A General Treatment of Solubility. 3. Principal Component Analysis (PCA) of the Solubilities of Diverse Solutes in Diverse Solvents

Alan R. Katritzky,*[,#] Indrek Tulp,[#,⊥] Dan C. Fara,[#] Antonino Lauria,[#] Uko Maran,*[,⊥] and William E. Acree, Jr.[§]

Florida Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, Florida 32611-7200, Department of Chemistry, University of Tartu, 2 Jakobi Street, Tartu EE2400, Estonia, and Department of Chemistry, University of North Texas, Denton, Texas 76203-5070

Received December 18, 2004

A phenomenological study of solubility has been conducted using a combination of quantitative structure−property relationship (QSPR) and principal component analysis (PCA). A solubility database of 4540 experimental data points was used that utilized available experimental data into a matrix of 154 solvents times 397 solutes. Methodology in which QSPR and PCA are combined was developed to predict the missing values and to fill the data matrix. PCA on the resulting filled matrix, where solutes are observations and solvents are variables, shows 92.55% of coverage with three principal components. The corresponding transposed matrix, in which solvents are observations and solutes are variables, showed 62.96% of coverage with four principal components.

INTRODUCTION

The phenomenon of solubility has interested mankind for thousands of years. Ancient Greeks tried to perceive why wine is miscible with water while olive oil is not. Modern science proposes rational explanations of the macroscopic solubility phenomenon based on microscopic properties of the matter. Statistical mechanics rigorously links these two realms through a probabilistic treatment of particle ensembles. Further development of Kirkwood's approach[1] as applied to nondissociating fluids resulted in a variety of simulation techniques with the most popular of them being molecular dynamics and the Monte Carlo method.[2] Contemporaneous practical chemistry elaborated qualitative concepts such as solvent polarity/nonpolarity, dipolarity, and protophilicity, which achieved quantitative realization in the form of various solvent polarity scales. Most of these scales are derived from spectral and electrical properties of substances, chemical kinetics, and equilibrium data.[3] Recently we have commenced a general analysis of these issues using a combination of chemometric techniques such as multilinear regression analysis (MLR) and principal component analysis (PCA).[4]

Our earlier efforts to use PCA to explain physicochemical phenomena have been successful. In an application of the PCA method to a chemical problem, we treated aromaticity as a multidimensional entity and found that the magnetic and structural components are orthogonal.[5,6] In a more recent study[7] 40 different polarity scales were treated as a set of variables (descriptors) for 40 various solvents. The square 40 × 40 matrix formed was subjected to a diagonalization procedure that partitioned the solvents into five groups and the solvent scales into seven groups, according to the nature of the solvent and the physical meaning of the polarity scales.

Other research groups have also used the PCA for the classification of solvation-related physicochemical properties. Fawcett and Krygowski[8] investigated thermodynamic heats of solution. Cramer[9] studied the aqueous solubility of 114 chemicals along with other molecular properties. Chastrette et al.[10] performed a PCA analysis of 83 solvents with respect to six empirical solvent scales and the semiempirically derived highest occupied and lowest unoccupied molecular orbital energies of each solvent.

All the above investigations expressed solvent properties in terms of different optical and chemical reactivity features (solvent scales) that are, generally speaking, not directly related to solute−solvent interactions. A highly relevant scale for a particular solvent would simply be free energies of solvation of numerous substances by this solvent or, more generally, mutual equilibrium solubilities of different substances in each other. To the best of our knowledge nobody ever tried to address the general problem of solubility from this point of view. In the framework of the current series of papers we treat the general problem of solubility as a problem of dimensionality, as a problem of structure−solubility relationships, and last as a problem of solvent/solute classification. Two recorded principal component analyses of solubility dealt with a rather limited number of experimental only data, and no solvent/solute classification was suggested. Dunn et al.[11] analyzed a 6 solvents × 50 solutes data matrix and found two principal components with the first one being highly correlated with the isotropic surface area. A recent study of gas−liquid partitioning coefficients carried out by

* Corresponding authors phone: (352)392-0554; fax: (352)392-9199; e-mail: katritzky@chem.ufl.edu; Web: http://ark.chem.ufl.edu (K.A.R.) and phone: (372)737-5254; fax: (372)737-5254; e-mail: uko@chem.ut.ee (M.U.).

[#] University of Florida.

[⊥] University of Tartu.

[§] University of North Texas.

10.1021/ci0496189 CCC: $30.25
Published on Web 06/07/2005

Reta et al.[12] over an 11 solutes × 67 solvents revealed two relevant principal factors too. Two of the solvents were selected as "test factors" because of their high correlations with most of the experimental data.

