Determination of Abraham Solute Parameters from Molecular Structure

Jesús Jover, Ramón Bosque, and Joaquim Sales*

Departament de Química Inorgànica, Universitat de Barcelona, Martí i Franquès, 1, 08028-Barcelona, Spain

Received February 6, 2004

The Abraham solute parameters are well-known factors for the quantitative description of solute/solvent interactions. A quantitative structure−property relationship (QSPR) is reported for the *E*, *S*, *A*, and *B* parameters of a large set of 457 solutes, of very different chemical nature. The proposed models, derived from multilinear regression analysis (MLRA) and computational neural networks (CNN), contain five descriptors calculated solely from the molecular structure of compounds. Good correlations were obtained for the four parameters studied, and the corresponding values of R^2 and standard deviations are better or similar than those derived from other theoretical bases. All models were validated by external prediction sets. The proposed QSPR models, both by MLRA and CNN, contain analogous descriptors encoding similar information, that agree with the accepted physicochemical meaning of the Abraham parameters; however, some descriptors which encode information that is not associated with this physicochemical meaning are also included in the QSPR models.

INTRODUCTION

The solute/solvent interactions are of major importance in chemistry and biochemistry according to the fact that, mainly, chemical reactions occur in solution. One of the most used methods to study these interactions is through empirical equations that relate a selected property with parameters of solutes and/or solvents.

Abraham and co-workers[1] have proposed the general solvation equation

$$\log \mathrm{SP} = c + eE + sS + aA + bB + vV$$

to correlate solute properties (SP), such as partitioning,[2,3] solubility,[4] characterization of the selectivity of micellar electrokinetic chromatography systems,[5] blood-brain distribution,[6] and human intestinal absorption,[7] with a standard set of five parameters. In this equation, *E* is an excess molar refraction that is obtained from the refractive index. *S* is the dipolarity/polarizability that can be obtained from gas−liquid chromatographic measurements on polar stationary phases or more generally from water/solvent partitions. The parameters *A* and *B* are the overall or effective hydrogen bond acidity and basicity, respectively, which are most easily obtained from water−solvent partitions. *V* is the McGowan characteristic volume that can promptly be calculated from bond and atom contributions.

These parameters represent the solute influence on various solute/solvent phase interactions. Hence, the coefficients *c, e, s, a, b*, and *v*, which are obtained via multiple linear regression against known log SP values, correspond to the complimentary effect on the phases on these interactions. The coefficients can be regarded as system constants which characterize and contain chemical information of the phase in question and can be interpreted as follows. The *e*-coefficient shows the tendency of the phase to interact with solutes through π and *n*-electron pairs. Usually the *e*-coefficient is positive, but for a phase which contains fluorine atoms, it can be negative. The *s*-coefficient represents the tendency of the phase to interact with dipolar/polarizable solutes. The *a*-coefficient denotes the hydrogen bond basicity of the phase, because acidic solutes will interact with basic phases, and the *b*-coefficient is a measure of the hydrogen bond acidity of the phase. The *v*-coefficient is a measure of the hydrophobicity of the phase, and it describes the dispersion interactions and cavitation forces.

Any application of the general solvation equation depends on the availability of the solute parameters, and the need to calculate them for new compounds will always be of primary importance. As explained earlier, the descriptors *E* and *V* can be calculated quite simply from structure, but the remaining three descriptors *S*, *A*, and *B* have to be determined experimentally, either directly from complexation measurements or indirectly via back-calculations from partition measurements. Then, it is not surprising that different attempts have been made to avoid the obtention of experimental data for the determination of new *S*, *A*, and *B* values. Such attempts include the work of Sevcik and co-workers[8] who have reported multilinear regression and neural network approaches to estimate the *S* parameter from a set of 333 compounds using 29 molecular descriptors. Platts et al. using ab initio and DFT methods have estimated *S*,[9] *A*,[10] and *B*[11] Abraham parameters for sets of 50−80 compounds. More recently, the same authors have also applied DFT methods to the estimation of *A* and *B* parameters for multifunctional acids and bases.[12] On the other hand, an additive model for the estimation of the five solute parameters *E, S, A, B*, and *V* has also been proposed.[13] This model was developed from a set of 81 atom and functional group fragments and intramolecular interactions for which an evaluation of their contribution to each parameter was carried out through a process of multiple linear regressions. The method gives good results for predicting parameters, but as with all group

* Corresponding author fax: +34934907725; e-mail: joaquim.sales@qi.ub.es.

10.1021/ci049943w CCC: $27.50 © 2004 American Chemical Society
Published on Web 04/17/2004

contribution methods it retains the basic disadvantage of being unable to resolve molecular details such as isomeric tautomeric and conformational effects and it is hard to apply on large complicated compounds with diverse functional groups.

Very recently, the five COSMOments of Klamt's COSMO-RF model,[14] sig2, sig3, Hbdon3, Hbacc3, and CSA, have been correlated with the five Abraham solute descriptors for a set of 470 compounds.[15] The descriptors sig2 and sig3 represent the polarity/polarizability of the solutes; sig2 descriptor, is an excellent measure of the overall electrostatic polarity of the solute, while sig3 is a measure of the asymmetry of the sigma profile, σ, being sigma a good local descriptor for molecular surface polarity. On the other hand, Hbacc3 and Hbdon3 (the hydrogen bond moments) are quantitative measures of the acceptor and donor capacities of the compounds, respectively, and the CSA descriptor can be regarded as the molecular surface. Although the correlation between E and the Klamt descriptors is very poor, the other four Abraham parameters are well correlated with the COSMOments. It has been concluded that the two sets of descriptors exhibit a large overlap as far as chemical information content. This information however is distributed differently in each set with the Abraham set incorporating extra information in the excess molar refraction parameter E.

