J. Phys. Chem. A **2005,** *109,* 10323−10341

The Classification of Solvents by Combining Classical QSPR Methodology with Principal Component Analysis

Alan R. Katritzky,*,† Dan C. Fara,† Minati Kuanar,† Evrim Hur,‡ and Mati Karelson*,§

Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, Florida 32611, U.S.A., Department of Chemistry, Faculty of Arts and Science, Osmangazi University, 26020 Eskisehir, Turkey, and Institute of Chemistry, Tallinn University of Technology, Ehitajate tee 5, Tallinn 19086, Estonia

Received: January 22, 2005; In Final Form: August 6, 2005

The results of a quantitative structure−property relationship (QSPR) analysis of 127 different solvent scales and 774 solvents using the CODESSA PRO program are presented. QSPR models for each scale were constructed using only theoretical descriptors. The high quality of the models is reflected by the squared multiple correlation coefficients that range from 0.726 to 0.999; only 18 models have $R^2 < 0.800$. This enables direct theoretical calculation of predicted values for any scale and/or for any organic solvent, including those previously unmeasured. The molecular descriptors involved in the models are classified and discussed according to (i) the origin of their calculation (i.e., constitutional, geometric, charge-related, etc.) and (ii) the commonly accepted classification of physical interactions between the solute and solvent molecules in liquid (condensed) media. A reduced matrix 774 (solvents) × 100 (solvent scales) was selected for the principal component analysis (PCA) by taking into account only the solvent scales with more than 20 experimental data points. The first 5 principal components account for 75% of the total variance. The robustness of the PCA model obtained was validated by the comparison models development for restricted submatrices of data and with the results obtained for the full data set. The total variance accounted for by the first three PCs, for the submatrices with the same number of solvent scales but different numbers of solvents, varies from 68.2% to 59.0%. This demonstrates that the total variance described by the first 3 components is essentially stable as the number of solvents involved varies from 100 to 774. Subsequently, a matrix with 703 diverse solvents and 100 solvent scales was selected for the general classification of the solvents and scales according to the scores and loadings obtained from the PCA treatment. Classification of the theoretical molecular descriptors, derived from the chemical structure alone, according to their relevance to specific types of intermolecular interaction (cavity formation, electrostatic polarization, dispersion, and hydrogen bonding) in liquid media enables a more easily comprehensible physical interpretation of the QSPR of molecular properties in liquids and solutions. The reported QSPR models for solvent scales with theoretical molecular descriptors and the results of the PCA analysis are potentially of great practical importance, as they extend the applicability of correlations with empirical solvent scales to many previously unmeasured systems.

1. Introduction

Solvents form the basis for much of the practice of chemistry. As every chemist knows, the selection of an appropriate solvent for a chemical reaction or physical measurement is often vital to its success. Consequently, the prediction and understanding of the influence and the nature of a solvent and its appropriate choice are of great importance. Attempts to link a single definitive solvent property, usually its “polarity”, to a single physical characteristic have met with little success. Indeed, solvent properties, especially solvent polarity, cannot be defined by a single parameter, whether physical/chemical or experimental/theoretical. Trying to understand the properties of solvents and to facilitate solvent choice has led to the development of many solvent scales. These scales are based on diverse physicochemical phenomena including reaction rates, solvatochromic effects, and reaction enthalpies, among others. The detailed mechanism of the influence of solvent on different physical or chemical processes is still under discussion. The same often applies to the individual solvent scales. Several analyses of solvent scales together with reviews and discussions on the subject have been published.[1,2]

Solvents influence chemical and physical processes by interacting with the solute through either van der Waals interactions or hydrogen-bonding dipole−dipole interactions or by providing solvent pockets or cages for encapsulating the solute. The functional groups present in the solvent molecule, their orientation, and the structure of the molecule as a whole play important roles. Solvents have been classified according to their polarity, acidity/basicity, electronic properties, and so forth. A popular classification of organic solvents, due to Parker, divides them into protic, dipolar aprotic, and apolar aprotic solvents, according to the dipolarity of the solvent molecules and their ability to act as hydrogen bond donors.[3]

---

* Corresponding authors. Phone 1 (352) 392-0554, fax 1 (352) 392-9199, e-mail katritzky@chem.ufl.edu (A.R.K.); phone + (372) 620-2814, fax + (372) 620-2828, e-mail mati.karelson@ttu.ee (M.K.).

† University of Florida.

‡ Osmangazi University.

§ Tallinn University of Technology.

10.1021/jp050395e CCC: $30.25

Most existing solvent scales have been constructed by choosing a model system and recording the changes in a measurable parameter, as the solvent is changed. Whereas the model processes have been carefully chosen to represent different intermolecular interactions in the system, no one scale can be universal and applicable for all kinds of properties. The empirical properties used to define solvent polarity scales include the following: (i) equilibrium and kinetic rate constants of chemical reactions of solutes, (ii) spectroscopic properties of solutes in different solvents comprising absorption and fluorescence energies, vibrational transition energies, and so forth, (iii) solvation energies and free energies of different solutes, (iv) macroscopic properties of solvents including dielectric constant, refractive index, molecular volume, polarizability index, and so forth, and (v) composite experimental parameters. Individual solvents are rarely represented in all common scales, and no single scale covers all the common solvents.

To date, more than a hundred quantitative solvent polarity scales have been proposed on the basis of diverse properties of solvents and solutes, including chemical reactivity, spectroscopic properties, or directly measured energies and/or free energies of solvation. It was realized early that no single solvent scale could offer a general correlation/classification for either solvent effects or solvents. Therefore, various multiparameter equations have been formulated by (i) using a combination of two or more existing scales or (ii) postulating different specific parameters to account for distinct types of effects. One of the earliest such multiparameter approaches was made more than 30 years ago by Fowler, Katritzky, and Rutherford.[4] Koppel and Palm[5] also defined a general four-parameter (Y, P, E, B) equation for characterizing specific and nonspecific solute−solvent interactions. Their work has been extended by Koppel and Paju[6] and more recently by Palm and Palm.[7,8] Kamlet and Taft[9] developed linear solvation energy relationships (LSER) using solvatochromic parameters, which successfully describe the relationship between several solvent scales. Significant contributions to the multiparameter approach of solvent properties were made by Famini et al.[10−12] with the definition of corresponding theoretically derived scales (theoretical linear solvation energy relationship, TLSER) and by Politzer and Murray[13] who introduced a general interaction properties function (GIPF) approach to represent, predict, and analyze condensed-phase macroscopic properties depending on noncovalent interactions, based on a detailed statistical characterization of the electrostatic surface potentials.

Other attempts have been made to formulate multiparameter equations, including (i) the proposal of a unified solution model that, by using QSPR, has also been correlated to theoretical molecular descriptors,[14−17] (ii) the investigation of the interrelation between different solvent parameters,[18] and (iii) the usage of different semiempirical scales as descriptors of nonspecific solvent effects to obtain a generalized solvent polarity scale.[19,20]

There have been several different proposals for the classification of solvents. Thus, Gramatica et al.[20] used k-nearest neighbor (KNN) and artificial neural network (ANN) methods and classified 152 solvents based on different sets of molecular descriptors. The 152 solvents were assigned to 5 different classes.

Relative recently, multivariate statistical analysis methods began to be used as tools for the classification and selection of organic solvents according to their solvent effects. The extraction of chemical information contained in the data set, relative to the importance of individual variables, can be achieved by multiple linear regression analysis, factor analysis (FA), principal component regression (PCR), principal component analysis (PCA), partial least squares (PLS), and other methods (NIPALS, etc.). PCA was used by several researchers in order to (i) characterize the ionicity[21] of molecules, (ii) assess the nucleophilicity and electrophilicity of radicals,[22] (iii) study the solvent effects on the CO stretching frequency of molecules,[23] (iv) characterize the solvent properties of gas chromatographic stationary phases,[24] and (v) analyze the gas−liquid partitioning of alkanes in several organic solvents.[25] These studies prove that PCA can be a useful tool in quantifying the solute and solvent effects.

Poole and Poole[26] used Abraham's solvation parameter model to characterize the solvent properties of 33 stationary phases in terms of their capacity for specific intermolecular interactions at 121.4 °C. A PCA together with clustering techniques classified the phases according to specific intermolecular interactions.[26] Lochmüller et al.[27] applied multivariate statistical techniques of principal component analysis and target transformation factor analysis to examine the reversed-phase high-performance liquid chromatography behavior of some 35 benzene derivatives in different solvent systems. Partition coefficient data for 50 solutes in 6 nonpolar solvent systems on analysis by PCA[28] showed that the relationship between solute structure and partitioning behavior for simple organic compounds depended on the isotropic surface area as the most important parameter.

A factor analysis performed by Martin et al.[29] for a data set consisting of 18 organic solvents characterized by 18 physicochemical parameters led to a classification similar to that of Parker, i.e, into protic, dipolar aprotic, and apolar aprotic solvents.[3]

Cramer found two principal components from a matrix of 6 physical properties of 114 pure liquid compounds, which are characteristics of the solvent molecule.[30] The two PCs have been used in the prediction of experimental values of 18 physical properties for 139 additional liquids of diverse structure.[31]

A general classification procedure for organic solvents proposed by Chastrette et al.[32] treats a basis set of 8 physicochemical properties with a set 83 organic solvents using principal component analysis. The 83 solvents are grouped into 9 classes by their clustered component values, using the discriminating power of the 8-descriptor properties. However, duality in the descriptor choice was found for some solvents. In another approach, a set of 83 solvents was classified into polar and nonpolar types based on the PCA analysis of 9-graph theoretical molecular descriptors.[33]

Maria et al.[34] used principal component analysis to determine the inherent dimensionality of the condensed-phase basicity of nonprotogenic organic molecules commonly used as solvents. Ten basicity scales were characterized by thermodynamic and spectroscopic measurements. The first and second factor PCA attributed to the electron delocalization and electrostatic characters, respectively, accounted for about 95% of the variance. Another nonhierarchical classification of 103 organic solvents used characteristics vector analysis of a set of 6 physical and empirical parameters.[35]

Pytela classified 51 solvents on the basis of clustering in the three-dimensional space formed by the empirical scales of solvent polarity/acidity, polarity/basicity, and polarity/polarizability parameters.[36]

In a previous study,[1] our group collected a representative set of 45 different solvent scales, which contained data for a total of 350 solvents; this data set was used to develop QSPR (quantitative structure−property relationship) models for the individual scales using theoretical molecular descriptors. The resulting QSPR equations for the different scales were compared in an attempt to elucidate solvents/solvent effects, similarities, and differences. The results gave considerable insight into both the nature of the scales and the nature of the solvents. The physical significance of the descriptors was consistent with the physical concepts used by the original authors to construct the scales. A PCA treatment was carried out with 40 solvent scales as variables, each having 40 data points for 40 solvents as objects. The first 3 principal components accounted for 74% of the total variance. The solvents and the solvent scales cluster according to the scores and loadings obtained from PCA treatment.[37] The aims of the present paper are (i) to extend the previous chemometric treatment to many more solvents and solvent effect scales in a continuing attempt to find a general model for solvent scales and (ii) to obtain a general classification procedure of solvents with regard to their solvent characteristics based on the multivariate statistical methods, in particular, principal component analysis (PCA). All experimental solvent scales have been measured for only a limited number of solvents. Moreover, the lists of solvents utilized to correlate different scales do not overlap. Therefore, to carry out a multivariate PCA analysis on all available scales, it is necesssary to evaluate the missing values in the respective (solvent) × (solvent scale) data matrix. As demonstrated earlier,[1] the development of QSPR based on theoretical molecular descriptors for each solvent scale and the prediction of missing data from these relationships can be used successfully to construct the necessary full PCA matrix. Thus, in the first stage of the present work, such QSPR models were developed for each solvent scale considered.

2. Data Set

The individual experimental solvent scales were taken from the literature.[2] The first table in the Supporting Information (SM 1) contains experimental values of the 127 solvent scales treated. The number of solvents involved in these 127 scales varies from 5 to several 100s.[2] Altogether, 774 solvents were represented in different scales. In the case of scale SPS_4, a few erroneous values reported in the reference[38] were recalculated for the development of the QSPR model. The physical background of the 127 solvent scales and literature references are given in SM 2. The 127 scales have been divided into 4 groups according to the type of physicochemical properties used to define them. The second and third columns of the table given in SM 2 list, respectively, our notation and the notation designated by the original authors of the scale. The next two columns of SM 2 give the size of the data set used by the developers of the given solvent scale and the initial number of solvents considered in our study. In some cases, the number of data points reported in SM 2 from the original sources for the solvent scales includes solvent mixtures, deuterated solvents, inorganic compounds, the salt form of the compounds, and so forth. From solvent scales where such data were included (SPS_9, SPS_{15}, SPS_{16}, SPS_{24}, SPS_{27}, SPS_{29}, SPS_{30}, SPS_{31}, SPS_{32}, SPS_{45}, SPS_{62}, SPS_{67}, SPS_{68}, SPS_{70}, SPS_{88}, SPS_{92}, SPS_{97}, SPS_{98}, SPS_{104}, and SPS_{115}), we excluded those experimental data points. In just 3 exceptional cases of SPS_7, SPS_{69}, and SPS_{92}, we used a smaller number of data points after we removed significant outliers from the QSPR regression line. Also, the preliminary QSPR modeling of the scales SPS_{19}, SPS_{81}, SPS_{87}, SPS_{88}, and SPS_{102} gave 2-point line plots, because 1 data point was located far from the others. To avoid this, data points situated outside the main experimental range were removed from the original datasets, and new QSPR models were developed. The next column of SM 2 gives a brief description of the method by which the scales were defined and developed by the original authors. In the last column, the literature sources from which the data were taken are listed. The list of solvents (774) and the number of solvent scales in which they are involved is also provided as Supporting Information (SM 3). Altogether, 149 solvents listed in SM 3 are represented in 10 or more of the 127 individual scales. The solvent molecular size varies from 3 atoms (i.e., water) to nearly 70 atoms (i.e., butyl oleate). The number of solvents represented in a single solvent scale varies from 1 to 111. The most common solvent used in the definition of a solvent scale is acetonitrile.

3. Methodology

3.1. QSPR Approach. The initial 3D geometrical structures of the 774 solvent molecules were developed using molecular mechanics force-field methods (MM+), and subsequently optimized by applying semiempirical quantum chemical AM1 (Austin model 1) method within the MOPAC v. 7.0[39] module incorporated in the CODESSA PRO package.[40] The resulting geometry together with other molecular characteristics from the output of quantum chemical calculations was automatically processed by CODESSA PRO software to calculate a large pool of theoretical molecular descriptors (up to 1100). All these descriptors (constitutional, geometrical, topological, electrostatic and charge-related, quantum chemical, and thermodynamical) are derived solely from molecular structure and do not require the knowledge of experimental data. The best multilinear regression (BMLR) algorithm, which is one of the statistical tools available in CODESSA PRO, was then used to find the best QSPR model with up to 5 descriptors, depending on the size of the data set of each scale. The methodology of the applied QSPR treatment has been described by Katritzky, Karelson, and Lobanov.[41] The successful use of the CODESSA PRO software as a tool for various QSPR studies has been reviewed elsewhere.[42]

3.2. PCA Treatment. The principal component analysis (PCA) of a matrix formed by the assembly of related properties for a large data set of structures provides insight into how these related properties depend on each other in a quantitative manner. PCA transforms a given set of data into principal components (PCs) that are orthogonal to each other. A matrix is first constructed comprising the correlations among the variables of interest. The eigenvalues and eigenvectors of the matrix are then determined. The eigenvectors so obtained are orthogonal, and the sum of the eigenvalues equals the original number of variables. Each eigenvector is a linear combination of the original variables and represents a principal component.