To address the general problem of solubility in a uniform and comprehensive manner, we have collected a huge body of experimental equilibrium solubility data and formed a matrix in which columns are solvents (variables of PCA) and rows are solutes (objects of PCA). A transposed matrix was also formed. Missing values of these matrices have been predicted from QSPR analyses. Applying the well-developed mathematical formalism of PCA, based on the matrix diagonalization procedure, we derived principal components (eigenvectors) of the matrices, which enable us to classify and group solvents and solutes regarding to their chemical nature.

The goal of the current work is to derive an intrinsic dimensionality of the general solubility phenomenon and to reveal important constitutional and structural factors responsible for the solvation behavior of chemical entities. Such analysis should be interesting and important both from the theoretical and the applied points of view, because most physiological and technological processes occur in solution, and solvents exert strong influences on the rates and even the outcome of these processes.[13]

SOLUBILITY VALUES AND COMPOUNDS IN THE DATA MATRIX

All solubility values are expressed in the logarithmic form of Ostwald solubility coefficients (*logL*).[14] The Ostwald coefficent is defined as the volume of saturating gas absorbed by a volume of the pure solvent at the same temperature and pressure of the measurements. The pure gaseous solute serves as the reference state for the calculated values.[15] Rows and column are ordered according to the number of data points that they contain, so the densest area of the matrix is located in its upper left corner.

Our initial general solubility matrix of 145 solvents times 388 solutes as provided in the Supporting Information of a previous publication[14] has now been revised and extended. First we excluded 13 solutes as follows: mono- and diatomic compounds (helium, argon, neon, xenon, krypton, hydrogen, nitrogen, oxygen, nitrogen oxide, and carbon monoxide), tetramethyl tin, ferrocene, and fullerene. These exclusions are due to the limited number of molecular descriptors that can be calculated for the small (mono- and diatom) molecules, and because of semiempirical parametrization (AM1[16] is not parametrized for the tin and other heavy metals). For 22 solvents, fullerene represented the only solute with a measured solubility. Consequently, those 22 solvents were also eliminated. Additionally we included 31 new solvents and 22 new solutes (Supporting Information SI-A). All together we added 1030 experimental solubility values, which led to a total of 4540 data points in the revised matrix. This revised matrix (154 solvents and 397 solutes) is given as Supporting Information SI-B.

METHODOLOGY

Our general methodological plan involves a combination of the QSPR and PCA approaches for analysis of solubilities. First, the QSPR method was used as a tool to fill the gaps in the small solubility matrix. Then the small matrix was analyzed by PCA, and the QSPR models were developed for the principal components. Both approaches were combined to predict the solubility for the remaining points in the huge matrix. The filled solubility matrices were finally analyzed using PCA. A detailed description of the methodology for the development of the QSPRs has been given in our previous publications.[14,17]

Principal component analysis (PCA) is one of the best-known multivariate exploratory techniques extensively used in different areas of chemistry.[18−20] The PCA reveals internal relations between characteristics of a class of compounds (objects) and hence enables drastic reduction of the dimensionality of the original raw data. This reduction is achieved by transforming to a new set of variables, the principal components, which are uncorrelated (orthogonal to each other), and which are ordered so that the first few, with descending importance, retain most of the variation in the total set of original variables.

In PCA, the initial data matrix, $\boldsymbol{D}$, is represented as the inner product of two matrices (eq 1):