Quantitative Structure−Property Relationship approach (QSPR) has become very useful in the prediction and interpretation of several physical and chemical properties. The basis of this methodology is the assumption that the behavior of compounds, as expressed by any measured physical or chemical property, can be correlated with molecular features of the compounds termed descriptors. Descriptors are numerical values used to describe different characteristics about a certain structure in order to yield information about the property being studied. While common approaches often need some intuitive vision to derive the relevant mathematical relationship, QSPR methods are based on statistically determined linear or nonlinear functional forms that relate the property of interest with descriptors. Recently, Katritzky et al.[16] have reviewed the applications of the QSPR approach to technologically relevant physical properties. Other properties have been also studied by this approach, such as the following: reaction rates,[17] solubilities,[18] NMR chemical shift,[19,20] chromatographic retention parameters,[21,22] and bond dissociation energy.[23] Regarding the solute/solvent interactions, the QSPR methodology has been used in the estimation of octanol/water partition, log P,[24] and in the treatment of solvent scales.[25]

The Abraham solute parameters are appropriate magnitudes to be analyzed by the QSPR approach since they encode information intrinsically related to chemical nature of the compounds, and, on the other hand, the values of these parameters are known for many compounds of very different chemical composition.

In this paper we study linear (multiple linear regression) and nonlinear (computational neural networks) QSPR models for the four Abraham solute descriptors, E, S, A, and B, of a large set of solutes. The obtained results are compared with the estimations done by other approaches (see before), mainly with the recent study using the COSMO-RS model. The descriptors contained in the proposed models are also analyzed, relating their physicochemical features with the accepted meaning for the Abraham parameters.

Table 1. Statistical Figures of the Abraham's Parameters

parameter	n	max	min	mean	SD
E	457	3.26	−0.55	0.54	0.52
S	457	2.25	−0.25	0.69	0.45
A	457	1.62	0.00	0.15	0.27
B	457	1.50	0.00	0.36	0.28

DATA AND COMPUTATIONAL METHODS

Data Set. We have used 457 solutes, of the set of 470 compounds whose Abraham parameters are available.[15] To have molecules that allow the calculation of a large number of molecular descriptors, several substances such as iodine, nitrogen, CO, SO_2, water, ammonia, etc. have been eliminated. The set contains compounds of very different chemical nature, aliphatics, aromatics, and heterocyclics, with different substituent groups such as the following: alcohol, ether, halogen, ester, aldehyde, ketone, amine, amide, nitro, nitrile, etc. Table 1 contains statistical information about the values of the Abraham parameters studied.

Structural Descriptors. The generation of the descriptors was performed with the CODESSA program.[26] The structures of the compounds were drawn with HyperChem Lite (Hypercube, Inc), and the geometries were fully optimized, without symmetry restrictions, using the semiempirical method AM1[27] implemented in the MOPAC 6.0 program.[28] In all cases frequency calculations have been performed in order to ensure that all the calculated geometries correspond to true minima. The MOPAC output files were used by the CODESSA program to calculate about 600 descriptors, which can be classified in five large classes: constitutional (number of various types of atoms and bonds, number of rings, molecular weight, etc.); topological (Wiener index, Randic indices, Kier-Hall shape indices, etc.); geometrical (moments of inertia, molecular volume, molecular surface area, etc.); electrostatic (minimum and maximum partial charges, polarity parameter, etc.); and quantum (reactivity indices, dipole moment, HOMO and LUMO energies, etc.). In the calculation of the electrostatic descriptors the program uses partial charges derived from the empirical approach proposed by Zefirov,[29] based on the Sanderson electronegativity. Many of these electrostatic descriptors are also calculated using the charges derived from the quantum-chemical methods. Between the electrostatic descriptors there is a specific class of descriptors, called Charge Partial Surface Area descriptors (CPSA), proposed by Jurs et al.[30] These descriptors are based on the surface area of the whole molecule and on the charge distribution in the molecule, so they combine shape and electronic information to characterize the molecule, and therefore they encode features responsible for polar interactions between molecules. This set of CPSA descriptors was further developed to account for any particular type of polar interaction such as hydrogen bonding interactions. Hydrogen-Bond Charged Partial Surface Area descriptors (HB-CPSA) were proposed in analogy with CPSA descriptors.[31] Hydrogen-bond donor groups are considered to be any heteroatoms (i.e. O, S, or N) possessing protons that can be donated. Acceptor groups include any functional group possessing sufficient electron density to participate in a hydrogen bond. To simplify the calculations the halogens, some double bonds, and some aromatic bonds were not included in the derivation of these descriptors. These types of descriptors are very appropriate for the present study.

Linear and nonlinear QSPR models were built using multiple linear regression analysis (MLRA) and computational Neural Networks (CNN). CNNs have the ability to generate nonlinear methods with the descriptors to produce predicted values comparable to experimental ones.

Multiple Linear Regression Models. The data set was split randomly into a 367 member training set (tset) and an external prediction set (pset) of 90 compounds. To find the best correlation models, the heuristic multilinear regression procedures available in the framework of the CODESSA program were used to find the best correlation models. These procedures provide collinearity control (i.e., any two descriptors intercorrelated above 0.8 are never involved in the same model) and implement heuristic algorithms for the rapid selection of the best correlation, without testing all possible combinations of the available descriptors. After the heuristic reduction the pool of descriptors was reduced to nearly 200.

The goodness of the correlation is tested by the coefficient regression (R^2), the *F*-test, and the standard deviation (SD). The stability of the correlations was tested against the cross-validated coefficient, R^2_{cv}, which describes the stability of a regression model obtained by focusing on the sensitivity of the model to the elimination of any single data point. The *t*-test and the level of significance of each coefficient as well as the standardized regression coefficients (beta) are also reported. To further validate the model other tests were performed for the descriptors, the pairwise correlations, and the variance inflation factors (VIF). The VIF values, defined as $(1-R^2)^{-1}$, were calculated to identify whether excessively high multicollinear coefficients existed among the descriptors; a VIF greater than 10 is indicative of multicollinearity. The statistics of the models have been done with the SPSS program.

The model which passed the statistical diagnosis with the smallest number of descriptors was chosen. When adding another descriptor did not improve significantly the statistics of a model, it was determined that the optimum subset size had been achieved. The optimum model size in this study was five descriptors. Validation of the model was performed on the external prediction set of compounds withheld from working set.

Computational Neural Network Models. The computations were performed with the ADAPT (Automated Data Analysis and Pattern recognition Toolkit) program,[32,33] including feature selection routines (genetic algorithm[34] and simulated annealing[35]) and CNN procedures.[36]

The use of CNNs requires a cross-validation set (cvset) to determine when to stop the training of the neural network, to prevent their overtraining, and to be sure that the network would have good and general predictive ability. From the initial training set of 367 compounds, a subset of 37 solutes has been selected to form the cvset. Thus, with the CNNs we have worked with a training set of 330 compounds, a cross-validation set of 37 compounds, and a test set of 90 compounds.