Eigenvectors and eigenvalues characterize the property of the square matrix derived from the initial data matrix, and they allow the calculation of the factor scores S and factor loadings L, respectively. The first principal component axis is constructed to account for a maximum amount of variance in the data. The second principal component axis accounts for a maximum amount of the remaining variance in the data under the constraint so that it must be orthogonal to the first component, and so forth, until all component axes are constructed.[43] The PCA tool is well explained in Reichardt's book.[43] The use of PCA for solvent classification has been reviewed.[43]

The scores and the loadings give the information necessary to reconstitute the original physical properties D of any solvent according to eq 1.

$$D = S_1L_1 + S_2L_2 + \cdots + S_nL_n \qquad (1)$$

Thus, the principal component model[37] may be described as follows:

$$X_{ik} = \vec{X}_{ik} + \sum_{a=1}^{A} t_{ia}P_{ak} + e_{ik} \qquad (2)$$

where $\vec{X}_{ik}$ is the mean scaled value of the experimental data (variables) (the scaling weights, W_k, connect $\vec{X}_{ik}$ with the unscaled data, $\vec{X}_{ik}{}' = W_k^{-1}\vec{X}_{ik}$), t_{ia} are scores, P_{ak} are loadings, e_{ik} are residuals, i is the chemical compound (object), k is the experimental measurement (variable), and a is the principal component. The number of principal components (PCs) of scores existing in a characteristic vector space is equal to, or less than, the number of variables in the data set. PCA allows the examination of a set of characteristics (variables) of a class of compounds (objects) and investigates the relations between them. The principal component score values and the principal component loadings are analyzed to characterize the objects and the variables, respectively. The plot of significant principal component score values and loadings allow the recognition of systematic patterns of the molecular characteristics.

4. Results and Discussion

4.1. QSPR Modeling: 127 Multilinear Regression Equations. The general form of the multilinear regression equations is shown in eq 3.

$$\mathrm{SPS}_i = a_0 \pm \sum_j a_jD_j \quad \text{with } i = 1, 2, ..., 127 \text{ and } j = 1, 2, ..., 5 \qquad (3)$$

where a_0 is the intercept and a_j the regression coefficient related to the descriptor D_j.

The QSPR models (127 total) were developed and reported as Supporting Information (see Table 1).

The descriptors involved in the QSPR models have been classified into six classes according to the origin of their calculation. For clarity, six particular symbols have been used in the following discussion for these descriptor classes. The symbols used relate to constitutional (C), geometrical (G), quantum chemical (Q), topological (T), thermodynamic (TH), and electrostatic descriptors (charge related type) (E) respectively (see Table 2). The theoretical background of each descriptor has been explained elsewhere.[41,44−46] The names of the descriptors involved in QSPR models are also listed in Table 2.

In all reported equations (Table 1), the descriptors are given in the decreasing order of Student test (t-test) values. The squared correlation coefficients for 127 models range from 0.726 to 0.999; 18 models have $R^2 < 0.800$ (see Table 3).

As demonstrated in Table 1, most of the QSPR models are characterized by statistically good correlation coefficient values. However, a good fit of a model depends on the quality of the experimental measurements used in the development of the solvent scales. In just 4 models, extreme outliers were found. Thus, for SPS_7, the 4-descriptors model built using the original number of 38 data points has a low squared correlation coefficient value of $R^2 = 0.55$ ($F = 10.11$) due to the presence of two large outliers, *N,N*-dimethylthioformamide and *N*-methylthiopyrrolidinone,[9] which have the highest and lowest observed values. The exclusion of these 2 outliers led to the QSPR model (eq 4) with a large statistical improvement (see Table 2 for the definition of each descriptor).

$$\mathrm{SPS}_7 = 1161.61 - 231.384Q_{11} + 22.062Q_{26} - 90.752E_{45} + 12.059Q_{23} \qquad (4)$$

$$N = 36,\ R^2 = 0.851,\ R^2_{\mathrm{CV}} = 0.80,\ F = 44.33,\ S = 13.81$$

In the case of SPS_{69}, the removal of a single major outlier (THF) resulted in a greatly improved regression equation with the squared correlation coefficient value increasing from 0.570 to 0.828. In the case of SPS_{92}, the elimination of two distinct outliers (*N*-methylformamide and *N*-methylacetamide) improved R^2 from 0.765 to 0.926. The exclusion of dimethyl sulfoxide from the SPS_{102} dataset as a weighted point led to a better fit of 0.840. Interestingly, most of the outliers are amides or thioamides.

Notably, the solvents used have wide structural variability, including molecules without carbon atoms (water, ammonia, hydrazine, and hydrogen sulfide) and molecules without hydrogen atoms, such as carbon tetrachloride. Nevertheless, the overall statistical quality of QSPR models for different solvent scales ranges from satisfactory to excellent.

The multilinear regression equations were used to predict the solvent scales for all 774 solvents. The minimum and maximum ranges of observed and predicted values for 127 solvent scales are given in Table 4. The ranges for the 127 solvent scales predicted using the proposed models (Table 1) indicats that 26 solvent scales have predicted values within the experimental range for 774 solvents. For another 101 solvent scales, the predicted range of values is at most 20% outside the experimental range values.

However, for the 27 solvent scales which possess less than 20 data points, somewhat greater discrepancies exist between the observed and predicted ranges as listed in Table 4. Prediction for those solvent scales of values for other solvents is less certain.

The QSPR model equations developed for the 127 solvent scales contain a total of 168 different descriptors. The frequency of appearance of descriptors of different classes in the total of 127 models (percentage, %) is shown in Figure 1. The 168 individual descriptors included in the 127 QSPR models comprise (i) 10 constitutional (applied 14 times), (ii) 2 geometrical (applied 3 times), (iii) 29 quantum chemical (applied 135 times), (iv) 22 topological (applied 47 times), (v) 13 thermodynamical (applied 23 times), and (vi) 92 electrostatic and of charged partial surface area (applied 203 times). Altogether, the molecular descriptors were applied 425 times.

The descriptors most frequently appearing are the total hybridization component of the molecular dipole (Q_{23}), the maximum atomic orbital electronic population (Q_{11}), the total dipole of the molecule (Q_{22}), the polarity parameter/square distance (E_{75}), and HOMO−1 energy (Q_5), which occur in 19, 17, 14, 12, and 10 of the QSPR models, respectively. Whereas most of these descriptors are directly related to electrostatic dipole−dipole interactions in condensed media, others can be ascribed to the specific hydrogen bond accepting ability of the solvent.

The electrostatic (E)-type descriptors (92) represent the largest group in all QSPR models, as they appear 203 times in QSPR models. These parameters describe the positively and negatively charged surface areas of the solvent molecules (both the absolute values and those relative to the total surface area). These descriptors are calculated proceeding either from empirical (Zefirov) or semiempirical quantum chemical charge distribution

TABLE 1: QSPR Models of the Solvent Scales[a]

solvent scale	equation	N	n	R^2	R^2_{cv}	s
AN	$SPS_1 = 5.280 + 0.516E_{47} - 88.655E_{18} + 0.734E_{86}$	52	3	0.903	0.875	5.648
B	$SPS_2 = 472.348 + 36.860Q_4 - 1105.350C_5 + 40.016C_4 + 48.897Q_{23} + 106.083E_{74}$	71	5	0.808	0.734	32.66
BCo	$SPS_3 = 1.016 - 0.180T_3 - 0.048E_{89} + 0.048Q_{22}$	25	3	0.929	0.895	0.075
B_{KT}	$SPS_4 = -0.063 - 0.008T_{22} + 0.025E_2 + 0.064T_{21} - 1.216E_{63}$	44	4	0.788	0.734	0.135
BPe	$SPS_5 = 0.261 + 0.531T_5 - 0.075E_4$	25	2	0.902	0.876	0.103
Co	$SPS_6 = 0.409 + 0.0997Q_{22} - 4.660E_{18} - 0.020C_3$	25	3	0.967	0.957	0.050
Cu$-\lambda_{max}$	$SPS_7 = 1161.610 - 231.384Q_{11} + 22.062Q_{26} - 90.752E_{45} + 12.059Q_{23}$	36	4	0.851	0.799	13.81
DCo	$SPS_8 = 1.805 - 0.616E_{21} - 0.017E_{73}$	23	2	0.908	0.882	0.076
D_s	$SPS_9 = 54.947 - 14.663C_2 + 61.234E_{59} + 755.614E_5 - 3.337Q_3 + 8.498Q_{23}$	56	5	0.768	0.717	7.481
$E_{(NR)}$	$SPS_{10} = 52.022 - 8.796E_{31} + 5.176E_{76} + 0.174E_3 - 1.272Q_{23} - 0.403Q_{22}$	82	5	0.829	0.798	1.424
E^*_{MLCT}	$SPS_{11} = 218.797 - 4.932T_{16} + 28.088T_5 + 4.326Q_{22}$	33	3	0.941	0.923	3.569
$E_{(CT)\pi}$	$SPS_{12} = 0.021 + 2.514E_{24} + 18.710E_{38} - 0.008T_{13} + 0.010TH_{10}$	28	4	0.823	0.744	0.132
$E_B{}^N$	$SPS_{13} = -0.068 + 0.0027E_{55} + 0.397T_1 + 0.715E_{75} + 0.006E_{86}$	52	4	0.951	0.931	0.047
$E_{CT(A)}$	$SPS_{14} = 20.982 + 27.092E_{15} + 5.004Q_6 + 4.001E_{75}$	23	3	0.922	0.856	0.684
$E_T(30)$	$SPS_{15} = 21.203 + 317.839E_{33} + 12.404E_{75} - 32.281E_{19} + 3.560T_5 + 4.779Q_{11}$	334	5	0.826	0.818	2.917
$E_T(N)$	$SPS_{16} = -1.478 + 0.589E_{31} - 1.369E_{19} - 0.069T_3 + 0.431E_{75} + 1.820Q_{16}$	334	5	0.821	0.811	0.092
$E_T{}^{SO}$	$SPS_{17} = 82.150 + 0.045E_{47} + 9.090E_{75} + 0.164E_{43} - 2.487E_{22}$	35	4	0.965	0.954	0.544
G	$SPS_{18} = 93.141 - 22.207T_3 + 47.263C_6 + 6.161T_{18}$	21	3	0.774	0.676	11.49
${}^2J_{119sn-117sn}$	$SPS_{19} = 480.143 - 106.589T_1 + 3.392Q_8$	18	2	0.933	0.903	4.011
K	$SPS_{20} = -93.597 + 183.926T_7 + 6.786Q_{10} - 2.064E_{73}$	25	3	0.837	0.740	16.38
NCo	$SPS_{21} = 0.407 + 0.731T_5 - 0.104E_4$	25	2	0.890	0.858	0.152
Ov	$SPS_{22} = 0.805 - 0.045E_{73} - 0.531E_{23} + 0.206Q_{22}$	25	3	0.945	0.910	0.155
P_s	$SPS_{23} = 2.273 + 3.170Q_{11} - 1.212T_{10} + 0.997E_1 - 4.181E_{17} + 0.0083E_{70}$	107	5	0.844	0.823	0.911
P_y	$SPS_{24} = 1.812 - 0.449T_3 + 1.446E_{24} - 1.547E_{12} + 0.192E_{65} + 0.0066E_{10}$	93	5	0.839	0.806	0.144
Qm	$SPS_{25} = -607.333 + 491.976T_4 + 192.873E_{39}$	19	2	0.764	0.671	166.8
SA	$SPS_{26} = -0.098 + 0.004E_{55} - 0.395E_{84}$	121	2	0.849	0.837	0.071
SB	$SPS_{27} = 0.125 + 0.025E_2 - 2.274E_{63} - 0.001E_{69} - 0.020E_{56} + 0.117Q_{23}$	200	5	0.828	0.816	0.126
SPP^N	$SPS_{28} = 0.449 + 0.452E_{18} + 0.045Q_{22} + 0.172Q_{11} - 0.027Q_7 + 0.083E_{40}$	100	5	0.870	0.816	0.058
Z	$SPS_{29} = 58.348 + 0.139E_{55} + 26.256E_{75} + 29.183Q_6 + 0.302E_{86}$	60	4	0.906	0.876	2.730
α	$SPS_{30} = 0.044 + 2.572E_{31} - 0.347E_{65}$	184	2	0.773	0.750	0.204
‘	$SPS_{31} = -0.025 - 2.200E_{63} + 1.329E_{36} + 0.017E_3 - 3.949E_{44} - 0.409Q_{18}$	184	5	0.756	0.735	0.147
π^*	$SPS_{32} = 1.135 - 0.323T_3 + 0.085Q_{22} + 0.001Q_{25} + 0.101Q_5 + 0.048TH_2$	216	5	0.751	0.736	0.145
π^*_{azo}	$SPS_{33} = 0.287 + 0.443E_{30} - 0.0015E_8 + 0.008E_{10}$	29	3	0.914	0.881	0.090
χR	$SPS_{34} = 47.341 - 3.754E_{30} + 0.037E_7 - 0.015G_1 - 11.841E_{75} + 0.171E_{88}$	58	5	0.852	0.805	1.343
$\int^{C_6H_5F}$	$SPS_{35} = 5.409 + 0.027C_1 + 0.044E_{72} - 1.073Q_{11}$	23	3	0.952	0.919	0.367
$\int_H{}^{p-NO_2}$	$SPS_{36} = 10.022 + 13.767E_{62} - 1.013E_6 - 2.135E_{75}$	29	3	0.846	0.723	0.394
Δ	$SPS_{37} = 1.671 + 0.801E_{30} + 0.418Q_{13} - 1.716E_{45} - 7.087E_{12} + 0.009E_{82}$	54	5	0.770	0.688	0.267
$\Delta\delta CHCl_3$	$SPS_{38} = -0.191 - 0.385E_{64} - 0.535Q_{23} + 0.048E_{56}$	28	3	0.820	0.731	0.234
$\Delta\nu_A$	$SPS_{39} = 0.566 + 0.116E_{10} - 0.791C_3 + 0.042T_{12}$	27	3	0.856	0.787	1.388
$\Delta\nu_D$	$SPS_{40} = -40.766 + 85.872Q_{23} - 48.478Q_{14} + 6.071E_3 + 751.427E_{32} + 16.236Q_{28}$	92	5	0.842	0.805	27.62
θ_{1K}	$SPS_{41} = 1.031 + 4.748E_{17} - 55.368E_{53} + 0.006E_{79} - 2.628Q_{11} - 0.391Q_5$	80	5	0.780	0.720	0.806
θ_{2K}	$SPS_{42} = 1.112 + 0.097Q_5 + 0.002C_1 - 0.315Q_{20} - 4.721E_{67} - 0.222E_{16}$	80	5	0.772	0.741	0.118
$\wedge$	$SPS_{43} = 543.266 + 27.397T_4 - 9.478Q_{22} - 7.206Q_7$	24	3	0.896	0.839	8.052
$\Lambda_F{}^{MHN12}$	$SPS_{44} = 396.810 - 40.576E_{19} - 4.886Q_{12}$	20	2	0.861	0.829	1.517
$\pi_1{}^*$	$SPS_{45} = 3.054 - 0.433T_3 - 0.001TH_7 + 0.014T_{14} + 0.438Q_{11} + 0.097Q_5$	95	5	0.873	0.847	0.134
$\pi_2{}^*$	$SPS_{46} = 2.285 + 0.448E_{45} + 2.820E_{41} + 0.00031G_1 + 0.129Q_5 + 0.001Q_2$	72	5	0.941	0.924	0.043
ν_{CE}	$SPS_{47} = 665.115 + 0.494G_2 - 1.569T_{11} - 6.049Q_{11}$	22	3	0.881	0.823	1.506
$\Delta\nu_{CI}$	$SPS_{48} = 34.901 + 3.640E_3 + 9.623Q_{23} + 578.580E_{14} - 1.701E_{37} - 0.053TH_7$	66	5	0.726	0.667	11.46
$\Delta\nu_{OH}$	$SPS_{49} = 170.098 + 21.815E_2 + 1119.650E_{74} + 90.806Q_{23} - 274.447T_5 + 361.860T_9$	66	5	0.794	0.751	80.40
C_{p-SCS}	$SPS_{50} = -11.221 + 2.517E_{17}$	5	1	0.806	0.517	0.266
CTTS	$SPS_{51} = 41453.900 + 15779.600Q_{18} + 112014.000E_{62}$	16	2	0.941	0.770	751.2
H	$SPS_{52} = 0.00023 + 2.432Q_6 + 3.177E_{11}$	11	2	0.999	0.984	0.011
Kq^{MMA}	$SPS_{53} = 26.898 + 1.796T_{13} - 2.790E_{82}$	12	2	0.906	0.770	8.112
log γK_c	$SPS_{54} = 8.840 + 0.002TH_7 - 0.186T_{19}$	11	2	0.975	0.952	0.056
m^*	$SPS_{55} = 14.777 + 0.149Q_5 - 6.562Q_{11}$	9	2	0.949	0.900	0.043
pK_{BH+}	$SPS_{56} = 15.579 + 1.723Q_4$	9	1	0.926	0.888	0.378
XX	$SPS_{57} = 23.578 + 1.542Q_5 + 0.018C_1 - 0.243Q_{10}$	20	3	0.923	0.867	0.745
$\int_N{}^{pyrrole}$	$SPS_{58} = 237.126 - 260.005Q_{15} - 404.399E_{42}$	13	2	0.924	0.871	1.282
δ	$SPS_{59} = 16.026 - 16.148C_8 - 2.809TH_{13}$	11	2	0.922	0.839	0.428
δ_0	$SPS_{60} = -10.648 + 0.280E_{46} - 0.1001E_{60}$	15	2	0.853	0.798	2.763
$\lambda_A{}^{NA}$	$SPS_{61} = 5.437 - 0.232TH_4$	7	1	0.879	0.734	1.067
T	$SPS_{62} = 13.448 + 0.101E_{46}$	7	1	0.960	0.940	2.479
$\lambda_A{}^{MS}$	$SPS_{63} = 457.294 - 0.016E_9 - 0.0063TH_7$	13	2	0.912	0.858	0.555
$\phi_f{}^{BBVB}$	$SPS_{64} = 0.591 - 2.101E_{59} + 0.044Q_8$	10	2	0.919	0.859	0.063
$\phi_f{}^{CEA}$	$SPS_{65} = 0.311 - 0.737E_{13}$	8	1	0.917	0.879	0.037
Φ	$SPS_{66} = -0.041 - 0.019TH_5 + 0.356T_1 - 0.001Q_{24}$	23	3	0.952	0.927	0.041
B-2	$SPS_{67} = -18.320 + 125.821Q_{23} + 140.448C_4 + 20.107E_4 - 878.479E_{63} + 8.552E_2$	113	5	0.798	0.774	74.36
C_B	$SPS_{68} = 6.782 + 0.750Q_4 + 2.189E_6 + 3.130T_2 + 0.146E_3 - 0.536E_{39}$	65	5	0.801	0.759	0.665