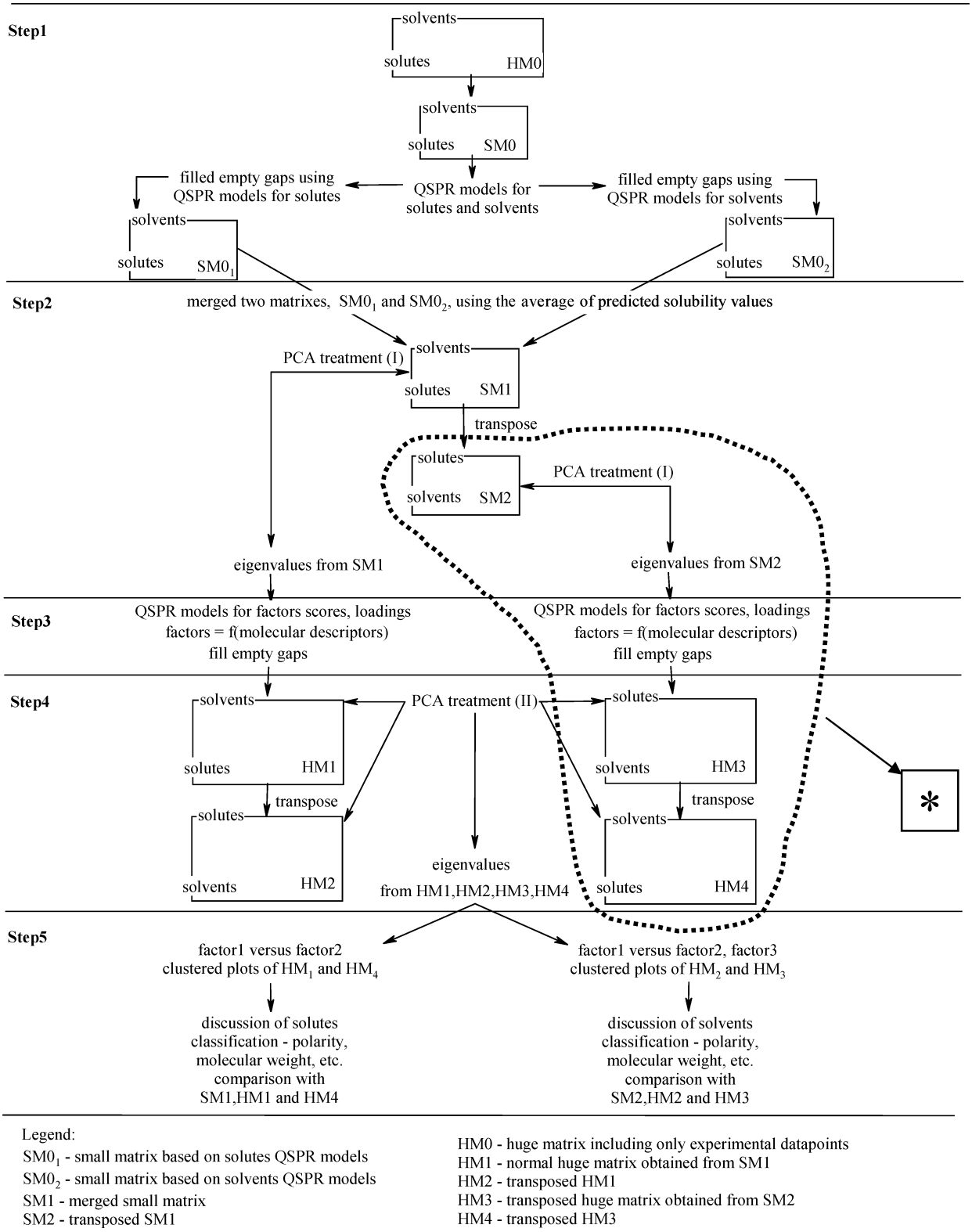
$$\boldsymbol{D} = \boldsymbol{RC} \tag{1}$$

The row matrix $\boldsymbol{R}$, named the score matrix, has the dimensionality $r \times n$, where r is the number of observations (i.e., compounds) in the initial data set, and n is the number of principal components (PC). The column matrix $\boldsymbol{C}$, named the loadings matrix, has the dimensionality $n \times c$, where c is the number of observable properties (variables) in the initial data set.

PCA can be highly useful for data classification and pattern recognition. In the two-dimensional plotting of a score vector against another score vector, compounds with similar properties as reflected in those two score vectors, are clustered. In the two-dimensional plotting of a loading vector against another loading vector, the initial statistical properties reflected in those two score loadings are clustered. The number of PCs (scores, loadings) existing in characteristic vector space can be equal to, or less than, the number of variables in the data set. The first principal component is defined as that giving the largest contribution to the respective PCA of linear relationship exhibited in the data. The second component may be considered as the second best linear combination of variables that accounts for the maximum possible of the residual variance after the effect of the first component is removed from the data. Subsequent components are defined similarly until practically all the variance in the data is exhausted. Principal component analysis for the current study was carried out with the SIMCA-P version 9.0 program package.[21]

Parts 1[14] and 2[17] of the project "General Treatment of Solubility" have been already published. Figure 1 describes the complete strategy for the project in five steps.