Two types of CNNs studies have been done. The first, Method I, is a linear/nonlinear hybrid model because it is based on the set of descriptors chosen by the multiple linear regression, but a nonlinear CNN model is developed from these descriptors. In the second method, Method II, the reduced descriptor pool is analyzed by the genetic algorithm.

Method I: Nonlinear Models using Best MLRA Descriptors. Descriptors from the best MLRA model were passed to a CNN. The CNNs used for this analysis are three-layer, fully connected, feed-forward networks, and they have been described in detail by Jurs et al.[36,37] The number of neurons of the input layer corresponds to the number of descriptors in the model. The number of hidden layers controls the flexibility of the network and was adjusted until the optimal network architecture was achieved. The output layer contains one neuron representing the predicted Abraham parameter in each case. A quasi-Newton method BFGS (Broyden-Flectcher-Golfarb-Shanno)[37] was used to train the network.

We have found that a 5−6−1 architecture give the best results. It should be noted that the ratio of training set observations to adjustable parameters should be kept above 2.0 to avoid overtraining.[38] The number of adjustable parameters (AP) is computed as $\mathrm{AP} = (\mathrm{IL}+1) \times \mathrm{HL} + (\mathrm{HL}+1) \times \mathrm{OL}$, where IL, HL, and OL denote the number of neurons in the input layer, hidden layer, and output layer, respectively.

Method II: Nonlinear Feature Selection and Nonlinear Modeling. The set of around 200 descriptors selected by CODESSA were imported to the ADAPT program and were subjected to the objective feature selection routines of this program, and a reduced pool of nearly 100 descriptors was obtained and used in the nonlinear feature selection. Both simulated annealing and genetic algorithm routines were used along with a CNN fitness evaluator to determine the optimal set of descriptors from the reduced pool. Once the best subsets of descriptors were found, they were trained by the same procedures outlined above. After testing several models, the best one was evaluated by the external prediction set compounds to show its ability to generalize. A 5−6−1 architecture was used for the CNN.

RESULTS AND DISCUSSION

Multiple Linear Regression Models. The QSPR analysis of the *E*, *S*, *A*, and *B* values for the 367 compounds of the training set resulted in the models containing five descriptors which are given in Tables 2−5, respectively. The obtained correlations are good with R^2 between 0.758 and 0.937 and from 0.817 to 0.973, for the training and the prediction sets, respectively. Standard deviations between 0.09 and 0.16 are obtained for the four parameters for both the training and the prediction sets. The values of the *t*-test and the signification levels show, clearly, that there are good correlations between the descriptors and the experimental values of the Abraham parameters. On the other hand, the values of the corresponding VIF are low enough to leave aside correlations between the different descriptors involved in the proposed models. Table S1 (Supporting Information) gives the experimental and calculated values for the four Abraham parameters. Figure 1 shows the plot of calculated versus experimental values of the four Abraham parameters.

Parameter *E*. The parameter *E* is an excess molar refraction that is calculated from the refractive index and the McGowan molecular volume of the solutes.[39] The solute