TABLE 1: (Continued)

solvent scale	equation	N	n	R^2	R^2_{cv}	s
D_H	$SPS_{69} = -12.993 + 12.524E_{39} + 187.614E_{25} - 0.696E_{54}$	24	3	0.828	0.760	7.273
E_B	$SPS_{70} = -0.668 - 4.704E_{49} + 1.920E_{51} - 0.020E_{73} + 1.726E_{59} + 1.117Q_{11}$	65	5	0.780	0.731	0.298
PA	$SPS_{71} = 303.259 - 615.588E_{68} - 118.380Q_{17}$	20	2	0.954	0.929	2.744
$\Delta_{acid}H$	$SPS_{72} = 8.345 + 55.576E_{18} - 36.663E_{75} - 21.361T_8 + 4.217Q_{14} + 1.452Q_{22}$	63	5	0.826	0.786	3.482
ΔH_ν	$SPS_{73} = -605.029 + 531.622E_{92} + 4330.840Q_{11}$	22	2	0.923	0.898	707.3
ΔH°_{solv}	$SPS_{74} = 7.305 + 19.339Q_{20} + 6.774Q_{23} - 2.716E_{57} - 2.618T_{17}$	35	4	0.845	0.797	2.430
ϵ°(SVB)	$SPS_{75} = 0.676 - 0.036Q_3 - 3.892E_{68} - 0.039T_{18}$	29	3	0.816	0.757	0.030
$-\Delta H^\circ_{BF3}$	$SPS_{76} = 260.565 + 38.066E_{30} - 350.145E_{63} - 127.345Q_{11} + 19.023Q_{23} - 10.659Q_{14}$	76	5	0.812	0.778	12.00
M	$SPS_{77} = 2.006 + 0.564Q_{23} - 5.293E_{25} - 0.107Q_3 - 0.178T_{10}$	34	4	0.817	0.731	0.167
$D1$	$SPS_{78} = 0.180 + 0.003E_{81} + 14.344E_{20}$	16	2	0.906	0.853	0.121
a_H	$SPS_{79} = -0.283 + 1.555E_{45} + 2.076E_{75}$	13	2	0.972	0.953	0.055
log K_f	$SPS_{80} = 3.595 - 71.641Q_{15}$	6	1	0.892	0.792	0.396
S_p	$SPS_{81} = 1.127 - 0.187TH_{13} + 0.002E_{47}$	12	2	0.990	0.979	0.010
$-\Delta S_s^\circ$	$SPS_{82} = 25.073 + 188.741C_9 + 23.697T_3$	8	2	0.985	0.977	1.852
X	$SPS_{83} = 0.299 - 0.018Q_3 - 0.019T_6$	28	2	0.923	0.900	0.002
DN	$SPS_{84} = -8.646 + 1.189E_2 + 56.617E_{74} + 10.315Q_{23} + 421.429E_5 + 16.927T_8$	110	5	0.763	0.731	6.267
D_π	$SPS_{85} = 1.785 + 0.925Q_{21} + 0.120E_{66} + 0.0007Q_{25} - 1.160Q_{11}$	34	4	0.754	0.656	0.341
log k_{DC}	$SPS_{86} = -4.147 - 0.047E_{47} + 0.645T_{20} + 0.473Q_{29}$	24	3	0.912	0.884	0.699
R_P	$SPS_{87} = 59.467 + 469.648Q_6 - 56.818E_{77}$	19	2	0.967	0.948	7.242
A	$SPS_{88} = 0.324 + 0.0062E_{47} - 0.0011E_{78} + 0.761E_{75} - 0.056E_{90}$	54	4	0.944	0.927	0.066
A_P	$SPS_{89} = 25.053 + 8417.340E_{34} - 0.263E_{35}$	18	2	0.956	0.937	1.794
BB'	$SPS_{90} = 0.765 - 0.010E_{47} - 0.267T_4 - 0.891Q_{18} + 2.037E_{36} + 3.424C_7$	55	5	0.772	0.705	0.160
B_P	$SPS_{91} = 17.690 - 0.022E_{69} - 0.293Q_3$	18	2	0.847	0.762	0.417
D	$SPS_{92} = 22.109 + 237.044Q_6 + 544.217E_{28} - 2.601Q_{10} - 0.0007T_{15}$	55	4	0.926	0.896	5.881
DC	$SPS_{93} = -158.775 - 3253.410E_{29} + 45.764TH_{13}$	22	2	0.948	0.924	6.481
E	$SPS_{94} = 31.305 + 0.115E_{55} - 0.766TH_9$	84	2	0.920	0.909	1.462
J	$SPS_{95} = -0.014 + 0.085Q_{22} + 0.162E_{40} + 0.366Q_{11} - 0.133TH_1$	57	4	0.846	0.815	0.090
log K	$SPS_{96} = -0.325 + 92.496E_{58} + 0.0006Q_{24}$	27	2	0.931	0.881	0.131
log L^{16}	$SPS_{97} = -0.192 + 0.058Q_1 + 0.0068E_{46} + 0.026E_3$	167	3	0.969	0.966	0.223
log P	$SPS_{98} = 1.092 + 1.054T_{17} + 7.647Q_{18}$	104	2	0.950	0.947	0.578
M	$SPS_{99} = -0.164 - 0.008Q_3 + 0.012TH_9 - 0.021TH_{11} + 0.00041E_{61}$	57	4	0.921	0.902	0.007
N	$SPS_{100} = -0.655 + 0.093Q_{22} + 2.585E_{41} + 0.303Q_{11} + 0.024Q_3$	57	4	0.847	0.815	0.094
P'	$SPS_{101} = 6.202 - 5.198Q_{12} + 2.933T_5 + 1.211E_{30} + 16.899E_{25} - 0.218Q_{27}$	78	5	0.849	0.827	0.854
q^-	$SPS_{102} = -0.395 - 0.958Q_{18} + 0.239Q_{11} - 0.020Q_{29}$	28	3	0.840	0.790	0.055
q^+	$SPS_{103} = 0.271 + 2.829E_{62} + 0.032Q_5 + 0.008TH_8$	29	3	0.938	0.924	0.017
S	$SPS_{104} = -0.183 - 0.0021E_{91} + 0.017E_{56} + 0.515E_{74} + 0.026T_{18}$	46	4	0.896	0.870	0.042
S'	$SPS_{105} = 0.936 + 2.577Q_6 + 0.671T_5 + 0.156Q_{13} + 0.005E_{48}$	46	4	0.901	0.877	0.164
V_{mc}	$SPS_{106} = 0.288 + 0.061C_{10} + 0.0023E_{69} - 0.077Q_{28}$	29	3	0.993	0.987	0.028
Xd^R	$SPS_{107} = 0.238 - 0.030Q_8 + 0.027TH_3 + 1.720E_{14} - 0.028Q_{23}$	52	4	0.880	0.849	0.025
Xe^R	$SPS_{108} = 0.240 + 0.049Q_8 + 0.101Q_{23} - 0.115T_5 + 0.032Q_{28} + 0.124E_{31}$	52	5	0.849	0.805	0.042
Xn^R	$SPS_{109} = 0.360 - 0.030E_{57} + 0.048Q_{14} - 0.062Q_{23} + 0.0022E_{82} - 0.002E_3$	52	5	0.842	0.786	0.026
X_d	$SPS_{110} = -0.009 - 0.143E_6 - 0.044Q_{26} + 3.694E_{26} + 0.0020E_{85} - 0.011Q_{22}$	72	5	0.819	0.768	0.028
X_e	$SPS_{111} = -2.565 + 0.007E_2 + 0.029Q_7 - 0.053Q_{14} + 2.975Q_{16} + 0.046Q_{23}$	72	5	0.822	0.782	0.041
X_n	$SPS_{112} = 0.412 - 7.363E_{29} - 0.297E_{75} + 0.193T_1 + 0.00030Q_{25} - 0.039Q_{23}$	72	5	0.810	0.769	0.029
$\Delta G_{6A}{}^\circ$	$SPS_{113} = 16.909 + 147.416E_{58} + 0.794Q_5 + 0.436Q_{29}$	25	3	0.790	0.702	0.881
ΔH^{acid}	$SPS_{114} = -13.454 + 16.918E_{45} + 1.212E_{56} - 4.263Q_9 + 30.048E_{74}$	36	4	0.893	0.809	2.480
$-\Delta H_f$	$SPS_{115} = 23.349 - 4.081Q_{19} + 26.757E_{63} + 0.812Q_4 + 13.424E_{62} - 1.302Q_{14}$	53	5	0.816	0.756	0.610
$\epsilon^\circ_{alumina}$	$SPS_{116} = 0.405 + 1.528E_{74} - 0.004E_{91} + 0.045E_{27}$	23	3	0.948	0.922	0.056
ϵ°_{silica}	$SPS_{117} = 0.819 + 1.479E_{74} - 0.003E_{87}$	19	2	0.947	0.922	0.047
e_b	$SPS_{118} = 0.199 - 0.0050Q_3 - 0.00013E_{46} + 0.036E_{74}$	29	3	0.873	0.826	0.004
$\Theta(\in_B)$	$SPS_{119} = 0.203 + 0.013E_{50} - 0.205Q_{18} - 0.00039E_8 + 0.032Q_{22}$	39	4	0.918	0.890	0.035
μ_D	$SPS_{120} = 0.519 + 0.562Q_{29} - 0.0078E_{71} + 2.493E_{52} + 0.088TH_5$	39	4	0.941	0.912	0.331
$\pi 1$	$SPS_{121} = 0.119 + 0.00021E_{69} - 0.00007TH_7 - 0.036E_{49}$	29	3	0.932	0.902	0.003
$\sigma 1$	$SPS_{122} = -11.577 + 0.411TH_9 + 0.0092TH_{12}$	25	2	0.985	0.982	0.139
ϵ_a	$SPS_{123} = 0.183 - 0.009Q_7 - 0.0011E_{83}$	29	2	0.991	0.989	0.002
γ_{SO_2}	$SPS_{124} = -1.035 - 0.121Q_4 - 0.00041E_{80}$	17	2	0.835	0.773	0.061
δ_H	$SPS_{125} = 21.980 + 33.888E_{15} + 32.882Q_6 - 1.649T_{10}$	30	3	0.928	0.908	2.038
Y	$SPS_{126} = -1.686 + 0.079E_{40} + 0.141Q_{11} + 0.025Q_{22} - 0.00038E_{78} + 1.883Q_{16}$	66	5	0.921	0.902	0.023
P	$SPS_{127} = -0.287 + 0.017TH_9 + 0.013Q_5 - 0.00001TH_6 + 0.493E_{38} + 0.00024E_{10}$	66	5	0.961	0.951	0.004

[a] Where N is the number of datapoints, n is the number of descriptors, R^2 is the squared correlation coefficient, R^2_{cv} is the squared cross-validated correlation coefficient, and s is the standard deviation.

and the van der Waals shape of the molecule. In the 127 equations, the count of hydrogen donor sites (Zefirov, E_3), the count of hydrogen donor sites (MOPAC PC, E_2), HASA H-acceptors surface area (MOPAC PC, E_{47}), minimum partial

