenter[38] and consists of a data space dimensionality reduction step using principal component analysis,[39−40] a partitioning step, and a compound selection step. A more thorough explanation can be found in our previous paper.[4] Table 4 lists the number of compounds selected for the training, monitoring, and test sets for each of the nine data sets. Approximately 65% of the compounds were selected for the training sets, 15% for the monitoring sets, and the remaining 20% for the test sets.

**3.5. Determination of the Best Models for Each Data Set.** For each of the eight data subsets all possible $4 - n$ descriptor models were built for the $n$ descriptors chosen for these subsets as listed in Table 3. Totals of 7814, 32 192, 15 914, 15 914, 64 839, 15 914, 64 839, and 64 839 models were built for data subsets 1−8, respectively. Because of the large number of models to be evaluated, the neural network training process was repeated only 10 times for each model. Instead of selecting the single best neural network from each model, averages of the 10 training, monitoring, and test sets were calculated and used to determine which models gave the best results. For each of the eight data subsets, approximately 1000 models were selected that had the lowest average training set and test set errors.

**Table 6.** Test Set Errors Associated with the Top 20 Models

| data set | | mean of test set errors | std dev of test set errors |
|---|---|---|---|
| 1 | hydrocarbons | 0.039 | 0.013 |
| 2 | halogenated hydrocarbons | 0.956 | 0.243 |
| 3 | ethers | 0.524 | 0.227 |
| 4 | carboxylics and esters | 1.037 | 0.205 |
| 5 | amines | 1.129 | 0.323 |
| 6 | carbonyls, etc. | 2.203 | 0.823 |
| 7 | alcohols | 3.016 | 0.822 |
| 8 | nitros, etc. | 4.951 | 1.976 |
| 9 | full data set | 1.800 | 0.319 |

The full neural network training in which 200 neural networks are calculated and the best neural network is selected was performed for the approximately 1000 models for the eight data subsets. This methodology is described in greater detail in our previous paper.[4] For the full data set of 454 compounds, the full neural network training was performed for the 293 models found to give the best results in the previous study for the set of 497 compounds.[4] For each of the data subsets, the models were ranked on the basis of training set and monitoring set errors and the model with the lowest errors was selected. To avoid overfitting, preference was given to the selection of a model with the fewest number of descriptors. The selected models are listed in Table 5.

There is a significant amount of variation in the test set error for the best models for each of the eight data subsets. Table 6 lists the mean and standard deviation of the test set errors for the top 20 models for each of the subsets as well as for the full data set. The nitros, alcohols, and carbonyls have the highest means and standard deviations for the test set errors. One possible reason for these errors is that the test sets selected are not perfect representations of the monitoring sets or the training sets. The underlying problem is that the points are not evenly enough distributed in the data space. This problem results from the use of a database with a limited number of compounds, rather than the

**Table 7.** Partitioning of the Data Spaces for the Selection of the Training Sets

| | data subset | no. of occupied blocks | no. of total blocks | percentage occupied |
|---|---|---|---|---|
| 1 | hydrocarbons | 31 | 320 | 9.69 |
| 2 | halogenated hydrocarbons | 45 | 2880 | 1.56 |
| 3 | ethers | 18 | 720 | 2.50 |
| 4 | carboxylics and esters | 55 | 1080 | 5.09 |
| 5 | amines | 26 | 504 | 5.16 |
| 6 | carbonyls, etc. | 25 | 2880 | 0.87 |
| 7 | alcohols | 46 | 9720 | 0.47 |
| 8 | nitros, etc. | 30 | 3600 | 0.83 |
| 9 | full data set | 281 | 1069200 | 0.026 |

generation of experimental data from a well-defined research program in which the compounds selected for measurement are selected in such a manner that they evenly cover the range of chemical structure types and dielectric constant values. Such an experimental program was unfortunately far beyond the practical scope of this research project, and it was felt that useful models could still be developed with the available data.