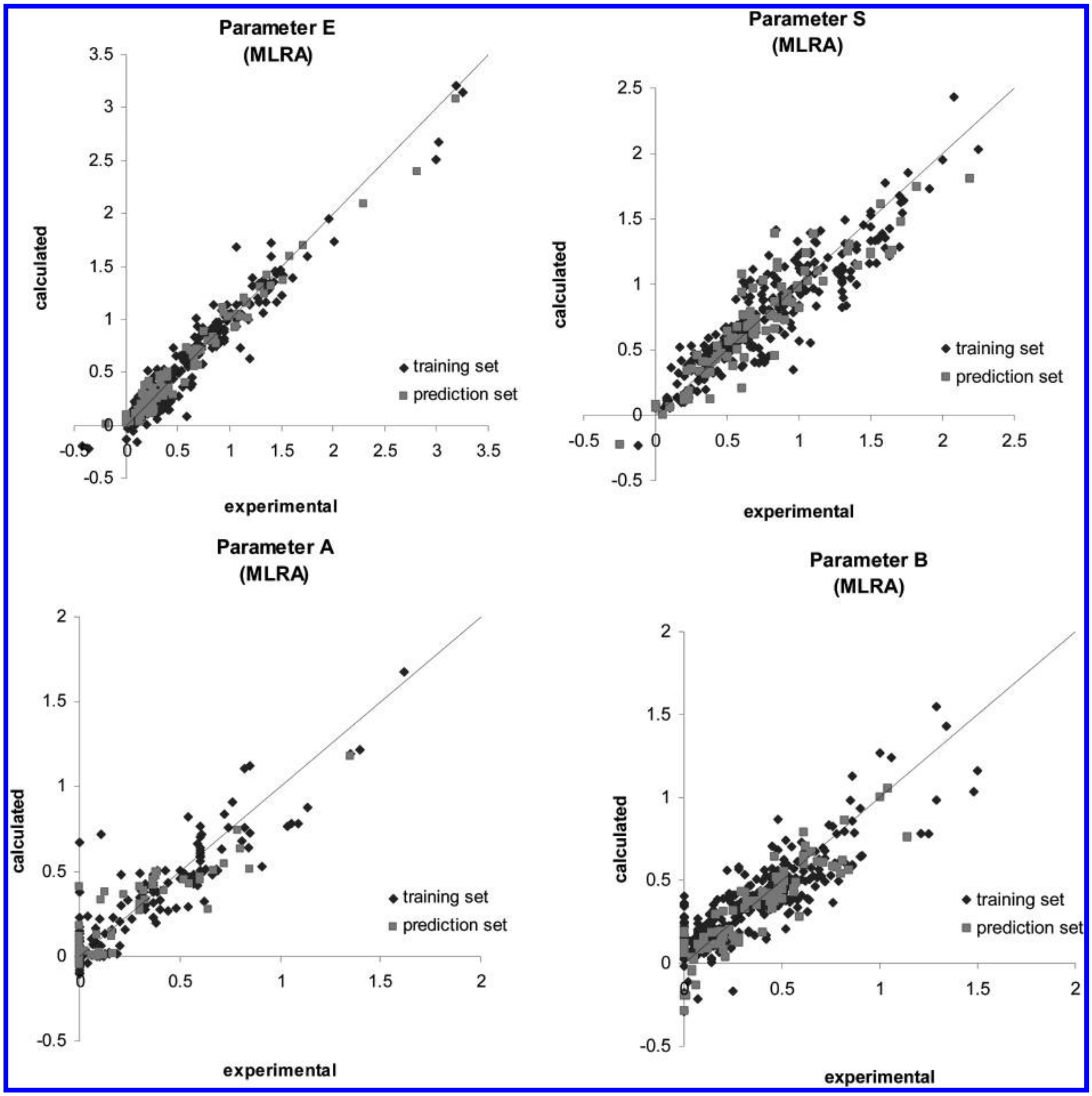

Figure 1. Plot of calculated vs experimental *E, S, A*, and *B* parameters, using MLRA of the tset and pset.

Table 2. Descriptors of MLRA Model for the *E* Parameter[a]

descriptor	coefficient	SD	beta	*t*-test	sig	VIF
intercept	0.065	0.025		2.62	0.009	
ALFA polarizability	0.023	4.00E-04	1.39	61.25	0.000	2.97
number of H atoms	−0.065	0.002	−0.69	−37.60	0.000	1.95
total molecular 1-center E−E repulsion	−0.001	3.00E-05	−0.44	−20.81	0.000	2.58
HDCA-2/TMSA (Zefirov)	140.22	9.005	0.23	15.57	0.000	1.24
relative molecular weight	0.017	0.001	0.21	12.46	0.000	1.61

[a] $R^2 = 0.937$, $F = 1081$, SD = 0.13, $n = 367$, $R^2_{cv} = 0.934$.

molar refraction, MR, is defined as

$$\mathrm{MR_X} = 10\frac{\eta^2 - 1}{\eta^2 + 2}V_\mathrm{X}$$

where η is the refractive index and V_X is the McGowan's volume, which is calculated additively from the number and kinds of atoms constituting the molecule and the number of bonds between these atoms. The *E* parameter is defined as the difference of $\mathrm{MR_X}$ of the solute and $\mathrm{MR_X}$ of an alkane with the same V_X.

The model (Table 2) contains two constitutional (number of hydrogen atoms and the relative molecular weight) and two quantum-chemical (ALFA polarizability and the total molecular 1-center E−E repulsion) descriptors. The values of these quantum-chemical descriptors are those calculated directly by the MOPAC program. The fifth descriptor, HDCA-2/TMSA, belongs to the HB-CPSA type. It is the quotient between the area-weighted surface charge of the hydrogen-bonding donor atoms in the molecule and the total molecular surface area (TMSA).

$$\text{HDCA-2} = \sum_{D} \frac{q_D\sqrt{S_D}}{\sqrt{S_{\text{TOT}}}}$$

In this case the charges are calculated by the Zefirov approach. As shown by the beta values, the most significant descriptor in the model is the ALFA polarizability, this descriptor together with the number of hydrogen atoms give a correlation with $R^2 = 0.846$ and SD = 0.19. The negative sign of the number of hydrogen atoms agrees with the low polarizability of this element. The relative molecular weight, that reflects the molecular volume, which is also related to the polarizability, is also present in the model, although with low significance. Also with low significance, the descriptor related to the hydrogen bond donor capacity, HDCA-2/TMSA, is present in the model, suggesting that some hydrogen bond donor capacity of the solutes are also encoded in this Abraham parameter. The prediction capacity of the model is very good, $R^2 = 0.972$, $F = 3113$, SD = 0.09.

In the work of Klamt,[15] the *E* parameter does not correlate well with the five COSMOments ($R^2 \approx 0.50$). However, the most significantly related moments are the hydrogen bond donor (Hbdon3) as well as sig2 and sig3. This fact suggests that the *E* parameter could encompass some kind of hydrogen bond interactions.

Parameter *S*. The model (Table 3) contains two constitutional (number of aromatic bonds, number of fluorine atoms) and two quantum-chemical (average valency of a C atom, total dipole moment) descriptors, being the total dipole moment of the molecule the calculated values by the MOPAC program. The fifth descriptor, weighted charged partial negative charged surface area, WNSA-3, is of the CPSA type

$$\text{WNSA-3} = \frac{\text{PNSA-3} \times \text{TMSA}}{1000}$$

where PNSA-3 is the atomic charge weighted partial negative surface area

$$\text{PNSA-3} = \sum_{A} q_A S_A$$

and TMSA is the total molecular surface area. In this case, the charges are derived from the Zefirov approach.

This *S* parameter is a combined dipolarity/polarizability descriptor of the solutes and measures the tendency to interact with dipolar and polarizable solvents. Besides the total dipole of the molecule, the proposed model contains descriptors that are related to the polarizability of the compounds such as the number of aromatic bonds and the number of F atoms. In general, the aromatic bonds make the compounds more polarizable, and the fluorine atoms, due to its high electronegativity, decrease the capacity of the molecule to be polarized. Accordingly, the coefficients of the number of aromatic bonds and of the number of F atoms are positive and negative, respectively. It is important to notice that despite the low number of fluoro derivatives contained in the set of solutes studied, 16 compounds, this descriptor is present in the proposed model. As indicated before, the WNSA-3 descriptor belongs to the CPSA descriptors that encode information about polar interactions. The number of aromatic bonds and the dipole moment are the most important descriptors in the model.