TABLE 2: Descriptors Used in the QSPR Models of the Solvent Scales

descriptor name	symbol	occurrence
Constitutional		
molecular weight	C_1	3
number of benzene rings	C_2	1
number of H atoms	C_3	2
number of N atoms	C_4	2
relative number of benzene rings	C_5	1
relative number of H atoms	C_6	1
relative number of rings	C_7	1
relative number of double bonds	C_8	1
relative number of triple bonds	C_9	1
total number of bonds	C_{10}	1
Geometrical		
gravitation index (all bonds)	G_1	2
shadow plane *ZX*	G_2	1
Thermodynamical		
internal heat capacity (300 K)/*n* atoms	TH_1	1
rot entropy (300 K)	TH_2	1
rot entropy (300 K)/*n* atoms	TH_3	1
thermodynamic heat of formation of the molecule at 300 K	TH_4	1
thermodynamic heat of formation of the molecule at 300 K/*n* atoms	TH_5	2
tot enthalpy (300 K)	TH_6	1
tot enthalpy (300 K)/*n* atoms	TH_7	5
tot entropy (300 K)/*n* atoms	TH_8	1
translational entropy (300 K)	TH_9	4
vib entropy (300 K)	TH_{10}	1
vib entropy (300 K)/*n* atoms	TH_{11}	1
zero-point vibrational energy	TH_{12}	1
zero-point vibrational energy/*n* atoms	TH_{13}	3
Topological		
average bonding information content (order 0)	T_1	4
average bonding information content (order 1)	T_2	1
average complementary information content (order 0)	T_3	7
average complementary information content (order 1)	T_4	3
average information content (order 0)	T_5	8
average information content (order 1)	T_6	1
average structural information content (order 0)	T_7	1
average structural information content (order 1)	T_8	2
average structural information content (order 2)	T_9	1
Balaban index	T_{10}	3
bonding information content (order 1)	T_{11}	1
complementary information content (order 0)	T_{12}	1
complementary information content (order 2)	T_{13}	2
information content (order 0)	T_{14}	1
Kier shape index (order 1)	T_{15}	1
Kier & Hall index (order 0)	T_{16}	1
Kier & Hall index (order 1)	T_{17}	2
Kier & Hall index (order 2)	T_{18}	3
Randic index (order 2)	T_{19}	1
structural information content (order 0)	T_{20}	1
structural information content (order 1)	T_{21}	1
Wiener index	T_{22}	1
Electrostatic		
count of H-acceptor sites (Zefirov PC)	E_1	1
count of H-donors sites (MOPAC PC)	E_2	6
count of H-donors sites (Zefirov PC)	E_3	7
difference (pos − neg) in charged part of charged surface area (Zefirov's PC)	E_4	3
difference (pos − neg) in charged part of partial charged surface area (MOPAC PC)	E_5	2
difference (pos − neg) in charged partial surface area (Zefirov's PC)	E_6	3
difference (pos − neg) in charged surface areas (MOPAC PC)	E_7	1
DPSA-1 difference in CPSAs (PPSA1 − PNSA1) (MOPAC PC)	E_8	2
DPSA1 difference in CPSAs (PPSA1 − PNSA1) (Zefirov PC)	E_9	1
DPSA-3 difference in CPSAs (PPSA3 − PNSA3) (MOPAC PC)	E_{10}	4
FHACA fractional HACA (HACA/TMSA) (MOPAC PC)	E_{11}	1
FHASA fractional HASA (HASA/TMSA) (MOPAC PC)	E_{12}	2
FHBSA fractional HBSA (HBSA/TMSA) (MOPAC PC)	E_{13}	1
FHDCA fractional HDCA (HDCA/TMSA) (MOPAC PC)	E_{14}	2
FHDSA fractional HDSA (HDSA/TMSA) (MOPAC PC)	E_{15}	2
FNSA1 fractional PNSA (PNSA-1/TMSA) (Zefirov PC)	E_{16}	1

TABLE 2: (Continued)

descriptor name	symbol	occurrence
Electrostatic		
FNSA-2 fractional PNSA (PNSA-2/TMSA) (MOPAC PC)	E_{17}	3
FNSA2 fractional PNSA (PNSA-2/TMSA) (Zefirov PC)	E_{18}	4
FNSA-3 fractional PNSA (PNSA-3/TMSA) (MOPAC PC)	E_{19}	3
FNSA3 fractional PNSA (PNSA-3/TMSA) (Zefirov PC)	E_{20}	1
FPSA-1 fractional PPSA (PPSA-1/TMSA) (MOPAC PC)	E_{21}	1
FPSA1 fractional PPSA (PPSA-1/TMSA) (Zefirov PC)	E_{22}	1
FPSA-2 fractional PPSA (PPSA-2/TMSA) (MOPAC PC)	E_{23}	1
FPSA2 fractional PPSA (PPSA-2/TMSA) (Zefirov PC)	E_{24}	2
FPSA-3 fractional PPSA (PPSA-3/TMSA) (MOPAC PC)	E_{25}	3
FPSA3 fractional PPSA (PPSA-3/TMSA) (Zefirov PC)	E_{26}	1
HA dependent HDCA-1 (MOPAC PC)	E_{27}	1
HA dependent HDCA-1/TMSA (MOPAC PC)	E_{28}	1
HA dependent HDCA-1/TMSA (Zefirov PC)	E_{29}	2
HA dependent HDCA-2 (MOPAC PC)	E_{30}	5
HA dependent HDCA-2 (Zefirov PC)	E_{31}	4
HA dependent HDCA-2/SQRT(TMSA) (MOPAC PC)	E_{32}	1
HA dependent HDCA-2/SQRT(TMSA) (Zefirov PC)	E_{33}	1
HA dependent HDCA-2/TMSA (Zefirov PC)	E_{34}	1
HA dependent HDSA-1 (Zefirov PC)	E_{35}	1
HA dependent HDSA-1/TMSA (MOPAC PC)	E_{36}	2
HACA-1 (MOPAC PC)	E_{37}	1
HACA-1/TMSA (Zefirov PC)	E_{38}	2
HACA-2 (MOPAC PC)	E_{39}	3
HACA-2 (Zefirov PC)	E_{40}	3
HACA-2/SQRT(TMSA) (Zefirov PC)	E_{41}	2
HACA-2/TMSA (MOPAC PC)	E_{42}	1
H-acceptors CPSA (version 2)	E_{43}	1
H-acceptors FCPSA (version 2)	E_{44}	1
H-acceptors FPSA (version 2)	E_{45}	5
H-acceptors PSA (version 2)	E_{46}	4
HASA H-acceptors surface area (MOPAC PC)	E_{47}	6
HASA-1 (MOPAC PC)	E_{48}	1
HASA-1/TMSA (Zefirov PC)	E_{49}	2
HASA-2 (MOPAC PC)	E_{50}	1
HASA-2/SQRT(TMSA) (MOPAC PC)	E_{51}	1
HASA-2/SQRT(TMSA) (Zefirov PC)	E_{52}	1
HASA-2/TMSA (MOPAC PC)	E_{53}	1
HBCA H-bonding charged surface area (MOPAC PC)	E_{54}	1
HBSA H-bonding surface area (MOPAC PC)	E_{55}	4
HDCA H-donors charged surface area (MOPAC PC)	E_{56}	4
H-donors CPSA (version 2)	E_{57}	2
H-donors FCPSA (version 2)	E_{58}	2
H-donors FPSA (version 2)	E_{59}	3
H-donors PSA (version 2)	E_{60}	1
HDSA H-donors surface area (MOPAC PC)	E_{61}	1
max partial charge (Zefirov) for all atom types	E_{62}	4
min partial charge (Zefirov) for all atom types	E_{63}	6
min (#HA, #HD) (MOPAC PC)	E_{64}	1
min (#HA, #HD) (Zefirov PC)	E_{65}	2
negatively charged part of charged surface area (MOPAC PC)	E_{66}	1
negatively charged part of partial charged surface area (MOPAC PC)	E_{67}	1
negatively charged part of partial charged surface area (Zefirov's PC)	E_{68}	2
PNSA-1 partial negative surface area (MOPAC PC)	E_{69}	4
PNSA1 partial negative surface area (Zefirov PC)	E_{70}	1
PNSA-2 total charge weighted PNSA (MOPAC PC)	E_{71}	1
PNSA2 total charge weighted PNSA (Zefirov PC)	E_{72}	1
PNSA-3 atomic charge weighted PNSA (MOPAC PC)	E_{73}	4
polarity parameter (Zefirov)	E_{74}	8
polarity parameter/square distance (Zefirov)	E_{75}	12
positively charged partial surface area (MOPAC PC)	E_{76}	1
positively charged partial surface area (Zefirov's PC)	E_{77}	1
positively charged surface area (MOPAC PC)	E_{78}	2
positively charged surface area (Zefirov's PC)	E_{79}	1
PPSA-1 partial positive surface area (MOPAC PC)	E_{80}	1
PPSA1 partial positive surface area (Zefirov PC)	E_{81}	1
PPSA-3 atomic charge weighted PPSA (MOPAC PC)	E_{82}	3
PPSA3 atomic charge weighted PPSA (Zefirov PC)	E_{83}	1
RNCG relative negative charge (QMNEG/QTMINUS) (MOPAC PC)	E_{84}	1

TABLE 2: (Continued)

descriptor name	symbol	occurrence
Electrostatic		
RNCS relative negative charged SA (SAMNEG*RNCG) (MOPAC PC)	E_{85}	1
RPCS relative positive charged SA (SAMPOS*RPCG) (MOPAC PC)	E_{86}	3
TMSA total molecular surface area (Zefirov PC)	E_{87}	1
WNSA1 weighted PNSA (PNSA1*TMSA/1000) (Zefirov PC)	E_{88}	1
WNSA-3 weighted PNSA (PNSA3*TMSA/1000) (MOPAC PC)	E_{89}	1
WNSA3 weighted PNSA (PNSA3*TMSA/1000) (Zefirov PC)	E_{90}	1
WPSA1 weighted PPSA (PPSA1*TMSA/1000) (Zefirov PC)	E_{91}	2
WPSA-3 weighted PPSA (PPSA3*TMSA/1000) (MOPAC PC)	E_{92}	1
Quantum Chemical		
ALFA polarizability (DIP)	Q_1	1
final heat of formation	Q_2	1
HOMO−LUMO energy gap	Q_3	8
HOMO energy	Q_4	5
HOMO − 1 energy	Q_5	10
image of the Onsager−Kirkwood solvation energy	Q_6	7
LUMO energy	Q_7	4
LUMO + 1 energy	Q_8	4
max antibonding contribution of one MO	Q_9	1
max atomic force constant	Q_{10}	3
max atomic orbital electronic population	Q_{11}	17
max bonding contribution of one MO	Q_{12}	2
max net atomic charge	Q_{13}	2
max $\pi-\pi$ bond order	Q_{14}	6
max $\sigma-\pi$ bond order	Q_{15}	2
max $\sigma-\sigma$ bond order	Q_{16}	3
min atomic orbital electronic population	Q_{17}	1
min net atomic charge	Q_{18}	6
no. of occupied electronic levels/# atoms	Q_{19}	1
topographic electronic index (all bonds)	Q_{20}	2
topographic electronic index (all pairs)	Q_{21}	1
tot dipole of the molecule	Q_{22}	14
tot hybridization comp. of the molecular dipole	Q_{23}	19
tot molecular 1-center E−E repulsion	Q_{24}	2
tot molecular 1-center E−N attraction	Q_{25}	3
tot molecular 2-center exchange energy	Q_{26}	2
tot molecular 2-center resonance energy	Q_{27}	1
tot molecular electrostatic interaction	Q_{28}	3
tot point-charge comp. of the molecular dipole	Q_{29}	4

TABLE 3: Distribution of the Solvent Scales According to the R^2 Value Overall QSPR Models

range of R^2 value	solvent scales
0.726−0.800	18
0.801−0.850	32
0.851−0.900	17
0.901−0.950	41
0.951−0.999	19

charge (Zefirov) for all atom types (E_{63}), and FNSA2 fractional PNSA (PNSA-2/TMSA) (Zefirov, E_{18}) were respectively involved 7, 6, 6, 6, and 4 times each, whereas HDCA H-donors charged surface area (MOPAC PC) occurred 4 times. The two polarity parameters (E_{74} and E_{75}) occurred 8 and 12 times, respectively. The frequent involvement of such descriptors is expected, as the charge distribution of molecules directly determines the solute−solvent interactions in liquid media.

The next most important descriptors involved in the QSPR models are of quantum chemical origin, appearing altogether 135 times. The HOMO−1 energy, LUMO energy, HOMO−LUMO energy gap, total hybridization component of the molecular dipole, total dipole of the molecule, and maximum atomic orbital electronic population can be directly related to the polarity of molecules.

Topological descriptors, which are directly calculated from the structures using graph theory, appear 47 times in the models. They represent the atomic and group connectivity in the molecules.

Charge-distribution-related descriptors that reflect the electrostatic interactions of molecules appear 45 times in QSPR model equations. These descriptors are also related to the polarity of the molecule.

Thermodynamic descriptors, which encode inherent characteristics of the molecule, appear 23 times. Constitutional and geometrical descriptors, which describe molecular shape and size, appear only 14 and 3 times, respectively.

Three different types of experimental techniques used for the measurement of the solvent polarity scales were utilized to divide these 127 solvent scales. Thus, 88 out of the 127 solvent scales were classified as follows: (i) spectroscopic measurements (67), (ii) equilibrium measurements (17), and (iii) kinetic measurements (4). The remaining 39 solvent scales were grouped into class (iv) entitled “other measurements” (see SM 2).

We now discuss the distribution of descriptors in the four individual classes of solvent scales:

(i) The 67 scales derived from spectroscopic measurements (SPS_1-SPS_{67}) such as $E_T(30)$, $E_T(N)$, α, β, π^*, π^*(azo), or χ_R (Table 1) should reflect primarily the polarity of the solvent. Electrostatic descriptors appear 110 times in the QSPR models developed for these 67 scales. Quantum chemical descriptors make the next most significant contribution, occurring 59 times. The remaining contributions comprise topological (38), constitutional (11), thermodynamical (9), and geometrical (3) descriptors.

TABLE 4: Maximum and Minimum Values of the Observed and Predicted Data for 127 Solvent Scales