One way to look at the distribution of the data points is to examine the results of the partitioning that is used to select the data points for the training, monitoring, and test sets. Table 7 shows this partitioning for the selection of the compounds for the training sets for each of the models selected as the best models for the eight data subsets and the full data set. Column 2 lists the number of occupied blocks in the partitioned data space, column 3 lists the total number of blocks in the partitioned data space, and column 4 lists the percentage of occupied blocks in the partitioned data space. A comparison of the percentages of occupied blocks in Table 7 to the mean errors and standard deviations in Table 6 shows that the data sets with higher errors tend to be the same data sets that have very unevenly spaced compounds in the corresponding data space (as indicated by low percentages of occupied blocks). As already discussed, larger numbers of compounds that more evenly cover the data spaces spanned by the independent variables would result in better models and would decrease the variation of the test set errors for the top models, especially for the nitros, alcohols, and carbonyls.

The results for the full data set are an apparent exception to this trend. The partitioning shows a very low percentage of the data space as occupied, indicating a need for a better distribution of data points. Indeed, this poor distribution was the original impetus for the division of the full data set into data subsets. The mean error and the standard deviation of the mean error for the test sets, however, are relatively good compared to those of the eight data subsets. One contributing factor is the fact that the mean test set error is an average of 454 values rather than 30−80 values. An even more important factor is revealed by an examination of the compounds that were actually selected for the test set for the general model. The compounds selected for a given test set tend to be very similar to one another because they are selected after the training and monitoring sets are selected and represent a consequently small range of variance in the data set. Consider the combined set of all the compounds in the eight individual test sets. Because these compounds were selected from each of the eight subsets, the combined test set will consist of compounds from each of the eight subsets and will be fairly representative of the full set of 454 compounds. There is no such assurance for the compounds selected for the general test set from the full data set, and as a result, the compounds in this general test set are much less representative of the full data set than the compounds in the combined test set. The compounds in the general test set are dominated by compounds that are easy to predict (such as hydrocarbons) and have inherently low errors.

**Table 8.** Results for the Data Subset of Hydrocarbons

| | general model | | | local model | | |
|---|---|---|---|---|---|---|
| | mean of absolute errors | mean of relative errors (%) | no. of compds | mean of absolute errors | mean of relative errors (%) | no. of compds |
| overall | 0.349 | 16.52 | 66 | 0.014 | 0.70 | 66 |
| training | 0.318 | 13.87 | 30 | 0.008 | 0.38 | 42 |
| monitor | 0.261 | 11.96 | 7 | 0.022 | 1.12 | 10 |
| test | 0.403 | 20.35 | 29 | 0.026 | 1.34 | 14 |

**Table 9.** Results for the Data Subset of Halogenated Hydrocarbons

| | general model | | | local model | | |
|---|---|---|---|---|---|---|
| | mean of absolute errors | mean of relative errors (%) | no. of compds | mean of absolute errors | mean of relative errors (%) | no. of compds |
| overall | 1.491 | 25.72 | 81 | 0.349 | 5.75 | 81 |
| training | 1.626 | 30.08 | 51 | 0.085 | 1.66 | 53 |
| monitor | 1.345 | 18.16 | 18 | 0.579 | 10.26 | 12 |
| test | 1.139 | 18.53 | 12 | 1.055 | 15.91 | 16 |

**Table 10.** Results for the Data Subset of Ethers

| | general model | | | local model | | |
|---|---|---|---|---|---|---|
| | mean of absolute errors | mean of relative errors (%) | no. of compds | mean of absolute errors | mean of relative errors (%) | no. of compds |
| overall | 2.915 | 62.21 | 27 | 0.153 | 3.02 | 27 |
| training | 3.007 | 60.94 | 21 | 0.088 | 0.70 | 19 |
| monitor | 1.445 | 52.16 | 1 | 0.287 | 8.23 | 4 |
| test | 2.822 | 69.54 | 5 | 0.330 | 8.86 | 4 |