Sevcik et al.[8] have analyzed the *S* parameter of a set containing 333 compounds by multiple linear regression models using molecular descriptors belonging to constitutional, electrostatic, quantum, and topological types. The calculations performed resulted in $R^2 = 0.940$ for the training set of 266 compounds, and $R^2 = 0.585$ for the prediction set of 67 compounds, for a model with 29 descriptors. When a model of 7 descriptors was used the results were poorer, with a R^2 of 0.741 and 0.450 for the training and prediction sets, respectively. In the work of Platts et al.[9] a set of 98 compounds were studied. The descriptors that best correlated with *S* are as follows: the molecular dipole moment, the polarizability, the CHelpG atomic charges, and the frontier orbital energies; for a training set of 58 compounds they obtain good correlations: $R^2 \approx 0.83$, SD ≈ 0.19, $F \approx 55$. The COSMOments of Klamt,[15] except for the area descriptor (CSA), give also good correlations with *S*. Thus, values of $R^2 = 0.777$, SD = 0.215, and $F = 404$ are obtained with the other four descriptors, being, clearly, the sig2 and the Hbdon3 the most strongly related. In contrast to our results, this approach suggests that the *S* parameter can contain some kind of hydrogen bond donor capacity.

The following statistics are obtained with the 90 compounds of the pset, $R^2 = 0.856$, $F = 522.8$, and SD = 0.16. Our results are better than those obtained by other calculations.

Parameter *A*. The model shown in Table 4 contains the electrostatic descriptor polarity parameter/square distance, pol/d^2, which is calculated as the difference between the maximum (q_{max}) and the minimum (q_{min}) charges factorized by the square of the distance between the atoms that bear minimum and maximum partial charges. The remaining four descriptors of the model are related to the hydrogen bonding interactions. One is the number of hydrogen-donor sites, and the other is HDCA-2, defined earlier. The other two are related to the acceptor capacity of the solutes: HACA-1/TMSA, which is the acceptor descriptor parallel to HDCA, and the Fractional Hydrogen Acceptor Surface Area, FHASA, which is the quotient HASA/TMSA; in both descriptors the partial charges were calculated from quantum calculations. The beta coefficients of these two descriptors are relatively low, but they have opposite signs, being that HACA-1/TMSA is negative.

It is not easy to relate the meaning and the sign of these descriptors with the significance of the *A* parameter. Table 4 shows that the hydrogen bond donor descriptor, HDCA-2, is clearly the most important; its beta coefficient is nearly three times the coefficient of the second most important descriptor, FHASA. The correlation with only HDCA-2 has one $R^2 = 0.784$, and the addition of the other four descriptors increases R^2 to 0.873. The results obtained with the pset are similar to those of the tset, $R^2 = 0.846$, $F = 485.5$, and SD = 0.09.

These results are not very different from those found by the approach of Klamt.[15] They obtained very good correlations of this parameter *A*: $R^2 = 0.926$, SD = 0.075, $F = 1941$, with the COSMOments, sig3, Hbdon3, and with Hbacc3, which represents the hydrogen-bond acceptor ability. The three descriptors have the same importance in the correlation, since their beta coefficients are nearly equal

Table 3. Descriptors of MLRA Model for the *S* Parameter[a]

descriptor	coefficient	SD	beta	*t*-test	sig	VIF
intercept	12.774	1.446		8.83	0.000	
number of aromatic bonds	0.068	0.002	0.61	30.37	0.000	1.10
total dipole of the molecule	0.134	0.007	0.43	18.45	0.000	1.45
WNSA-3 (Zefirov)	−0.083	0.007	−0.26	−11.78	0.000	1.33
number of F atoms	−0.229	0.023	−0.21	−9.88	0.000	1.24
average valency of a C atom	−3.225	0.367	−0.22	−8.78	0.000	1.64

[a] $R^2 = 0.868$, $F = 474$, SD = 0.16, $n = 367$, $R^2_{cv} = 0.861$.

Table 4. Descriptors of MLRA Model for the *A* Parameter[a]

descriptor	coefficient	SD	beta	*t*-test	sig	VIF
intercept	0.002	0.007		0.25	0.799	
HDCA-2 (Zefirov)	0.901	0.043	0.79	20.78	0.000	4.11
FHASA (quantum)	0.768	0.089	0.29	8.55	0.000	3.21
HACA-1/TMSA (quantum)	−2.161	0.346	−0.18	−6.25	0.000	2.46
count of H-donor sites	−0.008	0.001	−0.15	−6.38	0.000	1.67
polarity/square distance	0.422	0.068	0.15	6.23	0.000	1.55

[a] $R^2 = 0.873$, $F = 498$, SD = 0.10, $n = 367$, $R^2_{cv} = 0.863$.

Table 5. Descriptors of MLRA Model for the *B* Parameter[a]

descriptor	coefficient	SD	beta	*t*-test	sig	VIF
intercept	−0.292	0.025		−11.56	0.000	
PPSA-3 (Zefirov)	0.069	5.10E-03	0.46	13.39	0.000	1.71
relative number of N atoms	0.888	0.212	0.13	4.16	0.000	1.50
min net atomic charge	−0.851	0.066	−0.37	−12.83	0.000	1.21
count of H-acceptor sites	0.169	0.014	0.51	12.12	0.000	2.58
HBSA (quantum)	−2.39E-03	2.28E-04	−0.38	−10.41	0.000	1.99

[a] $R^2 = 0.758$, $F = 225.6$, SD = 0.14, $n = 367$, $R^2_{cv} = 0.743$.

(0.56, 0.56, and 0.52, respectively), suggesting that the *A* parameter contains information about both hydrogen bonding donor and acceptor characters. Working with a small set of 39 compounds, the ab initio and DFT calculations of Platts[10] find a good correlation: $R^2 = 0.899$, SD = 0.075, $F = 330$, with only one descriptor, the electrostatic potential at the donor H nuclear position, EP_{NUC}. This descriptor can be interpreted as an approximation to the binding energy of the hydrogen nucleus with the molecule.

Parameter *B*. The model of Table 5 is formed by one constitutional (relative number of nitrogen atoms), three electrostatics (PPSA-3, count of H-acceptors sites, HBSA), and one chemical-quantum (minimum net atomic charge) descriptors. Some of them are related to the hydrogen bonding capacity of the solutes, the number of hydrogen acceptor sites clearly reflects the basicity, and HBSA encode information about the donor and the acceptor hydrogen bonding, it is calculated as the addition of the respective hydrogen-bonding donor (HDSA), and hydrogen-bonding acceptor (HASA) descriptors, being the partial charges calculated by MOPAC. The relative number of nitrogen atoms is related to the basicity of the compounds. The beta coefficients show that the number of H-acceptor sites and PPSA-3 are the most important ones. The Partial Positive Surface Area, PPSA-3, is calculated by

$$\text{PPSA-3} = \sum_A q_A S_A$$

where q_A is the atomic partial charge, and S_A is the atomic solvent-accessible surface area. The charge has been calculated by the Zefirov's approximation. The prediction set gives the following results: $R^2 = 0.816$, $F = 391.6$, and SD = 0.11.