solvent scales	*N*	obsd (min)	obsd (max)	pred (min)	pred (max)
SPS$_{1}$	52	0	83.6	5.280	177.8
SPS$_{2}$	71	31	292	16.10	353.1
SPS$_{3}$	25	0.38	1.25	0.257	3.430
SPS$_{4}$	44	−0.040	0.9964	−46.31	0.847
SPS$_{5}$	25	0.38	1.36	−0.103	4.128
SPS$_{6}$	25	0.12	0.98	−0.656	9.511
SPS$_{7}$	36	413	763	477.4	769.4
SPS$_{8}$	23	1.23	2	1.190	3.491
SPS$_{9}$	56	6	69	−56.25	66.41
SPS$_{10}$	82	44.85	59.12	43.23	59.53
SPS$_{11}$	33	208.4	265.8	132.2	284.8
SPS$_{12}$	28	−0.044	1	−0.004	2.287
SPS$_{13}$	52	0	1	−0.011	0.950
SPS$_{14}$	23	20.3	27.5	20.98	38.20
SPS$_{15}$	334	30.7	65.3	30.42	63.50
SPS$_{16}$	334	0	1.068	−0.074	1.066
SPS$_{17}$	35	79.9	88.8	79.66	90.99
SPS$_{18}$	21	36	118	35.42	122.3
SPS$_{19}$	18	424.2	471.8	379.6	477.0
SPS$_{20}$	25	6	150	−6.753	298.2
SPS$_{21}$	25	0.57	1.86	−0.109	5.790
SPS$_{22}$	25	0.3	2.38	−1.175	6.374
SPS$_{23}$	107	2.41	12.1	2.737	17.93
SPS$_{24}$	93	0.52	1.95	0.328	2.248
SPS$_{25}$	19	118	1166	−430.7	1920.3
SPS$_{26}$	121	−0.026	0.717	−0.081	0.853
SPS$_{27}$	200	0.014	1	−0.184	1.273
SPS$_{28}$	100	0.214	1.009	−0.045	1.162
SPS$_{29}$	60	54	94.6	58.35	98.25
SPS$_{30}$	184	0	1.96	−0.396	2.612
SPS$_{31}$	184	−0.08	1.43	−0.135	1.299
SPS$_{32}$	216	−0.48	1.08	−0.433	1.318
SPS$_{33}$	29	−0.09	1.03	−0.405	2.273
SPS$_{34}$	58	33.6	50.9	20.28	55.08
SPS$_{35}$	23	1.9	10.3	−23.19	13.62
SPS$_{36}$	29	9	14.05	9.009	14.00
SPS$_{37}$	54	1.1	3.71	−1.874	4.289
SPS$_{38}$	28	−1.96	0.16	−2.085	0.287
SPS$_{39}$	27	−5.9	9	−17.16	14.48
SPS$_{40}$	92	−21	242	−56.30	345.3
SPS$_{41}$	80	−3.91	4.53	−10.19	6.109
SPS$_{42}$	80	−0.961	0.405	−0.940	0.614
SPS$_{43}$	24	510	613	506.0	659.0
SPS$_{44}$	20	384.9	399.8	386.6	401.4
SPS$_{45}$	95	0.627	2.127	0.684	2.306
SPS$_{46}$	72	0.797	1.624	0.514	1.953
SPS$_{47}$	22	652.8	665.4	642.2	686.3
SPS$_{48}$	66	0	91	−11.001	142.7
SPS$_{49}$	66	0	706	−19.327	1010.3
SPS$_{50}$	5	−12.21	−11.13	−17.084	−11.22
SPS$_{51}$	16	37861	46870	26147.0	63759.3
SPS$_{52}$	11	0	1	0.000	1.366
SPS$_{53}$	12	0.68	80	−145.4	489.0
SPS$_{54}$	11	8.95	9.83	6.984	10.71
SPS$_{55}$	9	0.22	0.76	−0.576	5.429
SPS$_{56}$	9	−3.9	−0.14	−6.454	1.974
SPS$_{57}$	20	2	11.5	−2.318	14.67
SPS$_{58}$	13	224.2	238.77	183.7	237.1
SPS$_{59}$	11	−2.54	2.8	−4.143	12.57
SPS$_{60}$	15	−13.1	9.9	−12.91	127.5
SPS$_{61}$	7	13.79	20.8	−15.19	211.3
SPS$_{62}$	7	13	40	13.45	63.30
SPS$_{63}$	13	450.6	455.7	441.0	461.4
SPS$_{64}$	10	0.216	0.78	−0.952	0.771
SPS$_{65}$	8	0.02	0.37	−0.427	0.311
SPS$_{66}$	23	−0.01	0.65	−0.100	0.626
SPS$_{67}$	113	0	727	−815.1	1063.3
SPS$_{68}$	65	0.1	5.73	−4.125	6.904
SPS$_{69}$	24	−14	40	−32.33	47.37
SPS$_{70}$	65	0.04	2.59	−0.196	3.822
SPS$_{71}$	20	165	207	146.0	263.2
SPS$_{72}$	63	−32.47	2.1	−106.2	11.85
SPS$_{73}$	22	6570	17010	6508.5	56534.6
SPS$_{74}$	35	1.46	24.11	−4.797	53.44
SPS$_{75}$	29	−0.072	0.185	−0.359	0.429
SPS$_{76}$	76	10	139.51	−3.622	234.0
SPS$_{77}$	34	−0.12	1.35	−0.691	1.595
SPS$_{78}$	16	−0.01	1.06	−1.218	2.916
SPS$_{79}$	13	0.15	1.14	−0.282	1.580
SPS$_{80}$	6	0	3.25	−10.48	3.595
SPS$_{81}$	12	0	1	−0.003	0.897
SPS$_{82}$	8	43.9	82.2	34.55	158.1
SPS$_{83}$	28	0.0020	0.0289	−0.025	0.191
SPS$_{84}$	110	0	61	−28.79	65.11
SPS$_{85}$	34	−1.56	0.704	−3.527	2.520
SPS$_{86}$	24	−5.13	2.8	−5.513	10.92
SPS$_{87}$	19	1.52	212	2.649	210.9
SPS$_{88}$	54	0	1.72	−0.595	1.657
SPS$_{89}$	18	19.2	48	15.08	49.15
SPS$_{90}$	55	−0.03	1.19	−0.815	1.420
SPS$_{91}$	18	10.6	14	3.618	15.13
SPS$_{92}$	55	1.9	109.5	−13.8	105.6
SPS$_{93}$	22	0	100	−102.5	117.3
SPS$_{94}$	84	−0.2	21.8	−2.850	22.49
SPS$_{95}$	57	0.231	0.984	0.213	1.290
SPS$_{96}$	27	−0.029	1.971	−0.200	4.265
SPS$_{97}$	167	0.922	6.705	0.410	16.69
SPS$_{98}$	104	−1.3	13.7	−3.623	14.47
SPS$_{99}$	57	0.169	0.27	0.147	0.274
SPS$_{100}$	57	−0.006	0.757	−0.115	1.029
SPS$_{101}$	78	0.1	10.2	0.585	11.19
SPS$_{102}$	28	0.0101	0.7204	0.001	1.308
SPS$_{103}$	29	0	0.2161	−0.022	0.557
SPS$_{104}$	46	−0.337	0.154	−1.953	0.199
SPS$_{105}$	46	1.11	3.1	1.425	3.536
SPS$_{106}$	29	0.3647	1.638	0.018	5.744
SPS$_{107}$	52	0.08	0.4	0.098	0.410
SPS$_{108}$	52	0.22	0.66	0.120	0.740
SPS$_{109}$	52	0.24	0.45	0.127	0.581
SPS$_{110}$	72	0.12	0.41	0.089	0.485
SPS$_{111}$	72	0.23	0.59	−0.619	0.642
SPS$_{112}$	72	0.24	0.53	0.209	0.505
SPS$_{113}$	25	6.35	13.15	6.194	13.33
SPS$_{114}$	36	−0.52	29.58	−10.61	32.21
SPS$_{115}$	53	4.07	9.5	−3.320	11.41
SPS$_{116}$	23	0	0.95	−3.501	1.299
SPS$_{117}$	19	0	0.7	−1.584	0.876
SPS$_{118}$	29	0.1128	0.1524	0.089	0.164
SPS$_{119}$	39	0.103	0.467	−0.100	0.769
SPS$_{120}$	39	0	4.7	−0.453	14.38
SPS$_{121}$	29	0.086	0.1316	0.055	0.176
SPS$_{122}$	25	2.77	8	2.789	11.63
SPS$_{123}$	29	0.1402	0.1911	0.135	0.215
SPS$_{124}$	17	0.023	0.507	−0.272	0.515
SPS$_{125}$	30	11.9	47.9	11.08	47.23
SPS$_{126}$	66	0.226	0.484	−0.101	0.607
SPS$_{127}$	66	0.169	0.26	0.088	0.273

(ii) For the 17 scales obtained from measurements of chemical equilibria (SPS$_{68}$−SPS$_{84}$), the most common descriptor types are charge-distribution-related electrostatic descriptors (27) and quantum chemical descriptors (18). Regression equations with satisfactory statistical parameters (R^2 between 0.763 and 0.828) were obtained for solvent scales as D_H, E_B, C_B, μ, and DN (Table 1).

(iii) The 4 QSPR models derived from kinetic measurements (D_π, log k_{DC}, R_P) contain electrostatic (3), quantum chemical (5), and topological (1) descriptors.

(iv) The most commonly appearing descriptors in the QSPR for the 39 solvent scales obtained from miscellaneous types of measurements (SPS$_{88}$−SPS$_{127}$ in Table 3) are charged partial surface area electrostatic descriptors (63) and quantum chemical

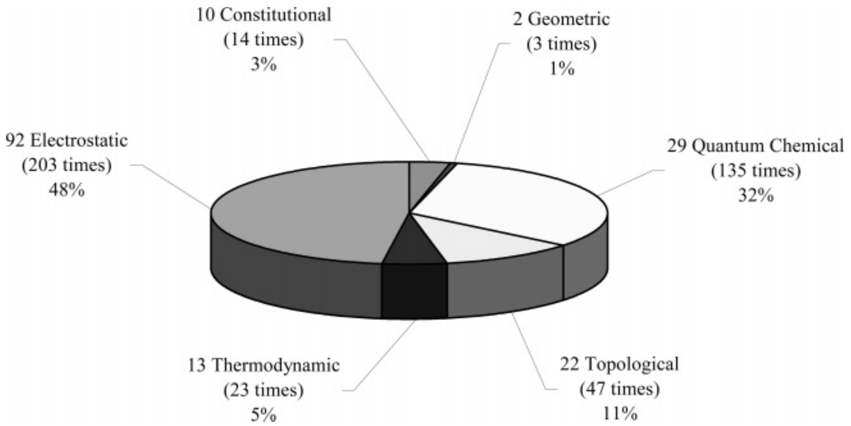


Figure 1. Classes of the 168 descriptors involved 425 times in 127 QSPR models; the percentages are calculated for the occurrence of descriptors.

descriptors (54). Also, the solvent scales defined as the hydrophobic fragmental constants (log P), gas−liquid partition coefficient (log L^{16}), dielectric constant (d), and strong dipole (X_n) are akin to charge-related and quantum chemical descriptors.

The QSPR models for some polarity solvent scales (α, β, π^*, $E_T(N)$, E_T^{SO}) are well-described by some charge-distribution-related descriptors.[44] For example, the model for the hydrogen bond donor acidity solvent scale (α) is satisfactorily described ($R^2 = 0.77$) by two charge-distribution-related descriptors, i.e., min (#HA, #HD) (Zefirov), and HA-dependent HDCA-2 (Zefirov). Both descriptors describe the hydrogen bond donor (and acceptor) ability of the molecule.

For a more precise illustration, let us consider the solvent scale Z (SPS$_{29}$), which is defined using the transition energies for the charge transfer band of the complex from 1-ethyl-4 methoxycarbonylpyridinium iodide as a measure of solvent polarity.[18] The QSPR model for this scale is given by eq 5.

$$\mathrm{SPS}_{29} = 58.348 + 0.139E_{55} + 26.256E_{75} + 29.183Q_6 + 0.302E_{86} \quad (5)$$

The 4-parameter regression model (eq 5) utilizes 3 electrostatic descriptors plus 1 quantum chemical descriptor and shows a good fit ($R^2 = 0.906$) for 60 datapoints. The 3 electrostatic descriptors, HBSA H-bonding surface area (MOPAC PC) (E_{55}), RPCS relative positive charged SA (SAMPOS*RPCG) (MOPAC PC) (E_{86}), and Zefirov's polarity parameter/square distance (E_{75}), are related to the polarity of the solvent molecule. The image of the Onsager−Kirkwood solvation energy (Q_6) describes the interaction of the molecule with the reaction field that it creates in the surrounding medium. Among the solvents considered, water is reported to have the highest Z value (94.6). The predicted value for water using eq 5 is 94.8, which fits well with the experimental data.

The intercorrelation of the above-mentioned four descriptors is given in Table 5.

TABLE 5: Descriptor Multicollinearity for Eqs 4 and 5[a]

dependent variable	independent variable	variance inflation factor
	Equation 4	
Q_{23}	Q_{11}	1.357
	Q_{26}	1.276
	E_{45}	1.244
Q_{11}	Q_{23}	1.211
	Q_{26}	1.201
	E_{45}	1.348
Q_{26}	Q_{23}	1.144
	Q_{11}	1.207
	E_{45}	1.307
E_{45}	Q_{23}	1.103
	Q_{11}	1.34
	Q_{26}	1.293
	Equation 5	
E_{75}	E_{86}	1.046
	E_{55}	1.031
	Q_6	1.031
E_{86}	E_{75}	1.319
	E_{55}	1.318
	Q_6	1.04
E_{55}	E_{75}	1.011
	E_{86}	1.025
	Q_6	1.035
Q_6	E_{75}	1.298
	E_{86}	1.038
	E_{55}	1.329

[a] See Table 2 for definition of all descriptors.

As one can see from Table 5, the lower values of the variance inflation factor (VIF) show that there are no severe multicollinearity effects present in the discussed models (eqs 4 and 5, respectively). However, it should be mentioned that there is no precise limit of variance inflation factor (or tolerance, defined as tolerance = 1/VIF) according to which the presence or absence of multicollinearity can be assessed.

However, the frequency of occurrence of various descriptors can also be analyzed using a different classification of molecular descriptors, based on their physical definition related to solvation. According to the commonly accepted classification of specific physical interactions between the solute and solvent molecules in different media,[9,47−50] these molecular descriptors can be classified as follows (see SM 4):

- Descriptors related to *cavity formation* in the condensed medium, which characterize the size and the shape of the molecule.
- Descriptors related to *electrostatic interactions* between the solute and solvent molecules, which characterize the charge distribution in the solute molecules.
- Descriptors related to *dispersion interactions* between the solute and solvent molecules, which characterize the electron orbital distribution and energetics.
- Descriptors related to *specific interactions (hydrogen bonding)* between the solute and solvent molecules, which characterize the charge distribution on the potentially hydrogen bonded atoms. In principle, the various counts of heteroatoms and the bonds with heteroatoms also belong to this class of descriptors, since the respective atoms potentially participate in the specific intermolecular interactions.

The analysis of the descriptors appearing in the QSPR models for solvent scales reveals that the class of descriptors that relate

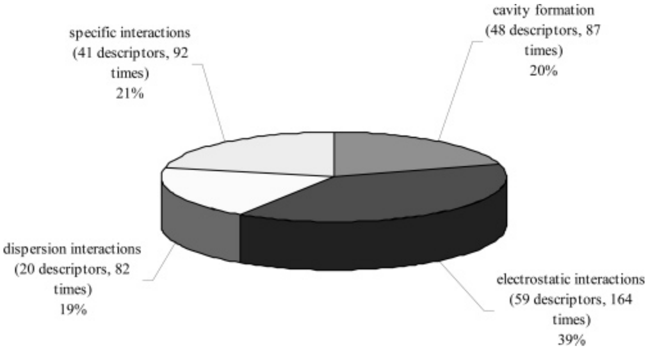


Figure 2. Classes of molecular descriptors based on their physical definition related to solvation (the percents are calculated for the occurrence).

to electrostatic interactions in liquid media have the largest occurrence: 59 individual descriptors. The other classes occur less frequently (48 descriptors related to the cavity formation, 41 descriptors related to hydrogen bonding, and only 20 descriptors related to the dispersion interactions), as shown in Figure 2.

The descriptors related to electrostatic interactions appear most frequently in the QSPR models for individual solvent scales—altogether 164 times. The other classes of molecular descriptors have similar frequencies of occurrence in the QSPR models: the descriptors related to cavity formation, 87 times; the descriptors related to dispersion interactions, 82 times; and the descriptors featuring the hydrogen bonding properties of compounds, 92 times (see SM 4).

Notably, for a given class, certain descriptors occur very frequently. Thus, of the 48 descriptors related to cavity formation, the 10 most frequent cover about 50% of all occurrences of this descriptor type (42 out of 87). The 12 most frequent descriptors related to hydrogen bonding in liquids appear in the QSPR models for solvent scales in 54 cases out of 92. In both cases, each of the most frequent descriptors appears at least 3 times in the QSPR models. In the case of descriptors associated with electrostatic interactions in liquid media, the 11 most frequent descriptors cover more than 50% of all occurrences (86 out of 164) of these descriptors in the QSPR models for solvent scales. Similarly, the majority of occurrences of the descriptors associated with dispersion interactions (58 out of 82) are represented by just 7 descriptors. In the last 2 cases, each of these descriptors occurs at least 4 times in the QSPR models.

Therefore, because of the high intercorrelation between similar descriptors, a much smaller number of individual molecular descriptors should be adequate to describe most of the variance in the solvent scales. The above classification also enables a better interpretation of the QSPR models related to the molecular properties or reactivity in condensed media.

4.2. PCA Treatment Applied to Various Matrices. A reduced matrix 774 (solvents) × 100 (solvent scales), taking into account only those solvent scales with more than 20 experimental data points, was selected for multivariate statistical analysis. The filled data matrix (774 × 100) with experimental and predicted values is given as Supporting Information (SM 1). This matrix was further divided into several submatrices on which the principal component analysis was performed. All the matrices were standardized and subjected to PCA using STATISTICA v. 6.0 software.[51]

For the sake of comparison with our previous PCA treatment of the solvent scales,[1] we initially selected a matrix with the same size of 40 (solvents) × 40 (solvent scales). As seen from Table 4, the present results lead to three principal components that account for 80.5% of the variance. This “improvement” (the previously reported three PCs accounted for 70.2% of the variance) can be easily explained by the refined QSPR models now utilized, and this encouraged us to extend the treatment. Indeed, in the initial reduced matrix (774 × 100), the first five principal components accounted for 75% of the total variance.