**Table 11.** Results for the Data Subset of Esters and Carboxylics

| | general model | | | local model | | |
|---|---|---|---|---|---|---|
| | mean of absolute errors | mean of relative errors (%) | no. of compds | mean of absolute errors | mean of relative errors (%) | no. of compds |
| overall | 1.683 | 37.04 | 85 | 0.447 | 10.18 | 85 |
| training | 1.623 | 32.93 | 65 | 0.248 | 5.73 | 57 |
| monitor | 1.911 | 50.24 | 13 | 0.659 | 12.67 | 11 |
| test | 1.812 | 50.72 | 7 | 0.978 | 23.49 | 17 |

**Table 12.** Results for the Data Subset of Amines

| | general model | | | local model | | |
|---|---|---|---|---|---|---|
| | mean of absolute errors | mean of relative errors (%) | no. of compds | mean of absolute errors | mean of relative errors (%) | no. of compds |
| overall | 1.326 | 21.34 | 37 | 0.351 | 7.02 | 37 |
| training | 1.470 | 24.44 | 29 | 0.060 | 1.61 | 26 |
| monitor | 1.112 | 12.42 | 5 | 0.470 | 9.88 | 4 |
| test | 0.286 | 6.28 | 3 | 1.366 | 25.46 | 7 |

Tables 8−15 compare the results for each data subset as obtained by the best local model for each data subset and by the best general model for the full set of 454 compounds. Table 16 examines the results for the full data set as a combination of the local models compared to the general

**Table 13.** Results for the Data Subset of Ketones, Etc.

| | general model | | | local model | | |
|---|---|---|---|---|---|---|
| | mean of absolute errors | mean of relative errors (%) | no. of compds | mean of absolute errors | mean of relative errors (%) | no. of compds |
| overall | 4.477 | 25.90 | 43 | 0.480 | 4.88 | 43 |
| training | 4.584 | 27.92 | 30 | 0.169 | 1.67 | 30 |
| monitor | 3.562 | 25.26 | 4 | 0.949 | 12.19 | 6 |
| test | 4.529 | 19.44 | 9 | 1.411 | 12.39 | 7 |

**Table 14.** Results for the Data Subset of Alcohols

| | general model | | | local model | | |
|---|---|---|---|---|---|---|
| | mean of absolute errors | mean of relative errors (%) | no. of compds | mean of absolute errors | mean of relative errors (%) | no. of compds |
| overall | 3.082 | 33.35 | 71 | 0.894 | 11.81 | 71 |
| training | 3.293 | 34.87 | 46 | 0.279 | 5.39 | 46 |
| monitor | 1.900 | 16.39 | 8 | 2.285 | 23.46 | 10 |
| test | 3.066 | 37.22 | 17 | 1.853 | 23.74 | 15 |

**Table 15.** Results for the Data Subset of Nitros, Nitriles, and Amides

| | general model | | | local model | | |
|---|---|---|---|---|---|---|
| | mean of absolute errors | mean of relative errors (%) | no. of compds | mean of absolute errors | mean of relative errors (%) | no. of compds |
| overall | 3.424 | 16.12 | 44 | 0.996 | 4.42 | 44 |
| training | 3.495 | 17.35 | 32 | 0.158 | 0.66 | 31 |
| monitor | 3.338 | 15.11 | 7 | 1.175 | 4.06 | 6 |
| test | 3.088 | 9.67 | 5 | 4.556 | 21.41 | 7 |

**Table 16.** Results for the Full Data Set

| | general model | | | local model | | |
|---|---|---|---|---|---|---|
| | mean of absolute errors | mean of relative errors (%) | no. of compds | mean of absolute errors | mean of relative errors (%) | no. of compds |
| overall | 2.151 | 28.59 | 454 | 0.468 | 6.52 | 454 |
| training | 2.318 | 29.85 | 304 | 0.148 | 2.65 | 304 |
| monitor | 1.757 | 24.06 | 63 | 0.842 | 10.76 | 63 |
| test | 1.854 | 27.47 | 87 | 1.314 | 17.00 | 87 |

model. For Tables 8−16, the results for the general model are given in columns 2−4 and the results for the local models are given in columns 5−7. The general model was developed without regard to the functional groups of the compounds in the data set. For the purposes of these tables, however, the compounds along with their predicted versus actual dielectric constants for the general model were sorted into the eight data subsets for which the individual local models were developed. The resulting residuals were used to calculate the statistics recorded in columns 2−4. As can be seen from the column headers, means of the absolute values of the errors and means of the relative errors are calculated for each of the training, monitoring, and test sets, as well as for the combined sets of all compounds in the training, monitoring, and test sets. It should be noted that the number of compounds in column 4 does not match the number of compounds in column 7 because the selection of the compounds for the training, monitoring, and test sets for the general model was based on the entire set of compounds rather than the eight data subsets as determined by functional group.