Step 1. The QSPR models for the solubility of single solutes in a range of solvents ("models for solvents")[14] and also the QSPR models for specified solutes in single solvents ("models for solutes")[17] using the general solubility matrix (HM0) have been further developed and revised in the present study. The newly developed QSPRs were used to predict the missing values of the *logL* for a so-called small



Legend:

SM0$_1$ - small matrix based on solutes QSPR models
SM0$_2$ - small matrix based on solvents QSPR models
SM1 - merged small matrix
SM2 - transposed SM1
HM0 - huge matrix including only experimental datapoints
HM1 - normal huge matrix obtained from SM1
HM2 - transposed HM1
HM3 - transposed huge matrix obtained from SM2
HM4 - transposed HM3

Figure 1. General treatment of solubility. *Initially considered as another possible way to fill HM0, but abandoned, as explained in the text.

matrix, (SM0), with size 87 solvents × 91 solutes. The SM0 included only those solvents and solutes for which at least 15 experimental solubility values were available. Two small matrices were obtained by predicting *logL* values first horizontally (using the "QSPR models for solutes"), denoted as SM0$_1$, and second vertically (using the QSPR "models for solvents"), denoted as SM0$_2$.

Step 2. The SM0$_1$ and SM0$_2$ matrices were merged into one general small matrix (SM1) according to the following rules:

(i) The predicted values within the prediction range were selected. The prediction range is defined as ±15% of the distribution range of the experimental data points for each model considered.

(ii) If the predicted value was in range in both $SM0_1$ and $SM0_2$, we used a model-weighted average of the two values, as follows

$$\overline{logL} = \frac{k_X X + k_Y Y}{k_X + k_Y} \quad (2)$$

where X is the solubility value predicted using solute model (horizontally), Y is the solubility value predicted by the solvent model (vertically), and k is the appropriate statistical coefficient of the QSPR model calculated by eq 3

$$k = \frac{n}{(1 - R^2)^2 N^2} \quad (3)$$

where n is the number of data points used in developing the QSPR model, R^2 is squared correlation coefficient for the model, and N is the number of descriptors in the model. We note that coefficient k as defined includes the model inflation factor (MIF) used by Peterangelo et al.[22] to evaluate quantitative structure−activity relationships.

$$\mathrm{MIF} = \frac{1}{(1 - R^2)} \quad (4)$$

(iii) If the solubility value predicted from the solute model was out of the range of experimental values, then the value predicted from the solvent model was taken and vice versa.

(iv) If the predicted value was out of range for both the solute and solvent models, then the solvents were ordered according to E_{T30} (the polarity scale with the largest number of data points), and an average value of the solubility was calculated by a "left-right three neighbor weighted average" method we developed as defined by eq 5

$$\overline{logL} = \frac{3(X_{1l} + X_{1r}) + 2(X_{2l} + X_{2r}) + X_{3l} + X_{3r}}{12} \quad (5)$$

where X_{il} is the left neighbor value and X_{ir} is the right neighbor value. The first neighbor relative to the gap was weighted with 3, the second one with 2, and the third one (most distant from missing value) was weighted with 1. For the cases for which experimental values of the E_{T30} scale were not available, the missing E_{T30} values were predicted by using a QSPR model developed for E_{T30} polarity scale by our group.[23] The filled matrix, SM1, was transposed, resulting into the respective matrix, SM2. At this stage, a first PCA treatment was applied to the normal (SM1) and transposed (SM2) small matrices. The data in SM1 and SM2 was normalized and centralized to give the data equal importance.

Step 3. The so-called "backward procedure" for the calculation of *logL* using the results of the PCA treatment (I) was applied (Figure 1). This procedure comprised the following pathway:

(i) QSPR models were developed for the significant scores (S), loadings (L), for the standard deviation (SD), and mean of the each column of the matrix ($\bar{M}$), using the set of theoretical molecular descriptors used in the step 1 for the development of the QSPRs. In this procedure all the above-mentioned factors and statistical parameters were defined as properties and loaded into Codessa Pro.[24] The corresponding regression equations were developed using the best multilinear regression (BMLR) algorithm.[25,26]

(ii) A general equation given below, as eq 6, was next used to calculate *logL* values:

$$logL = \sum_i (S_i \times L_i) SD + \bar{M} \quad (6)$$

Step 4. The empty gaps in the general solubility matrix HM0 were filled using either the revised solvent/solutes QSPR models where the prediction was in range or the appropriate form of the eq 6. The resultant matrix HM1 (397 solutes × 154 solvents) and its transposed matrix, HM2, were developed, and the PCA treatment was applied in both cases. To verify the accuracy of the prediction, the predicted solubility values of *logL* were compared statistically with the experimental solubility values (*see* Results and Discussion).

Step 5. The last phase of this "General Treatment of Solubility" project will analyze and discuss plots of the factors scores and loadings in the frame of a general classification of solutes and solvents for all matrices. A direct comparison of these plots and an interpretation of the physical meaning of the principal components will also be attempted. Conclusions from step 5 and their discussion will be reported in subsequent publications.