The robustness of the PCA model obtained was validated by the development of models for restricted submatrices of data and their comparison with the results obtained for the full data set. The following criteria were used for various divisions of the original matrix into submatrices: (i) varying the number of solvents, but keeping the number of solvent scales constant (set 1), (ii) keeping the number of solvents constant, but varying the number of solvent scales (sets 2−5). Set 1 was built starting from the selected matrix of 40 (solvents) × 40 (solvent scales) by adding first the solvents that possess experimental values measured for most of the solvents scales. The last solvents added were those with fewer experimental values reported. Sets 2−5 were obtained either by comparison of the number of experimental datapoints in the scales considered or by the type of experimental measurement used in the development of the scale as follows:

- Set 2. Scales with $N \geq 50$ experimental datapoints.
- Set 3. Scales with $N < 50$ experimental datapoints.
- Set 4. Scales developed on the basis of spectroscopic measurements.
- Set 5. Scales developed on the basis of thermodynamic data, including data on chemical reactivity and solvation energies.

The percentages of variance and total variance for these different submatrices are listed in Tables 6 (set 1 of submatrices) and 7 (sets 2−5 of submatrices). The number of solvents was constant (703) for each of sets 2−5, because for 71 out of the 774 solvents, the predicted values fell outside the 20% error range values of experimental data and were excluded. These 71 solvents, which include some of the fluorinated compounds, specifically perfluoro compounds, and several phosphorus- and sulfur-containing solvents, are given in SM 5.

4.2.1. PCA Analysis of the Set 1 Submatrices. As seen from Table 6, for submatrices with the same number of solvent scales but different number of solvents, the total variance for the first three PCs varies from 68.2% to 58.96%. This demonstrates that the total variance described by the first 3 components is rather stable as the number of solvents involved changes from 100 to nearly 774. Consequently, the matrix with 703 diverse solvents and 100 solvent scales was selected for further general discussion. In this case, the first principal component accounts for 33.93% of the total variance and second and third principal components are associated with 15.85% and 11.48% of the total variance, respectively. The first 5 principal components account cumulatively for 75.44% of the variances, and the first 3 principal components and hence the essential properties of the solvents are associated cumulatively with 61.27% of the total variance.

4.2.2. PCA Analysis of the Submatrices of Sets 2−5. The results listed in Table 7 show that, in the case of the submatrices selected according to the number of experimental data points available for a given solvent scale, there is no significant difference in the total variance for the first three PCs. This finding demonstrates that using a QSPR with theoretical molecular descriptors gives consistent predictions for the missing solvent scale data. Even when most of the data (>90%) in the PCA matrix are predicted, the principal components determining the solvent scales remain practically the same.

TABLE 6: Variances Covered by up to Five Components for Seven Submatrices with the Same Number of Solvent Scales

matrices	% of variance					cumulative variance				
set 1	PC1	PC2	PC3	PC4	PC5	PC1	PC2	PC3	PC4	PC5
40 × 40	46.54	19.72	14.25	5.13	3.40	46.54	66.25	80.51	85.63	89.04
100 × 100	35.67	19.61	12.93	7.01	5.09	35.67	55.28	68.20	75.21	80.31
200 × 100	37.06	17.91	11.77	7.08	5.58	37.06	54.97	66.73	73.81	79.39
250 × 100	35.50	16.92	12.64	6.52	6.19	35.50	52.42	65.06	71.58	77.77
500 × 100	34.33	17.15	11.28	7.26	6.11	34.33	51.48	62.76	70.01	76.13
703 × 100	33.93	15.85	11.48	7.92	6.25	33.93	49.78	61.27	69.18	75.44
774 × 100	29.64	16.94	12.38	10.30	5.91	29.64	46.58	58.96	69.25	75.17

TABLE 7: Variances Covered by the Five Components for Various Submatrices with the Same Number of Solvents

matrices		% of variance					total variance				
		PC1	PC2	PC3	PC4	PC5	PC1	PC2	PC3	PC4	PC5
set 2	703 × 52 ($N^a \geq 50$)	36.67	19.65	11.89	8.49	4.33	36.67	56.31	68.20	76.69	81.02
set 3	703 × 48 ($N^a < 50$)	34.81	13.37	11.84	7.81	6.19	34.81	48.18	60.03	67.84	74.03
set 4	703 × 50 (SS^b)	40.88	16.83	10.55	7.57	4.60	40.88	57.72	68.26	75.83	80.44
set 5	703 × 22 (TS^c)	23.73	17.86	15.09	10.71	6.98	23.73	41.60	56.68	67.39	74.37

[a] *N*: number of experimental datapoints used for the development of the solvent scale. [b] SS: solvent scales defined using spectroscopic measurements. [c] TS: solvent scales defined using thermodynamic measurements (see SM 1).

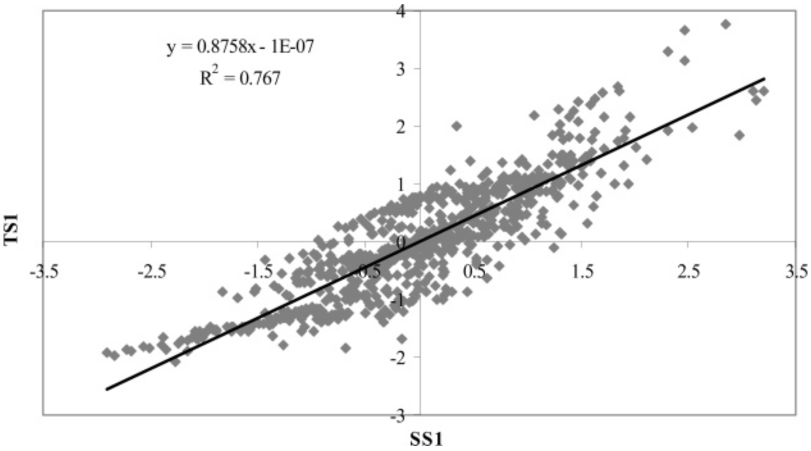


Figure 3. First principal component score (SS1) of (703 × 50) SS matrix vs first principal component score (TS1) of (703 × 22) TS matrix.

TABLE 8: Correlations between the First Three Principal Components Scores for the Sets 4 and 5

scores	R^2	scores	R^2	scores	R^2
SS1 vs TS1	0.7670	SS2 vs TS1	0.1365	SS3 vs TS1	0.0114
SS1 vs TS2	0.0417	SS2 vs TS2	0.3587	SS3 vs TS2	0.2755
SS1 vs TS3	0.0819	SS2 vs TS3	0.1319	SS3 vs TS3	0.3048

As already discussed, the solvent effect scales can also be divided into distinct classes on the basis of the experimental techniques used for their measurement. From a theoretical standpoint, there should be a significant difference between the scales obtained by using different experimental procedures. Explicitly or implicitly, various solvent scales reflect the variation of the solvation energy of individual solutes in different solvents. Traditionally, the solvation energy of a solute molecule E_{solv} in a given solvent can be presented as follows:[47,48]

$$E_{\text{solv}} = E_{\text{cav}} + E_{\text{disp}} + E_{\text{elst}} + E_{\text{H-bond}} \tag{6}$$

where each term corresponds to a certain type of intermolecular interaction in the condensed media. Thus, E_{cav} denotes the energy of the cavity formation for the solute in the solvent, E_{disp} is the dispersion energy, and E_{elst} is the electrostatic energy of interaction of the solute with the surrounding solvent molecules. The term $E_{\text{H-bond}}$ accounts for the energy of hydrogen bond formation between the solute and solvent molecules. In the case of thermodynamic measurements (equilibrium constants of chemical reactions, solvation energies of test solutes in different solvents, etc.), all the above-listed terms should contribute fully to the given solvent scale. Each of these terms should also be included in the scales determined by using spectroscopic measurements. For instance, the size of the molecule changes (increases) as a result of the Franck−Condon excitation of the solute molecule, and thus, the difference in the solute cavity formation energy in the ground and excited states, respectively, will be reflected in a spectroscopic solvent scale. However, the spectroscopic excitation process is practically instantaneous, and as a result, the position and the orientation of the solvent molecules around the chromophoric solute will not change. This means that only short relaxation time (electron redistribution) processes contribute to the various terms of solvation (eq 6).

Consequently, it would be of great theoretical interest to examine the possible difference in the principal components obtained from the PCA treatment including only a given type (spectroscopic or thermodynamic) of scale. Thus, 50 scales were classified as “spectroscopic” (denoted as SS) and 22 scales as “thermodynamic” (denoted as TS). Some scales have been developed as a combination of data obtained by using different experimental techniques, and these were kept separate from both of these sets. Results for separate PCA treatments for spectroscopic (set 4) and thermodynamic scales (set 5) are also given in Table 7. Overall, the PC coverages in both cases look similar. Detailed examination of the scores (solvent characteristics) of

TABLE 9: Variation of the Scores of the Five Principal Components for Each Group of Solvents

				PC1		PC2		PC3		PC4		PC5	
class	subclass	no. of solvents	type of solvents	mean[a]	SD[b]	mean[a]	SD[b]	mean[a]	SD[b]	mean[a]	SD[b]	mean[a]	SD[b]
I		81	hydrocarbons	−1.69	0.53	0.24	0.59	−0.56	0.82	0.52	0.71	−0.12	0.91
	Ia	37	saturated hydrocarbons	−2.13	0.34	−0.32	0.15	−1.11	0.62	0.54	0.72	0.39	0.85
	Ib	44	unsaturated hydrocarbons	−1.31	0.32	0.71	0.37	−0.09	0.67	0.49	0.71	−0.56	0.70
II		80	halo hydrocarbons	−0.35	0.47	1.29	0.59	−0.33	0.76	0.44	0.75	−0.58	0.85
	IIa	47	chloro hydrocarbons	−0.42	0.43	1.47	0.53	−0.55	0.60	0.34	0.77	−0.40	0.70
	IIb	16	bromo hydrocarbons	−0.55	0.29	0.82	0.48	0.31	0.52	0.42	0.63	−1.10	0.58
	IIc	8	iodo hydrocarbons	−0.35	0.35	0.72	0.51	0.42	0.61	0.23	0.59	−1.60	0.35
	IId	4	fluoro hydrocarbons	0.45	0.42	1.61	0.17	−0.26	0.13	0.63	0.18	0.36	0.63
	IIe	5	mixed hydrocarbons	0.32	0.37	1.71	0.16	−1.60	0.35	1.56	0.54	0.23	1.25
III		58	saturated, unsaturated, and cyclic ethers	−0.52	0.49	0.07	0.60	−0.35	0.70	−0.23	0.64	0.28	0.59
IV		67	esters and polyesters	0.28	0.67	0.66	0.51	−0.18	0.83	−0.13	0.88	1.17	0.59
V		84	aldehydes, ketones, and amides	0.25	0.61	−0.01	0.64	0.51	0.81	−0.78	0.71	0.51	0.54
	Va	13	aldehydes	0.50	0.43	0.28	0.45	−0.07	0.63	−1.11	0.54	0.40	0.40
	Vb	51	ketones	−0.01	0.45	0.01	0.74	0.53	0.70	−0.54	0.71	0.73	0.42
	Vc	20	amides	0.77	0.68	−0.23	0.38	0.81	0.99	−1.15	0.56	0.03	0.58
VI		36	nitriles and nitro compounds	0.43	0.74	0.68	0.83	−0.04	0.89	−0.89	0.79	0.66	0.70
	VIa	26	nitriles	0.19	0.61	0.34	0.70	0.08	0.94	−0.97	0.68	0.51	0.74
	VIb	10	nitro compounds	1.05	0.70	1.55	0.39	−0.34	0.68	−0.67	1.02	1.07	0.36
VII		125	hydroxylic compounds	0.95	0.77	−1.05	0.81	−0.43	1.04	1.09	0.66	−0.08	0.68
	VIIa	58	monohydric alcohols	0.49	0.59	−1.42	0.70	−0.66	0.71	0.71	0.47	0.12	0.59
	VIIb	32	phenols	0.85	0.47	−0.30	0.45	0.71	0.57	1.46	0.23	−0.66	0.42
	VIIc	13	di- and trihydroxy alcohols	1.74	0.42	−1.59	0.42	−1.09	0.58	1.35	0.71	0.21	0.67
	VIId	8	mixed alcohols	1.93	0.39	−1.25	1.06	−0.40	0.71	1.29	1.07	−0.72	0.56
	VIIe	14	organic acids, water	1.74	0.75	−0.61	0.63	−1.50	1.27	1.45	0.89	0.50	0.46
VIII		100	amines and pyridines	0.11	0.77	−0.53	0.90	0.61	0.83	−0.52	0.73	−0.83	0.60
	VIIIa	73	amines	0.00	0.80	−0.81	0.83	0.47	0.83	−0.56	0.74	−0.78	0.62
	VIIIb	27	pyridines	0.41	0.57	0.23	0.63	0.99	0.73	−0.44	0.73	−0.96	0.53
IX		49	sulfuro compounds	0.00	0.84	−0.09	0.71	0.69	0.81	−0.46	0.70	−0.83	1.17
	IXa	16	thiols	−0.40	0.62	−0.52	0.44	0.35	0.52	−0.20	0.58	−1.25	0.53
	IXb	13	sulfides	−0.60	0.54	−0.14	0.69	0.43	0.72	−0.39	0.51	−1.46	0.50
	IXc	13	sulfoxides	0.86	0.51	0.48	0.74	1.31	0.93	−0.87	0.75	0.62	1.29
	IXd	7	thio compounds	0.44	0.77	−0.09	0.56	0.83	0.69	−0.41	0.95	−1.41	0.46
X		12	phosphorus compounds	0.54	0.70	−0.29	0.75	1.52	1.33	−0.65	1.32	1.60	1.51
XI		11	miscellaneous	0.37	0.81	−0.16	0.70	0.36	1.32	−0.94	0.89	0.87	0.76

[a] Mean value of loadings for each group. [b] Standard deviation of loadings for each group.

the first 3 components from 2 treatments demonstrated that there is a significant correlation ($R^2 = 0.767$) between the first principal component scores from the treatment of the data matrices for the spectroscopic (set 4) and thermodynamic (set 5) solvent scales (see Figure 3). However, for the remaining correlations between PCs scores, the R^2 values vary from 0.0114 (SS3 vs TS1) to 0.3587 (SS2 vs TS2) (see Table 8).

This is an important result demonstrating that, in addition to others, one single strong interaction contributing to the solvation energy in different solvents is important in both the spectroscopic and thermodynamic measurements. Moreover, analysis of the correlation between the scores of the first PC of the complete matrix (703 × 100) with those of the two submatrices obtained on the basis of the type of the measurements shows very good fits of $R^2 = 0.9882$ and $R^2 = 0.8256$ for the scores of the first PC for the spectroscopic (703 × 50) and thermodynamic solvent scales (703 × 22), respectively (see SM 6).