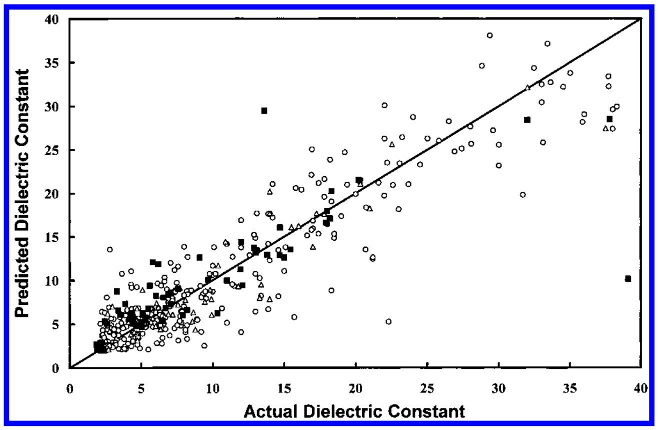

**Figure 2.** Predicted versus actual dielectric constants for the general model. The members of the training set are represented by open circles, the members of the monitoring set are represented by open triangles, and the members of the test set are represented by filled squares.

An examination of Tables 8−15 shows that, for each of the eight subsets, the local models yield better results than the general model for the mean of the absolute errors and the mean of the relative errors of the training set. The alcohol data subset (see Table 14) is the only subset for which the local model gives worse results for the monitoring set for absolute or relative errors. The nitros (see Table 15) and the amines (see Table 12) are the only two data subsets for which the local model gives worse results than the general model for the test sets. The comparison for the amine data set, however, is not a very fair one because only three of the compounds in the test set for the general model are amines, while seven compounds were selected for the test set for the local model for the amine data subset. It should also be noted that the results for the test set for the local versus general model for the halogenated hydrocarbon data subset (see Table 9) are very similar.

A comparison of the general model versus the combined local models for the full data set (see Table 16) gives a clearer picture of the improved accuracy obtained by dividing the data set into subsets and using a local model for each subset. The results for the combined local models are much better than the results for the general model. The impact of these results is especially significant for the test set because the test set for the general model has a large number of hydrocarbons (see Table 8), which, as already discussed, gives an advantage to the general model.

The predicted versus actual dielectric constants for the set of all 454 compounds are displayed in Figure 2 (the general model) and Figure 3 (the combined local models), where the members of the training set are represented by open circles, the members of the monitoring set by open triangles, and the members of the test set by filled squares. The diagonal line drawn through both plots illustrates where the points would fall if there were no error. A comparison of these two plots clearly demonstrates a large improvement for the training and monitoring sets by use of the eight local models. There is some degree of scatter for the points in the test set relative to the training and monitoring sets for the local models (Figure 3), but these results are much better than the results for the test set for the general model (Figure 2). For the combined local models, there are 2, 6, 24, 41, and 57 compounds in the test set with relative errors greater