Klamt[15] has obtained similar correlations ($R^2 = 0.879$, SD = 0.10) with their COSMOments. Sig2 and sig3 are the most important descriptors, Hbacc3 and Hbdon3 have a lower contribution to the regression, but their relative relevance is nearly the same. These results show again that the *B* parameter contains information about the donor and acceptor hydrogen bond capacities of the solutes. On the other hand, the approach of Platts[11] does not give good results with the parameter *B*. None of the isolated base properties, such as atomic charges, multipoles, local properties of the electron density, electrostatic potential and its derivatives, correlate well with *B* ($R^2 \approx 0.11-0.52$, $n = 50$). However, some family dependent models can be constructed for bases with O, N, S, or C as acceptor atoms, separately; thus, with a set of 17 nitrogen bases, better correlations with $R^2 = 0.78$ are obtained.

Computational Neural Networks Models: Method I. The five descriptors of the models derived by MLRA for each Abraham's parameter were imported to the ADAPT program and to CNNs routines. Architecture 5−6−1 gave good results. This neural network contains 43 adjustable parameters, corresponding to a ratio of 7.7 for training set observations (330) to adjustable parameters, well above the minimum acceptable rate of 2.

Table 6 show the regression coefficients, R^2, and the standard deviations, SD, of the regressions obtained for the tset, cvset, and pset. The obtained results are better than those derived by MLRA for all the parameters, both for the tset and the pset. The most important improvements are obtained

Table 6. Statistical Results of Method I

parameter	R^2 (tset)	SD (tset)	R^2 (pset)	SD (pset)	R^2 (cvset)	SD (cvset)
E	0.966	0.09	0.977	0.09	0.919	0.10
S	0.903	0.14	0.874	0.16	0.845	0.12
A	0.954	0.06	0.936	0.07	0.915	0.06
B	0.860	0.10	0.847	0.11	0.785	0.12

Table 7. Statistical Results of Method II

parameter[e]	R^2 (tset)	SD (tset)	R^2 (pset)	SD (pset)	R^2 (cvset)	SD (cvset)
E[a]	0.977	0.08	0.981	0.08	0.927	0.09
S[b]	0.916	0.13	0.878	0.16	0.887	0.10
A[c]	0.953	0.06	0.871	0.09	0.906	0.06
B[d]	0.929	0.08	0.861	0.11	0.942	0.06

[a] Relative molecular weight; Randic index (order 3); HDCA-2 (Zefirov); total molecular 1-center E−E repulsion; ALFA polarizability. [b] ALFA polarizability; HDCA2/TMS (Zefirov); number of occupied electronic levels; total dipole of molecule; total molecular electronic interaction. [c] Relative positive charged surface area (Zefirov); total dipole of molecule; HASA (quantum); HDSA-2 (quantum); minimum atomic orbital electronic population. [d] Number of oxygen atoms; number of nitrogen atoms; maximum net atomic charge for a C atom; DPSA-2 (quantum); number of occupied electronic levels/number of atoms. [e] Descriptors involved.

for the *A* and *B* parameters. The SD of the tset of *A* and *B* parameters improve from 0.09 to 0.06 and from 0.12 to 0.10, respectively. Table S1 (Supporting Information) gives the calculated values for the four Abraham parameters.

Computational Neural Networks Models: Method II. A genetic algorithm routine using CNN fitness evaluator was applied to a 5−6−1 architecture for the data set. An extensive search of the reduced descriptor pool yielded several five-descriptor models. These fully nonlinear models were then trained and tested as above. Table 7 gives the models and statistics of the nonlinear correlations for the training, cross-validation, and prediction sets of the Abraham parameters. Good correlations have been found for all the parameters and sets studied. For the training set the R^2 and SD values have been improved for the four parameters, mainly for the *A* and *B* parameters. For the prediction set, better results have also been obtained. These CNN models have many descriptors which are of the same type that those derived by MLRA, as will be shown in detail below. The values obtained for the four parameters are shown in Table S1 (Supporting Information) and plotted in Figure 2.

Parameter *E*. Three of these five descriptors were also present in the MLRA model: relative molecular weight, ALFA polarizability—both clearly related to the molar refraction which is the main information contained in the *E* parameter—and total molecular 1-center E−E repulsion. The fourth descriptor present, HDCA-2/TMSA (HB-CPSA), is very similar to the HDCA-2, and the fifth one, is the topological Randic index (order 3).

Parameter *S*. In this case, there is one descriptor, total dipole of molecule, which was also present in the MLRA model. This descriptor and the ALFA polarizability are evidently related to the dipolarity/polarizability meaning of the *S* parameter. The presence of HDCA-2/TMSA in the model suggests that H-donor character of the solutes can be encoded in the physical meaning of the *S* parameter. This possibility was already found in the COSMO model as it has already been mentioned.

Parameter *A*. The two HB-CPSA descriptors of the model, HASA and HDS-2, describe acceptor and donor H-bonding interactions, respectively. The fact that the *A* parameter can imply opposite hydrogen bonding interactions was also found in the MLRA and COSMO methods.

Parameter *B*. The genetic algorithm model contains two constitutional descriptors, number of oxygen and nitrogen atoms, that undoubtedly are related to the hydrogen basicity of the solutes; the MLRA model contains the relative number of nitrogen atoms. There is one CPSA descriptor, DPSA-2, defined as the difference between descriptors for the positive and negative partial charged atoms in the molecule, which is analogous to the PPSA-3 found in the MLRA approach. The other two descriptors are the maximum net atomic charge for a C atom and the number of occupied electronic levels divided by the number of atoms in the solute.

CONCLUSIONS

Multiple linear regression and computational neural networks have been used in order to develop useful models for the estimation of *E, S, A*, and *B* parameters of solutes proposed by Abraham. Although both methods give good results, the best ones have been obtained with the CNN approach, in particular with *A* and *B* parameters, which give poorer multiple linear regressions. The results presented in this work are comparable or better than those derived from other theoretical calculations.