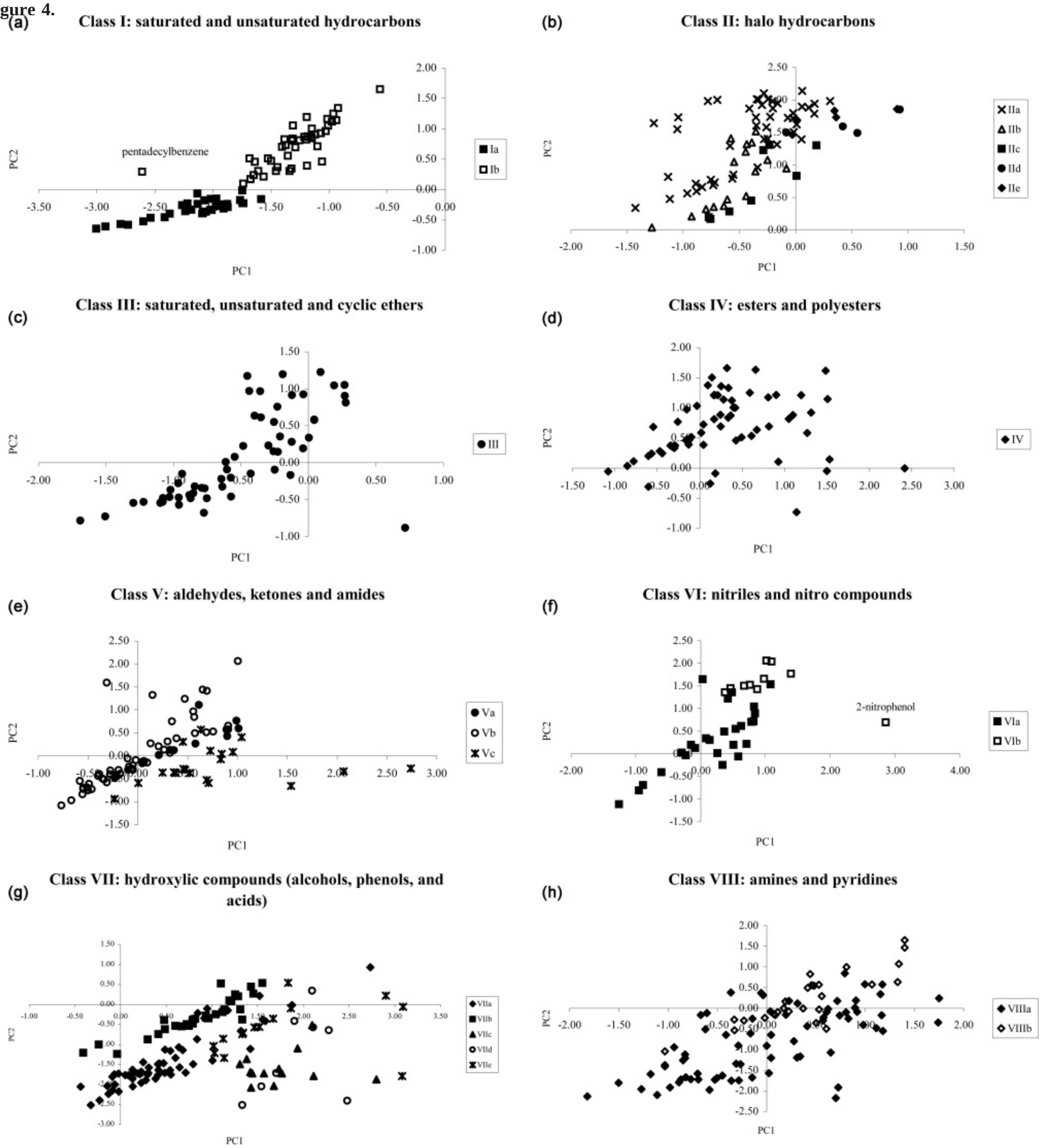
4.2.3. Analysis of Scores of (703 × 100) Matrix—Classification of Solvents. The scores of the first 5 principal components for the 703 solvents are given as Supporting Information (SM 7). Initially, these 703 solvents were classified according to the presence of various functional groups into 11 classes of compounds. Table 9 shows the subdivision of 7 of these classes into subclasses.

The analysis of the variation of the scores of the first 3 principal components indicates that the observed clusters support the initial division of the solvents. Figure 4a−k shows the plot of the scores of the second component versus the scores of the first component for each of these 11 classes.

Class I, hydrocarbons (Figure 4a), has a very large negative score extending to −3.01 for the first component and a relatively high value for the second component (1.65). The saturated hydrocarbons are clustered in the left bottom quadrant, having negative values for both principal components: from −3.01 to −1.58 for PC1 and from −0.65 to −0.02 for PC2. The unsaturated hydrocarbons present negative values for PC1 (from −2.61 to −0.56) and positive values for PC2 (from 0.10 to 1.65). These clusters support our previous observation[37] that *n*-hexane and cyclohexane (the only hydrocarbons included) with large negative scores for the first and second component are distinct from other solvents.

Class II, halo hydrocarbons (Figure 4b), clustered from medium-large negative values (−1.43) to medium positive ones (0.93) for the first component, with moderate positive values (from 0.04 to 2.14) for the second component. As one can see from Figure 4b, almost all chloro hydrocarbons are clustered in the left upper quadrant and have the highest values for the second component. They are followed by the bromo hydrocarbons (from −1.28 to −0.08 for PC1 and from 0.04 to 1.52 for PC2) with medium values for the second component and by the iodo hydrocarbons, which are present in both left and right upper quadrants, having both negative and positive values for the first component (from −0.77 to 0.19). Except for fluorobenzene (−0.08 for PC1 and 1.50 for PC2), the other fluoro hydrocarbons are located in the right upper quadrant. They have small to moderate positive values for PC1 (0.42 to 0.93) and medium to large values for PC2 (1.49 to 1.85). Five organic compounds that contain various numbers of diverse halogens

Figure 4.


(a) Class I: saturated and unsaturated hydrocarbons

(b) Class II: halo hydrocarbons

(c) Class III: saturated, unsaturated and cyclic ethers

(d) Class IV: esters and polyesters

(e) Class V: aldehydes, ketones and amides

(f) Class VI: nitriles and nitro compounds

(g) Class VII: hydroxylic compounds (alcohols, phenols, and acids)

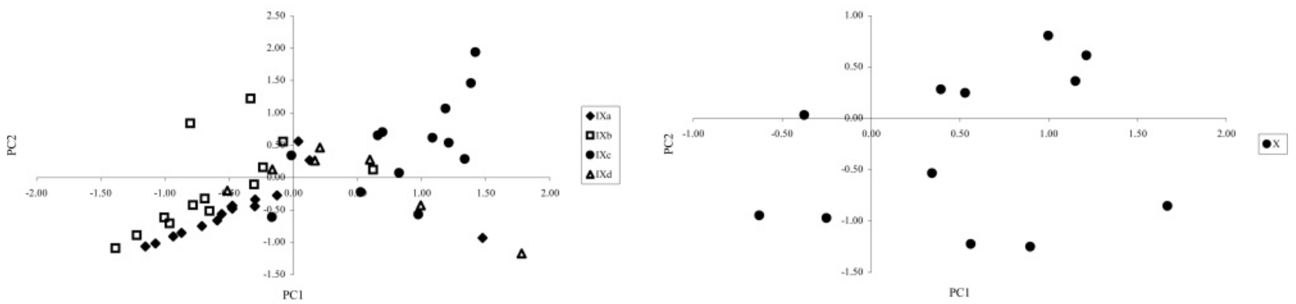
(h) Class VIII: amines and pyridines

in their molecule represent the fifth subclass, the mixed halo hydrocarbons. It was observed that their position in the cluster is influenced by the type and number of halogen atoms present in the molecule. For example, three compounds, 1,1,1-trichlorotrifluoroethane, 1,2,2-trichlorotrifluoroethane, and 1,2-dichlorohexafluorocyclobutane, are located together with the subclass IId (fluoro hydrocarbons) in the right upper quadrant (from 0.35 to 0.90 for PC1 and from 1.73 to 7.86 for PC2).

Class III, represented by saturated, unsaturated, and cyclic ethers (Figure 4c), is clustered below Class II: from −1.69 to 0.28 for the first component and from −0.78 to 1.23 for the second component. 5-Acetyl-5-methyl-1,3-dioxane is for some reason an outlier, by having a larger positive value for PC1 (0.72) and a bigger negative value for PC2 (−0.88). The third component values range from high-moderate negative ones (−1.88) to medium positive (1.56) for ethers and from (−2.01) to 1.75 for halo hydrocarbons.

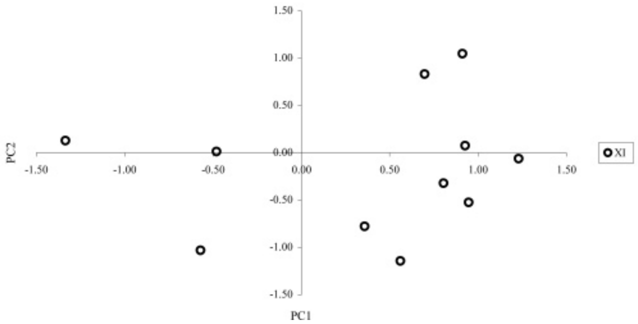
Class IV, esters and polyesters (Figure 4d), are grouped between −1.08 and 2.42 for PC1 and between −0.74 and 1.66 for PC2. Most of them show positive values for both PCs. The cluster has approximately the same shape as those formed by Class I, Class II, and Class III, and a preference of the esters for the right upper quadrant of the plot of PC2 vs PC1 is observed.

Class V is represented by aldehydes, ketones, and amides (Figure 4e). Most of the aldehydes (subclass Va) and ketones (subclass Vb) are uniformly spread alongside the line that passes through the origin from the left bottom quadrant (medium negative values for both principal components) to the right upper quadrant (medium positive values for PC1 and PC2). Subclass



(i) **Class IX: thiols, sulfides, sulfoxides and thio compounds**

(j) **Class X: phosphorus compounds**



(k) **Class XI: miscellaneous**

Figure 4. **(Continued)** (a−k) Scores of the second PC plotted vs scores of the first PC of (703 × 100) matrix for each group of solvents.

Vc (amides) is clustered from −0.23 to 2.74 for the first component and from −0.94 to 0.57 for the second component. Similarly, with the previous reported results,[37] formamide has a very large positive score (previously 2.02, now 2.74) for the first component. Acetone and acetophenone take positive values for both the first and second components, being located in the right upper quadrant, which supports the previous study.[37]

Class VI, nitriles and nitro hydrocarbons (Figure 4f), clustered by following the same trend as Class V: from medium negative values (left bottom quadrant) to medium positive values (right upper quadrant). The values of PC1 range from −1.26 to 1.08 for subclass VIa (nitriles) and from 0.37 to 2.85 for subclass VIb (nitro hydrocarbons). The highest positive values for PC2 achieved by nitriles is 1.65, but is 0.41 units less than the maximum value of PC2 for nitro hydrocarbons. 2-Nitrophenol seems to be an outlier for subclass VIb (2.85 for PC1 and 0.70 for PC1), probably due to the presence of hydroxy groups, which can form strong hydrogen bonds.

Class VII, hydroxylic compounds (Figure 4g), have large positive values extending to 3.09 for the first component and large negative values to −2.52 for the second component. The monohydric alcohols (subclass VIIa) have the larger negative value (−0.44) for PC1 in this class, followed by the phenols (subclass VIIb) with −0.41. Also, the monohydric alcohols have a larger negative value (−2.52) for the second component. However, the plot reveals that mono-, di-, and trialcohols are clustered in the right bottom quadrant, most of them having positive values for PC1 and negative values for PC2. Moreover, it seems that the organic acids are localized in the middle of a hypothetical triangle formed by phenols ("up corner"), monohydric alcohols ("left corner"), and di- and trihydroxy alcohols ("right corner"). Water has a very large positive value for PC1 (3.08) and a large negative value for PC2 (−1.79), similar to the previous results.[37]

Class VIII was divided into two subclasses: (i) VIIIa (amines) and (ii) VIIIb (pyridines). Figure 4h shows that this class is distributed between −1.82 and 1.75 units (for the principal component) and from −2.18 to 1.65 units (for the second component). Comparison between these two subclasses indicates that pyridines have high positive values for PC2 (1.65) and amines have large negative ones (−2.18). Also, most of the amines are spread between the left and right bottom quadrants, while the pyridines are situated almost completely in the right upper quadrant.

Class IX, the thiols (subclass IXa), sulfides (subclass IXb), sulfoxides (subclass IXc), and thio compounds (subclass IXd), have large negative to large positive scores (−1.39 to 1.78) for the first component and large negative to large positive scores (−1.17 to 1.94) for the second component. The thiols and sulfides are mostly clustered together in the left bottom quadrant, while the sulfoxides appear on the right upper quadrant. The thio compounds seem to be present in each of the quadrants but at lowest scores, either positive or negative, or both, for the principal and second components (see Figure 4i).

Class X, the phosphorus compounds (Figure 4j), which contain phosphates (3), phosphites (3), phosphane (1), phosphonate (1), phosphonic acids (2), phosphoric acid (1), and phosphorothioic triamide (1), and Class XI (Figure 4k), which includes a mix of diverse compounds, show a spread over all four quadrants. This is expected, since Class XI contains compounds with vastly different chemical functionalities.

4.2.4. Analysis of Loadings of (703 × 100) Matrix: Classification of Solvent Scales. The loadings of the 100 polarity scales in the first 5 principal components are given as Supporting Information (SM 8). Figure 5 shows the loadings of the second component plotted versus the loadings of the first component , and Figure 6 shows the loadings of the third component plotted against the loadings of the first component of the (703 × 100) matrix. The clustering of the scales in the space defined by the first 5 components suggests their classification into 5 distinct groups as listed in Table 10.

Group I consists of 31 solvent scales that are essentially defined (i) on the basis of the relative band intensities I/III for various fluorescence spectra and (ii) as expressions of the

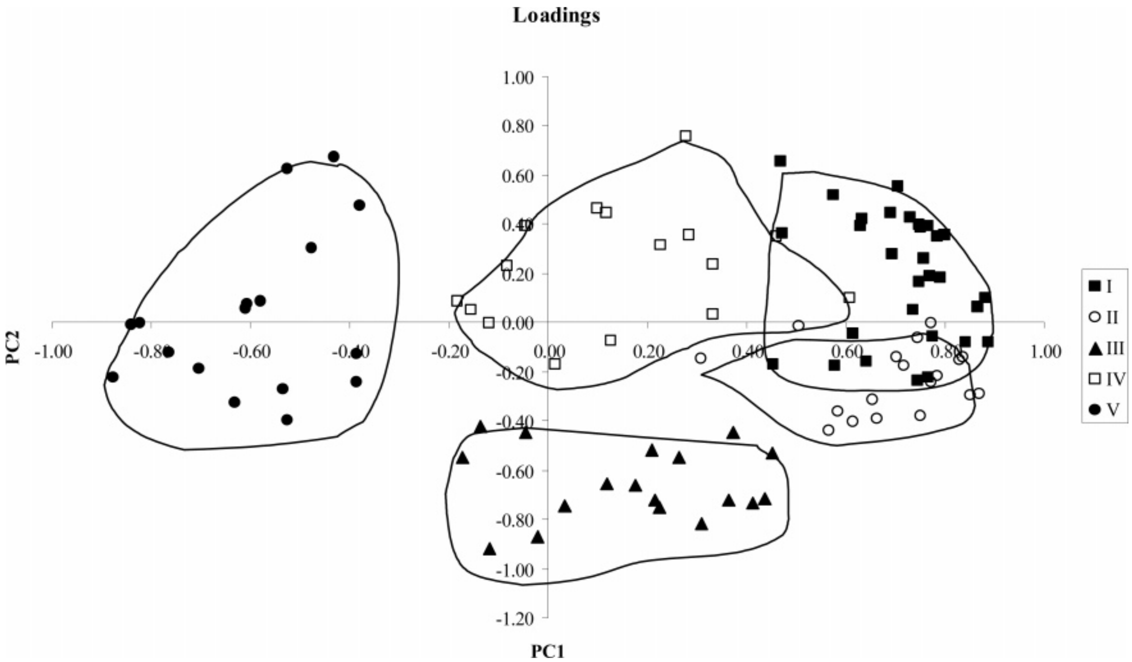


Figure 5. Loadings of the second PC plotted vs loadings of the first PC of (703 × 100) matrix.