**Table 17.** Most Frequently Used Descriptors for the Top Exploratory Models

| subset 1, hydrocarbons | | subset 2, halogenated hydrocarbons | | subset 3, ethers | | subset 4, esters, etc. | | subset 5, amines | |
|---|---|---|---|---|---|---|---|---|---|
| descriptor | freq | descriptor | freq | descriptor | freq | descriptor | freq | descriptor | freq |
| all | 961 | all | 960 | all | 955 | all | 947 | all | 960 |
| dipole | 910 | dipole | 929 | dipole | 824 | o | 831 | polar | 900 |
| polar | 898 | polar | 762 | FPSA-1 | 682 | polar | 721 | dipole | 712 |
| FNSA-3 | 599 | PNSA-2 | 649 | O | 629 | FPSA-1 | 632 | ${}^{0}\chi^{v}$ | 650 |
| FPSA-1 | 553 | ${}^{7}\chi^{v}$ | 587 | FNSA-3 | 554 | FNSA-3 | 627 | H-bond | 646 |
| C | 543 | ${}^{7}\chi^{v}_{c}$ | 583 | WNSA-3 | 536 | dipole | 611 | N | 627 |
| WNSA-3 | 533 | FNSA-3 | 574 | FNSA-2 | 533 | FNSA-2 | 573 | ${}^{3}\chi^{v}$ | 582 |
| ${}^{0}\chi^{v}$ | 533 | WNSA-3 | 569 | PNSA-2 | 497 | C | 514 | WNSA-3 | 574 |
| ${}^{3}\chi^{v}$ | 521 | FPSA-1 | 527 | C | 480 | WNSA-3 | 477 | FNSA-2 | 518 |
| FNSA-2 | 502 | C | 523 | ${}^{0}\chi^{v}$ | 450 | ${}^{0}\chi^{v}$ | 463 | FNSA-3 | 501 |
| PNSA-2 | 500 | ${}^{2}\chi^{v}$ | 514 | ${}^{2}\chi^{v}$ | 428 | PNSA-2 | 456 | ${}^{2}\chi^{v}$ | 485 |
| ${}^{7}\chi^{v}$ | 438 | FNSA-2 | 505 | ${}^{3}\chi^{v}$ | 419 | ${}^{3}\chi^{v}$ | 428 | C | 464 |
| ${}^{2}\chi^{v}$ | 437 | ${}^{0}\chi^{v}$ | 503 | polar | 348 | ${}^{2}\chi^{v}$ | 416 | ${}^{7}\chi^{v}_{c}$ | 411 |
| ${}^{7}\chi^{v}_{c}$ | 436 | ${}^{3}\chi^{v}$ | 498 | ${}^{7}\chi^{v}_{c}$ | 269 | ${}^{7}\chi^{v}$ | 412 | PNSA-2 | 382 |
| | | Cl | 445 | ${}^{7}\chi^{v}$ | 160 | ${}^{7}\chi^{v}_{c}$ | 353 | NH-1 | 359 |
| | | Br | 332 | | | | | FPSA-1 | 288 |
| | | | | | | | | NH-2 | 279 |

| subset 6, ketones, etc. | | subset 7, alcohols | | subset 8, nitros, etc. | | set 9,[a] full data set | |
|---|---|---|---|---|---|---|---|
| descriptor | freq | descriptor | freq | descriptor | freq | descriptor | freq |
| all | 965 | all | 987 | all | 959 | all | 191 |
| dipole | 891 | PNSA-2 | 855 | O | 862 | dipole | 191 |
| O | 666 | dipole | 707 | dipole | 693 | H-bond | 190 |
| FNSA-2 | 572 | ${}^{0}\chi^{v}$ | 650 | WNSA-3 | 652 | N | 188 |
| polar | 569 | WNSA-3 | 632 | ${}^{7}\chi^{v}_{c}$ | 611 | O | 175 |
| FPSA-1 | 565 | FPSA-1 | 631 | PNSA-2 | 595 | WNSA-3 | 152 |
| PNSA-2 | 540 | O | 605 | FPSA-1 | 562 | FNSA-3 | 151 |
| C | 523 | FNSA-2 | 561 | polar | 519 | PNSA-2 | 95 |
| WNSA-3 | 521 | ${}^{3}\chi^{v}$ | 554 | FNSA-2 | 516 | FNSA-2 | 93 |
| FNSA-3 | 487 | FNSA-3 | 530 | N | 497 | C | 92 |
| ${}^{2}\chi^{v}$ | 474 | polar | 521 | FNSA-3 | 484 | ${}^{0}\chi^{v}$ | 89 |
| ${}^{0}\chi^{v}$ | 445 | ${}^{2}\chi^{v}$ | 505 | H-bond | 432 | FPSA-1 | 85 |
| ${}^{7}\chi^{v}$ | 412 | OH-1 | 443 | C | 431 | ${}^{2}\chi^{v}$ | 58 |
| ${}^{3}\chi^{v}$ | 384 | C | 426 | ${}^{7}\chi^{v}$ | 375 | polar | 44 |
| ${}^{7}\chi^{v}_{c}$ | 363 | ${}^{7}\chi^{v}$ | 357 | ${}^{0}\chi^{v}$ | 345 | ${}^{3}\chi^{v}$ | 43 |
| | | OH-2 | 322 | ${}^{3}\chi^{v}$ | 318 | ${}^{7}\chi^{v}_{c}$ | 38 |
| | | ${}^{7}\chi^{v}_{c}$ | 312 | ${}^{2}\chi^{v}$ | 241 | ${}^{7}\chi^{v}$ | 28 |