The MLRA models contain descriptors that agree with the accepted physicochemical meaning of the Abraham parameters. Thus, the molar excess, *E*, is described mainly by the ALFA polarizability descriptor, being also important the relative molecular weight and the number of hydrogen atoms. The dipolarity/polarizability, *S*, is explained mainly by three descriptors: total dipole moment, number of aromatic bonds and number of fluorine atoms, the last descriptor appears with a negative sign, in agreement with the high electronegativity and consequent low polarizability of this element. The hydrogen bond acidity, *A*, is strongly related to the hydrogen donor descriptor HDCA-2, while the hydrogen bond basicity, *B*, mainly depends on the number of H-acceptor sites and the relative number of nitrogen atoms. The models derived by the CNN approach contain descriptors that are either the same or belonging to the same type, thus encoding similar information.

On the other hand, some of the descriptors obtained from both MLRA and CNN approaches encode physicochemical information that is not explicitly described in the Abraham

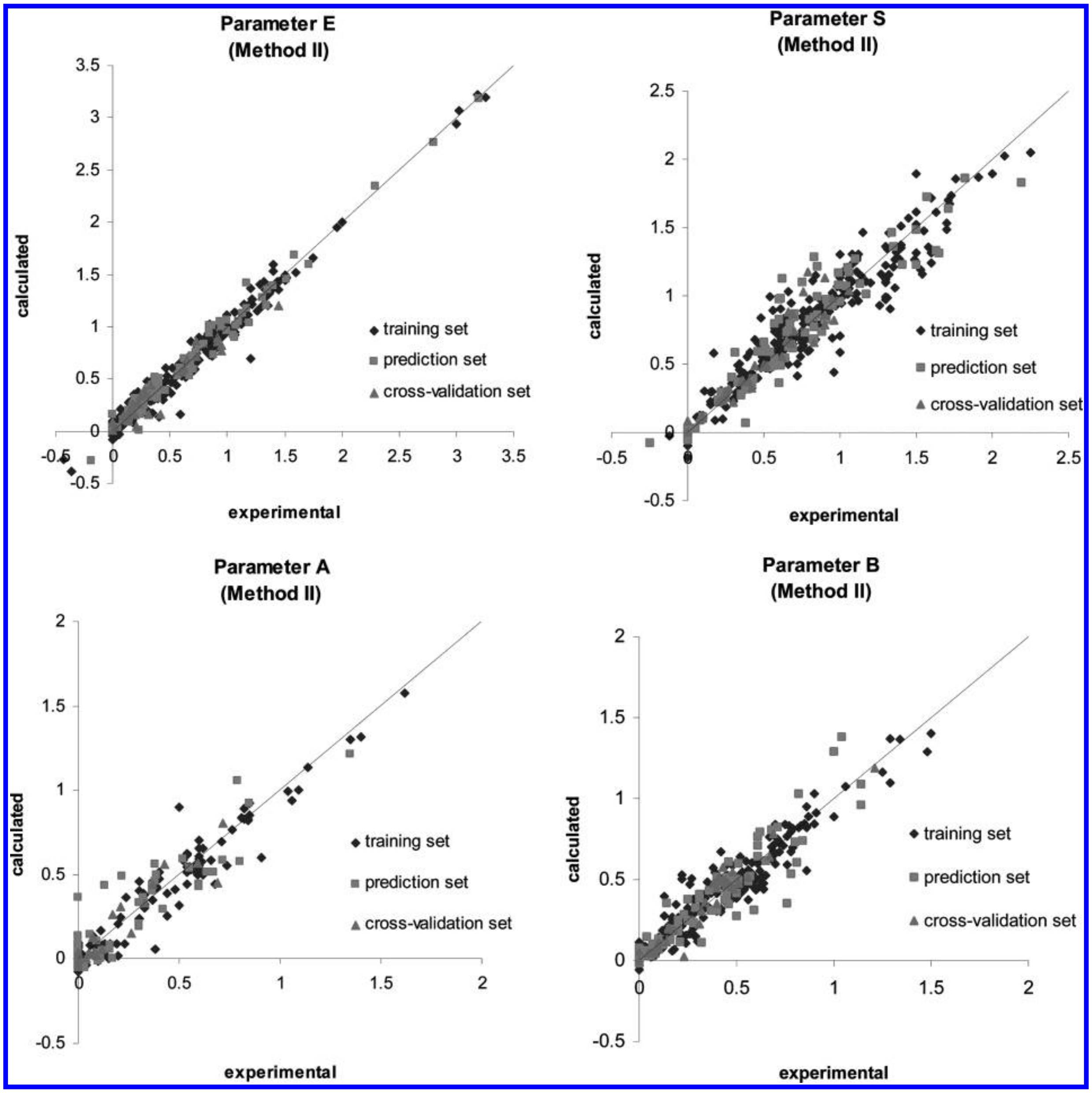

Figure 2. Plot of calculated vs experimental *E, S, A*, and *B* parameters, using Method II of the tset, pset, and cvset.

parameters. Thus, the model for the *E* parameter contains the HDCA-2/TMA descriptor, which describes donor hydrogen interactions; hydrogen-bonding interactions also contribute to the parameter *S*, while the models for both *A* and *B* parameters enclose descriptors that represent donor as well as acceptor capabilities. Results of this kind have also been found in the COSMO-RF model. These facts can be explained taking into account the huge complexity of solute/solvent interactions and also the difficulty of isolating closely related physicochemical properties by performing experimental measurements. Thus, although the Abraham parameters reflect mainly the initially proposed meaning, the participation of other characteristics and/or capacities of the solutes cannot be ruled out.

ACKNOWLEDGMENT

The authors thank Prof. M. H. Abraham (University College London) for the values of the parameters of the solutes studied, Prof. Peter C. Jurs (Pennsylvania State University) for giving us access to the ADAPT program, and Dr. Phil Moiser for helping us in the use of the program. Financial support from the Catalan Government (Grant 2001 SGR 00052) is gratefully acknowledged.

Supporting Information Available: Table S1 with a listing for the molecule name, the CAS registry number, the experimental and calculated values by MLRA, and Method I and Method II for the *E*, *S*, *A*, and *B* Abraham parameters of the 457 solutes studied. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Abraham, M. H. Scales of Hydrogen-bonding: Their Construction and Application to Physicochemical and Biochemical Processes. *Chem. Rev.* **1993**, *22*, 73−83.