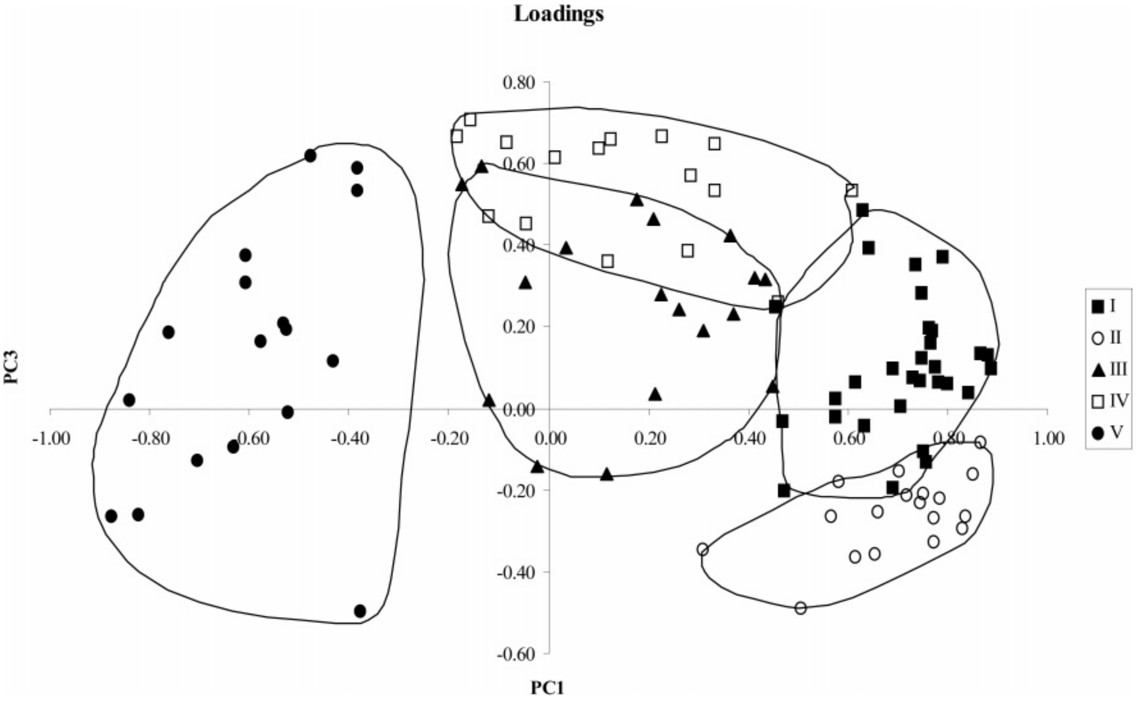


Figure 6. Loadings of the third PC plotted vs loadings of the first PC of (703 × 100) matrix.

dielectric constant. As one can see from Figures 5 and 6, these polarity scales have medium positive loadings (0.46 to 0.89) for the first component and comparable small loadings for the second and third components (−0.24 to 0.65 and −0.20 to 0.49, respectively).

Group II contains 18 polarity scales. According to the revealed clusters from Figures 5 and 6, these scales have small to medium positive loadings (0.31 to 0.87) for the first component and small negative loadings for the second (−0.44 to 0.00) and third (−0.49 to −0.08) components.

Eleven out of these eighteen scales have been reported in our previous study[37] as being clustered on the right bottom quadrant of the plot of loadings 1 vs loadings 2. As one can see from Figures 5 and 6, the same cluster is present at the same location: the right bottom quadrant characterized by positive loadings for the principal component and negative loadings for the second and third components, respectively. The majority of the scales from Group II are strongly influenced by the solvent stabilization of charge transfer in the UV/vis absorption spectral maxima of large and highly polarized conjugated systems.

Group III, consisting of 18 solvent scales, shows small negative to medium positive loadings for the first component (−0.17 to 0.45), large to medium negative loadings for the second component (−0.92 to −0.42), and small negative to moderate positive loadings for the third component (−0.16 to 0.59). These observations are in agreement with our previous paper[37] and support the conclusion that those scales, which reflect the solvent basicity, are clustering at small to moderate positive loadings for the first component (0.07 to 0.59).

TABLE 10: Variation of the Loadings of the Five Principal Components for Each Group of Scales

			PC1		PC2		PC3		PC4		PC5	
group	no. of scales	type of scales	mean[a]	SD[b]	mean[a]	SD[b]	mean[a]	SD[b]	mean[a]	SD[b]	mean[a]	SD[b]
I	31	relative band intensities I/III for various fluorescence spectra; expressions of dielectric constant	0.71	0.11	0.20	0.26	0.10	0.17	−0.10	0.22	0.09	0.18
II	18	solvent stabilization of charge transfer in the UV/vis absorption spectral maximum of large and highly polarized conjugated systems	0.70	0.14	−0.23	0.14	−0.26	0.09	0.23	0.26	−0.06	0.19
III	18	solvent basicity	0.17	0.21	−0.65	0.15	0.26	0.22	−0.13	0.15	−0.02	0.23
IV	16	solvent refractive index; related to equilibrium measurements	0.14	0.23	0.22	0.24	0.55	0.13	0.12	0.31	−0.07	0.37
V	17	miscellaneous	−0.59	0.16	0.02	0.32	0.12	0.31	0.06	0.31	0.13	0.24

[a] Mean value of loadings for each group. [b] Standard deviation of loadings for each group.

Group IV consists of 16 solvents scales that have small negative to medium positive loadings for the first and second components (−0.18 to 0.61 and −0.17 to 0.75, respectively) and small to moderate positive loadings for the third component (0.26 to 0.71). Most of these scales are related to equilibrium measurements and reflect the solvent refractive index (polarizability).

Group V contains 17 solvent scales related to strong dipole, surface tension of hard sphere liquids, or molecular volume or are defined on the basis of the shift of pure chloroform relative to that of chloroform in dilute solution and so forth. These scales are clustered on the left bottom and upper quadrants and process large negative to small negative loadings for the first component (−0.88 to −0.38), medium negative to medium positive loadings for the second component (−0.40 to 0.67), and moderate negative to positive loadings for the third component (−0.50 to 0.62). They deviate strongly from the other 4 major groups.

In conclusion, we observed that the use of an extended matrix of solvents and solvent scales (703 × 100) led us again to the already suggested idea[37] that complex solvation phenomena may involve nonlinear inter-relations between different mechanisms that cannot be “sensed” by the traditional linear/multilinear approach.

5. Conclusions

The wide diversity of solvent scales that have been developed reflects a limited number of possible physical interactions between dissolved molecules in condensed media. Unfortunately, it is extremely difficult to devise experimental procedure(s) that would measure the energy of any single type of intermolecular interaction in a liquid or in solution. Therefore, it is of substantial cognitive importance to estimate, even if only qualitatively, the number of possible different interaction mechanisms between molecules that result in observable solvent effects on spectra or chemical reactivity of compounds in solution. In principle, such an estimate can be obtained by applying the principal component analysis to an appropriately large solvent−solvent scale matrix. Unfortunately, large parts of such matrices based on experimental data are empty, as individual solvent scales are always determined for limited (and often not overlapping) sets of solvents.

The results of the present work demonstrate that, by employing large ranges of theoretical molecular descriptors, it is possible to obtain reliable and robust QSPR models for most solvent scales and, accordingly, to predict the missing values of solvent scales for the solvents that are experimentally unavailable. Consequently, the results of our present work can be used for the development of the full PCA matrix involving solvent scales and the determination of the dimensionality of intermolecular interactions in liquids and solutions.

In addition, the results of the present work differentiate the different classes of theoretical molecular descriptors in their ability to describe solvent effects. Importantly, the CPSA descriptors as well as other charge-distribution-related descriptors reflect the major part of the intermolecular interactions in solution. This is in accordance with the general concept that electrostatic interactions in many cases determine most solvent effects (ref 48 and references therein). However, for a number of solvent scales, the topological and even constitutional descriptors are statistically significant. In some cases, this may happen when limited sets of structurally similar solvents are used and these descriptors intercorrelate with the charge-distribution-related descriptors. On the other hand, both topological and constitutional descriptors depend on the size of the molecule and thus may be directly related to the effects accompanying the molecular cavity formation in the liquid. Interestingly, the geometrical descriptors are the least frequent in the QSPR equations for solvent scales.

In conclusion, while the results of the present work are applicable to the further analysis of solvent effects in condensed media, they also give a deeper insight into the possible interactions causing these effects.

For the first time, we have attempted to classify the theoretical molecular descriptors derived from the chemical structure alone according to their relevance to different intermolecular interactions in liquid media. In fact, almost all descriptors can be related to one of the generally accepted types of the intermolecular interactions (cavity formation, solvent electrostatic polarization, dispersion interaction, and specific hydrogen bond formation). Such classification enables significant insight into the physical interpretation of the QSPR of molecular properties in liquids and solutions, thus enhancing their cognitive value.

Both our QSPR models derived for individual solvent scales and the results of the PCA analysis should have large practical applicability, since they allow prediction of the solvent scale values for many solvents previously unmeasured. Almost all of the enormous number of existing correlations of molecular properties and chemical reactivity in solution with empirical solvent scales can now be extended with some confidence to a much wider selection of solvents.

Acknowledgment. The authors are very graceful to Ms. Hongfang Yang and Mr. Kaido Tämm for their contribution to the literature survey and the building of data matrix process, and to Dr. Charles Dennis Hall for critically reading the Manuscript. The Estonian Science Foundation grant no. 4548 is acknowledged for the partial support of this work.

Supporting Information Available: SM 1 − experimental values, matrices, and submatrices, SM 2 − list of solvent scales with references, SM 3 − list of solvents, SM 4 − classification of the descriptors, SM 5 − list of excluded solvents, SM 6 − plots of principal scores for matrices of different sizes, SM 7 − score values of the first five PCs, SM 8 − loading values of the first five PCs. This material is available free of charge via the Internet at http://pubs.acs.org.

References and Notes

(1) Katritzky, A. R.; Tamm, T.; Wang, Y.; Sild, S.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 684.

(2) Katritzky, A. R.; Fara, D. C.; Wang, H.; Tämm, K.; Karelson, M. *Chem. Rev.* **2004**, *104* (1), 175.

(3) Parker, A. J. *Chem. Rev.* **1969**, *69*, 1.

(4) Katritzky, A. R.; Fowler, F. W.; Rutherford, J. D. *J. Chem. Soc., B* **1971**, *3*, 460.

(5) Koppel, I. A.; Palm, V. A. *Advances in Linear Free Energy Relationships*; Chapman, N. B., Shorter, J., Eds.; Plenum Press: London, 1972; Chapter 5, pp 254−258.

(6) Koppel, I. A.; Paju, A. I. *Org. React.* **1974**, *39*, 121.

(7) Palm, N.; Palm, V. *Org. React.* **1997**, *31*, 141.

(8) Palm, V.; Palm, N. *Org. React.* **1993**, *28*, 125.

(9) Kamlet, M. J.; Abboud, J. L. M.; Abraham, M. H.; Taft, R. W. *J. Org. Chem.* **1983**, *48*, 2877.

(10) Famini, G. R.; Wilson, L. Y. *Theor. Comput. Chem*. **1994**, *1*, 213.

(11) Headley, A. D.; Starnes, S. D.; Wilson, L. Y.; Famini, G. R. *J. Org. Chem.* **1994**, *59*, 8040.

(12) Cronce, D. T.; Famini, G. R.; De Soto, J. A.; Wilson, L. Y. *J. Chem. Soc., Perkin Trans. 2* **1998**, *6*, 1293.

(13) Politzer, P.; Murray, J. S. *Fluid Phase Equilib.* **2001**, *185* (1−2), 129.

(14) Drago, R. S. *J. Chem. Soc., Perkin Trans. 2* **1992**, *10*, 1827.

(15) Joerg, S.; Drago, R. S.; Adams, J. *J. Chem. Soc., Perkin Trans. 2* **1997**, 2431.

(16) Mu, L.; Drago, R. S.; Richardson, D. E. *J. Chem. Soc., Perkin Trans. 2* **1998**, 159.

(17) Katritzky, A. R.; Mu, L.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 756.

(18) Marcus, Y. *Chem. Soc. Rev.* **1993**, *22*, 409.

(19) Catalan, J. *J. Org. Chem.* **1997**, *62*, 8231.

(20) Gramatica, P.; Navas, N.; Todeschini, R. *Trends Anal. Chem.* **1997**, *18*, 461.

(21) Meister, J.; Schwarz, W. H. E. *J. Phys. Chem.* **1994**, *98*, 8245.

(22) Heberger, K.; Lopata, A. *J. Org. Chem.* **1998**, *63*, 8646.

(23) Arnaudov, M.; Simeonov, V.; Yontchev, R.; Tsakovski, S. *Spectrosc. Lett*. **2001**, *34*, 93.

(24) Poole, S. K.; Poole, C. F. *J. Chromatogr., A* **1995,** *697*, 429.

(25) Castells, C. B.; Reta, M. R. *Anal. Chim. Acta* **2003**, *488*, 107.

(26) Poole, S. K.; Poole, C. F. *J. Chromatogr., A* **1995**, *697*, 415.

(27) Lochmüller, C. H.; Breiner, S. J.; Reese, C. E.; Koel, M. N. *Anal. Chem*. **1989**, *61*, 367.

(28) Dunn, W. J., III; Koehler, M. G.; Grigoras, S. *J. Med. Chem.* **1987**, *30*, 1121.

(29) Bohle, M.; Kollecker, W.; Martin, D. *Z. Chem.* **1977**, *17*, 161.

(30) Cramer, R., III. *J. Am. Chem. Soc.* **1980**, *102*, 1837.

(31) Cramer, R., III. *J. Am. Chem. Soc.* **1980**, *102*, 1849.

(32) Chastrette, M.; Rajzmann, M.; Channon, M.; Purcell, K. F. *J. Am. Chem. Soc.* **1985**, *107*, 1.

(33) Kuanar, M.; Mishra, B. K. *Ind. J. Chem., Sect. A* **1997**, *36*, 571.

(34) Maria, P. C.; Gal, J. F.; De Franceschi, J.; Fargin, E. *J. Am. Chem. Soc.* **1987**, *109*, 483.

(35) Zalewski, R. I.; Kokocinska, H.; Reichardt, C. *J. Phys. Org. Chem.* **1989**, *2*, 232.

(36) Pytela, O. *Collec. Czech. Chem. Commun.* **1990**, *55*, 644.

(37) Katritzky, A. R.; Tamm, T.; Wang, Y.; Slid, S.; Karelson, M. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 692.

(38) Krygowski, T. M.; Milcrazek, E.; Wrona, P. K. *J. Chem. Soc., Perkin Trans. 2* **1980**, *11*, 1563.

(39) Stewart, J. J. P. MOPAC 7.0; QCPE #455; http://qcpe.chem.indiana.edu/.

(40) http://www.codessa-pro.com.

(41) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. *Chem. Soc. Rev.* **1995**, *24*, 279.

(42) Katritzky, A. R.; Fara, D. C.; Petrukhin, R. O.; Tatham, D. B.; Maran, U.; Lomaka, A.; Karelson, M. *Curr. Top. Med. Chem.* **2002**, *2*, 1333.

(43) Reichardt C. *Solvents and Solvent Effects in Organic Chemistry*, 3rd ed.; VCH: New York, 2003.

(44) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience: New York, 2000.

(45) Stanton, D. T.; Jurs, P. C. *Anal. Chem*. **1990**, *62*, 2323.

(46) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. *Chem. Rev.* **1996**, *96*, 1027.

(47) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027.

(48) Karelson, M. *Adv. Quantum Chem.* **1997**, *28*, 141.

(49) Sinanoglu, O. *Intermolecular Forces in Gases and in Condensed Phases. Modern Quantum Chemistry − Istanbul Lectures*; Sinanoglu, O., Ed.; Academic Press: New York, 1965; Vol. II.

(50) Hansch, C.; Leo, A. *Exploring QSAR Fundamentals and Application in Chemistry and Biology*; American Chemical Society: Washington, DC, 1995.

(51) STATISTICA, release 6; Statsoft, Inc., 1997. http://www.statsoft.com/.