[a] From Table 10 of ref 4.

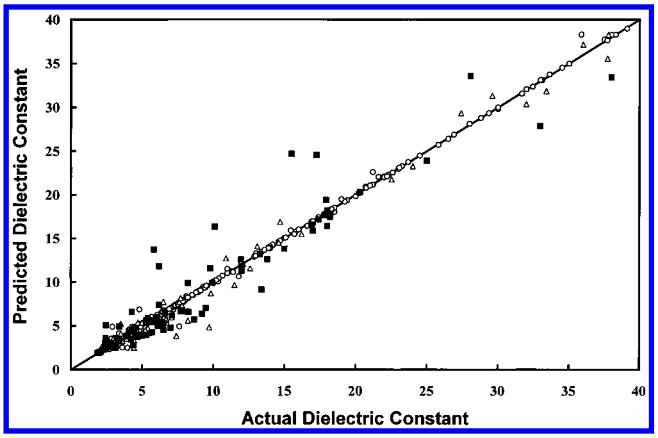

**Figure 3.** Predicted versus actual dielectric constants for the combined local models. The members of the training set are represented by open circles, the members of the monitoring set are represented by open triangles, and the members of the test set are represented by filled squares.

than 100%, 50%, 20%, 10%, and 5%, respectively. The corresponding results for the general model are 4, 11, 41, 58, and 77 compounds. A Mann−Whitney test was performed to compare the test set results for the combined local models versus the general model. The results of this test showed that a mean error of 1.31 (the combined test sets) is statistically smaller than a mean error of 1.85 (the general test set). Highlighting these results is the fact that the majority of the compounds (for example, hydrocarbons) in the general test set should be easier to predict accurately than the compounds in the test set for the combined local models.

**3.6. Cluster Analysis.** Some cluster analysis was used to study the distribution of the compounds for the best model for each of the nine data sets. The number of neighbors within a certain Euclidean squared distance (0.005, 0.01, and 0.05) was calculated for each compound in the data sets. It was observed that the compounds with the largest errors in the test sets for each of the data sets do not necessarily correspond to compounds in the most sparse area of the respective data space. It is likely, however, that fuller data sets with compounds evenly covering the data space would result in test sets with smaller errors and better reproducibility between similar models. Cluster analysis was able to reveal some data sets that should be further subdivided. For instance, the compounds in data set 8 tend to form two separate clusters. One of the clusters contains amides and nitriles, and the other cluster contains only nitros.

**3.7. Most Frequently Used Descriptors.** A study was performed of the descriptors that are most useful for building models for each of the nine data sets. Table 17 lists the descriptors used for each of the data sets in order of importance. The number of models in which a particular descriptor was used in the best approximately 1000 models is listed in parentheses following each descriptor code. The results for the full data set in Table 17 (set 9) are taken from our previous paper for a set of 191 models.[4] The actual number of exploratory models used for each data set is listed in the first row of Table 17.