(2) Abraham, M. H.; Chadha, H. S.; Whinting, G. S.; Mitchell, R. C. Hydrogen bonding. 32. An analysis of water-octanol and water-alkane partitioning and the delta log P parameter of Seiler. *J. Pharm. Sci.* **1994**, *83*, 1085−1100.

(3) Abraham, M. H.; Zissimos, A. M.; Acree, W. E. Partition of solutes from the gas phase and from water to wet and dry di-*n*-butyl ether: A linear free energy relationship analysis. *Phys. Chem. Chem. Phys.* **2001**, *3*, 3732−3736.

(4) Abraham, M. H.; Le, J. The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. *J. Pharm. Sci.* **1999**, *88*, 868−880.

(5) Fuguet, E.; Ràfols, C.; Bosch, E.; Abraham, M. H.; Rosés, M. Solute−solvent interactions in micellar electrokinetic chromatography. Selectivity of lithium dodecyl sulphate-lithium perfluorooctanesulfonate mixed-micellar buffers. *J. Chromatogr. A* **2002**, *942*, 237.

(6) Platts, J. A.; Abraham, M. H.; Zhao, Y. H.; Hersey, A.; Ijaz, L.; Butina, D. Correlation and prediction of a large blood-brain distribution data set-an LFER study. *Eur. J. Med. Chem.* **2001**, *36*, 719−730.

(7) Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J.; Luscombe, C. N.; Boutina, D.; Beck, G.; Sherbone, B.; Cooper, I.; Platts, J. A. Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure−activity relationship (QSAR) with the Abraham descriptors. *J. Pharm. Sci.* **2001**, *90*, 749−784.

(8) Svozil, D.; Sevecik, J. G. K.; Kvasnicka, V. Neural Network Predcition of the Solvatochromic Polarity/Polarizability Parameter $\pi^{H}{}_{2}$. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 338−342.

(9) Lamarche, O.; Platts, J. A.; Hersey, A. Theoretical prediction of the polarity/polarizability parameter $\pi^{H}{}_{2}$. *Phys. Chem. Chem. Phys.* **2001**, *3*, 2747−2753.

(10) Platts, J. A. Theoretical prediction of hydrogen bond donor capacity. *Phys. Chem. Chem. Phys.* **2000**, *2*, 973−980.

(11) Platts, J. A. Theoretical prediction of hydrogen bond basicity. *Phys. Chem. Chem. Phys.* **2000**, *2*, 3115−3120.

(12) Lamarche, O.; Platts, J. A. Complementary nature of hydrogen bond basicity and acidity scales from electrostatic and atoms in molecules properties. *Phys. Chem. Chem. Phys.* **2003**, *5*, 677−684.

(13) Platts, J. A.; Butina, D.; Ahraham, M. H.; Hersey, A. Estimation of Molecular Linear Free Energy Relation Descriptors Using a Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 835−845.

(14) Klamt, A. Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *J. Phys. Chem.* **1995**, *99*, 2224−2235.

(15) Zissimos, A. M.; Abraham, M. H.; Klamt, A.; Eckert, F.; Wood, J. A Comparison between the Two General Sets of Linear Free Energy Descriptors of Abraham and Klamt. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1320−1331.

(16) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally Diverse Quantitative Structure−Property Relationship Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1−18.

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

(18) McElroy, N. R.; Jurs, P. C. Prediction of Aqueous Solubility of Heteroatom-Containing Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1237−1247.

(19) Clouser, D. L.; Jurs, P. C. The Simulation of ^{13}C nuclear magnetic resonance spectra of dibenzofurans using multiple regression analysis and neural networks. *Anal. Chim. Acta* **1996**, *321*, 127−135.

(20) Bosque, R.; Sales, J. A QSPR Study of the ^{31}P NMR Chemical Shifts of Phosphines. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 225−232.

(21) Baczek, T.; Kaliszan, R. J. *J. Chromatogr. A* **2003**, *987*, 29.

(22) Bosque, R.; Sales, J.; Bosch, E.; Rosés, M.; García-Alvarez-Coque, M. C.; Torres-Lapasió, J. R. A QSPR Study of the *p* Solute Polarity to estimate Retention in HPLC. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1240−1247.

(23) Bosque, R.; Sales, J. A QSPR Study of O−H Bond Dissociation Energy in Phenols. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 637−642.

(24) Grombar, V. K.; Enslein, K. Assesment of *n*-Octanol/Water Partition Coefficient: When Is The Assessment Reliable? *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1127−1134.

(25) Katritzky, A. R.; Tamm, T.; Wang, Y.; Karelson, M. QSPR Treatment of Solvent Scales. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 684−691.

(26) Katritzky, A. R.; Lovanov, V. S.; Karelson, M. *CODESSA, Reference Manual V 2.13*; Semichem and the University of Florida, 1997.

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

(28) Stewart, J. P. P. *MOPAC 6.0 Quantum Chemistry Program Exchange;* QCPE, No. 455, Indiana University, Bloomington, IN, 1989.

(29) Zefirov, N. S.; Kirpichenok, M. A.; Izmailov, F. F.; Trofimov, M. I. Calculation Schemes for atomic electronegativities in molecular graphs within the framework of Sanderson Principle. *Dokl. Akad. SSSR* **1987**, *296*, 883−887.

(30) 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.

(31) Stanton, D. T.; Egolf, L. M.; Jurs, P. C. Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 306−316.

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

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

(34) Luke, B. T. Evolutionary Programming Applied to the Development of Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279−1287.

(35) Sutter, J. M.; Dixon, S. L.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure−Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.

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

(37) 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.

(38) Livingstone, D. J.; Manallack, D. T. Statistics Using Neural Networks: Chance Effects. *J. Med. Chem.* **1993**, *36*, 1295−1297.

(39) Abraham, M. H.; Whiting, G. S.; Doherty, R.; Shuely, W. J. A New Method for the Characterisation of GLC Stationery Phases- The Laffort Data Set. *J. Chem. Soc., Perkin Trans. 2* **1990**, 1451−1460.

CI049943W