Different data sets do indeed use different descriptors. The dipole moment is the first or second descriptor in order of importance (frequency of use) for all data sets except the esters (see Table 17), and is used in each of the models selected as best in Table 5. Polarizability is important for most of the data sets for which the majority of compounds have low dielectric constants. For high dielectric constant compounds, the polarizability loses its importance. The CPSA descriptors are useful for most of the data sets. The molecular connectivity descriptors find some use, but are not as important as the CPSA descriptors. The count of the number of oxygens is important for the ethers, esters, ketones, nitros, and the full data set. The added descriptors—counts of Cl, Br, OH-1, OH-2, NH-1, and NH-2—found relatively little use, although some of the top models (Table 5) do use some of these descriptors. It can be seen that the descriptors found to be most useful for the full data set are descriptors which distinguish different classes of compounds from one another. The descriptors for data sets 1−8, however, are tailored more to each particular data set.

## 4. CONCLUSION

Figures 2 and 3 and Tables 8−16 clearly show that the division of the set of 454 compounds into eight subsets—based on functional group—allows the development of local targeted models that result in much more accurate predictions of a test set of compounds than does the general model. There is still an uneven distribution of compounds in the data spaces defined by the independent variables for each of the eight best models. Although not as severe as the distribution for the general model, this uneven distribution results in the development of similar models with a significant variance of test set errors. Further improvement would most likely be obtained by the addition of compounds to improve the distribution of the points in the data spaces. Given a sufficient number of additional compounds, some of the eight subsets could be further divided into more homogeneous subsets. The CPSA descriptors were very useful in the local models as well as the general model. Some added simple descriptors such as counts of Cl, OH-1, OH-2, and NH-1 were selected for the best models, although they were not predominant in the top 1000 models.

## ACKNOWLEDGMENT

This work was supported in part by a grant of high-performance computing (HPC) time on the SGI Origin 2000 cluster from the DoD HPC Center at the Army Research Laboratory at Aberdeen Proving Grounds, MD. Dr. Gary Small of the Center for Intelligent Chemical Instrumentation at Ohio University is acknowledged for providing the database of three-dimensional coordinates and the accompanying software. R.C.S. acknowledges support through the National Academy of Sciences/National Research Council Postdoctoral Associateship Program at the Army Research Laboratory.

## REFERENCES AND NOTES

(1) Hasted, J. *Aqueous Dielectrics*; Chapman and Hall: London, 1973.

(2) Tute, M. In *Comprehensive Medicinal Chemistry*; Hansch, C., Ed.; Pergamon Press: Oxford, 1990; Vol. 4, pp 1−31.

(3) Hansch, C. A Quantitative Approach to Biochemical Structure−Activity Relationships. *Acc. Chem. Res.* **1969**, *2*, 232−239.

(4) Schweitzer, R. C.; Morris, J. B. The Development of a Quantitative Structure Property Relationship (QSPR) for the Prediction of Dielectric Constants using Neural Networks. *Anal. Chim. Acta* **1999**, *384*, 285−303.

(5) *CRC Handbook of Chemistry and Physics*; Weast, R., Ed.; CRC Press: Boca Raton, FL, 1985; pp E-49−E-74.

(6) *Handbook of Organic Chemistry*; Dean, J., Ed.; McGraw-Hill: New York, 1987; pp 4-45−4-79.

(7) *Table of Dielectric Constants of Pure Liquids*; NBS Circular 514; National Bureau of Standards: Washington, DC, 1951.

(8) Schweitzer, R.; Small, G. Enhanced Structural Encoding Algorithm for Database Retrievals of Carbon-13 Nuclear Magnetic Resonance Chemical Shifts. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 310−322.

(9) Burket, U.; Allinger, N. *Molecular Mechanics*; American Chemical Society: Washington, DC, 1982.

(10) Frisch, M.; Frisch, A.; Foresman, J. *Gaussian 94 User's Reference*; Gaussian: Pittsburgh, 1995.

(11) Stanton, D. L.; 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.

(12) Randic, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6614.

(13) Kier, L.; Hall, L. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(14) Kier, L.; Hall L. *Molecular Connectivity in Structure−Activity Analysis*; Chemometrics Series; Research Studies Press, Wiley: New York, 1986; Vol. 9.

(15) Randic, M. Chemical Structure - - What is "She"? *J. Chem. Educ.* **1992**, *69*, 713−718.

(16) Hansen P.; Jurs, P. Chemical Applications of Graph Theory Part I. Fundamentals and Topological Indices. *J. Chem. Educ.* **1988**, *65*, 574−580.

(17) Zupan, J.; Gasteiger, J. Neural Networks: A New Method for Solving Chemical Problems or just a Passing Phase? *Anal. Chim. Acta* **1991**, *248*, 1−30.

(18) Jansson P. Neural Networks: An Overview. *Anal. Chem.* **1991**, *63*, 357A-362A.

(19) Masters, T. *Practical Neural Network Recipes in C++*; Academic Press: San Diego, 1993.

(20) Masters, T. *Advanced Algorithms for Neural Networks: A C++ Sourcebook*; Wiley: New York, 1995.

(21) Jones, W.; Hoskins, J. Back-Propagation: A Generalized Delta Learning Rule. *BYTE* **1987**, 155−162.

(22) Wythoff, B. Back-propagation in Neural Networks. A Tutorial. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 115−155.

(23) Spining, M.; Darsey, J.; Sumpter, B.; Noid, D. Opening Up the Black Box of Artificial Neural Networks. *J. Chem. Educ.* **1994**, *71*, 406−411.

(24) Sumpter, B.; Getino, C.; Noid, D. Theory and Applications of Neural Computing in Chemical Science. *Annu. Rev. Phys. Chem.* **1994**, *45*, 439−481.

(25) Svozil, D.; Kvasnicka, V.; Pospichal, J. Introduction to Multi-Layer Feed-Forward Neural Networks. *Chemom. Intell. Lab. Syst.* **1997**, *39*, 43−62.

(26) Long, J.; Gregoriou, V.; Gemperline, P. Spectroscopic Calibration and Quantitation Using Artificial Neural Networks *Anal. Chem.* **1990**, *62*, 1791−1797.

(27) Egolf, L.; Jurs, P. Prediction of Boiling Points of Organic Heterocyclic Compounds Using Regression and Neural Network Techniques. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 616−625.

(28) Zupan, J.; Gasteiger, J. *Neural Networks for Chemists: An Introduction;* VCH: Weinheim, 1993.

(29) Rumelhart, D.; Hinton, G.; Williams, R. In *Microstructures of Cognition;* Rumelhart, D. E., McClelland, J. L., Eds.; MIT Press: Cambridge, 1986; Vol. 1, Chapter 8.

(30) Broyden, C. The Convergence of a Class of Double-Rank Minimization Algorithms. *J. Inst. Math. Appl.* **1970**, *6*, 76−90, 222−231.

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

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

(33) Shanno, D. Conditioning of Quasi-Newton Methods for Function Minimization. *Math. Comput.* **1970**, *24*, 647−656.

(34) Nocedal, J. Updating Quasi-Newton Matrixes with Limited Storage. *Math. Comput.* **1980**, *35*, 773−782.

(35) Xu, L.; Ball, J.; Dixon, S.; Jurs, P. Quantitative Structure−Activity Relationships for Toxicity of Phenols Using Regression Analysis and Computational Neural Networks. *Environ. Toxicol. Chem.* **1994**, *13*, 841−851.

(36) Egolf, L.; Wessel, M.; Jurs, P. Prediction of Boiling Points and Critical Temperatures of Industrially Important Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 947−956.

(37) Wessel, M.; Jurs, P. Prediction of Reduced Ion Mobility Constants from Structural Information Using Multiple Linear Regression Analysis and Computational Neural Networks. *Anal. Chem.* **1994**, *66*, 2480−2487.

(38) Carpenter, S.; Small, G. Selection of Optimum Training Sets for Use in Pattern Recognition Analysis of Chemical Data. *Anal. Chim. Acta* **1991**, *249*, 305−321.

(39) Spragg, R.; Lidiard, D.; Rupert, A. Principal Component Analysis. *Chem. Br.* **1991**, 821−824.

(40) Wold, S.; Geladi, P.; Esbensen, K. Principal Component Analysis *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37−52.

CI0000070