RESULTS AND DISCUSSION

Revision of the QSPR Models (Step 1). Significant revisions in the previously reported QSPR models[14,17] were made for the following reasons: (i) new experimental solubility data became available since our last QSPR modeling, and this lead to the extension of the general solubility matrix to 397 solutes × 154 solvents, as described above; (ii) there were mistakes in the 3D structural representation (used in ref 14) of nine of the compounds as detailed in the Supporting Information (SI-C); (iii) 16 experimental *logL* values were corrected in the general matrix (see Supporting Information SI-D); (iv) new features implemented into Codessa Pro were utilized: e.g. the BMLR algorithm[25−27] for the development of the QSPR models have been added and the pool of calculated descriptors has been increased with up to 40 hydrogen-bonding descriptors.

The QSPR were reconstructed for 87 solvents (vertical series) and 91 solutes (horizontal series). For this, a common descriptor pool for the solvents and solutes series was developed. The initial descriptor pool calculated by Codessa Pro consisted of 1101 theoretical molecular descriptors. In depth analysis of this pool lead to a set of rules that eliminated descriptors "inappropriate and irrelevant" for the current modeling task. Descriptors were eliminated as follows:

(i) 662 descriptors that are related to the specific atoms are eliminated because not all compounds from the data set include them. Examples are numbers of atoms, energy partitioning terms, etc. The number of carbon atoms and relative number of carbon atoms descriptors were retained because they were applicable for most of the data series. Only seven compounds out of 434 in the matrix do not contain carbon atom.

(ii) 38 charge distribution related descriptors that were derived from quantum-chemical calculations because Mul-

Table 1. Comparison of Performance of New and Previous QSPR Models

no. of data points		QSPR models with 2−5 parameters								squared correlation coefficient, R^2				average no. of descriptors	
		2		3		4		5 or more		range		average			
prev	now	prev	now	prev	now	prev	now	prev	now	prev	now	prev	now	prev	now
									Solvents						
3307	4183	3	22	22	47	18	13	26	5	0.837−0.998	0.868−0.998	0.955	0.961	4	3
									Solutes						
2409	3424	1	1	21	28	18	21	40	41	0.604−0.996	0.676−0.994	0.908	0.920	4	4

liken charge distribution scheme together with AM1 parametrization does not give correct estimates, in particular for halogens. Therefore the charge distribution related descriptors calculated according to Zefirov's approach (on the basis of electronegativities) were used instead.

(iii) 81 H-bonding descriptors based on quantum-chemical calculations and that have Zefirov analogues;

(iv) 81 reactivity indexes that also relate to specific atoms;

(v) 3 Kier shape indexes of various order because they cannot be calculated for molecules as H_2S, H_2O, NH_3 and CH_4.

(vi) 9 moments of inertia descriptors because they have excessively high values for small 3 atom structures;

(vii) and additionally 12 constitutional descriptors (number of multiple bonds, number of rings, etc.) and 3 normal vibration mode descriptors which were deemed not sufficiently relevant.

All together 889 descriptors were excluded. The final common descriptor pool thus consists of 212 whole molecule descriptors: 8 constitutional, 91 electrostatic, 12 geometrical, 29 quantum-chemical, 35 thermodynamic, and 37 topological descriptors. A detailed list of the descriptors used in this study is given in Supporting Information SI-E.

During the development of models, of the total of 4540 experimental data points 52 were designated as major outliers and were not used in development of QSPR models.

The solvent series had 15 outliers: *haloperidol* (142) from the series of toluene (30), 1-propanol (45), 2-propanol (46); *piroxicam* (144) from the series of dichloroethane (35), ethanol (44), 1-pentanol (51), 1-octanol (59), acetone (76), acetophenone (91), diethyl ether (104), and acetic acid (108); *benzoic acid* (118) from the nitrobenzene (120) series; *4-nitroaniline* (121) and *4-nitro-N,N-dimethylaniline* (147) from the series of nitromethane (122); and *phenanthrene* (10) from the carbon disulfide (161) series. The series of solutes had the remaining 37 outliers: *water* (116) from the following series 7, 24, 25, 29, 30, 31, 32, 34, 35, 42, 58, 65, 66, 92, 93, 119, 121, 134, 135, 142, 143, 164, 181, 185, 186, 188, 204, 288, 339, and 343 (see Supporting Information SI-F); *1-octanol* (59) from the *trans*-stilbene (2) series; *chloroform* (92) and *m*-cresol (260) from the 2-butanone (70) series; *2-butanol* (48) from the series of benzoic acid (118) and 4-nitroaniline (121); and *2-methyl-1-propanol* (49) and *1-hexanol* (55) from the series of 4-nitroaniline (121). Inclusion of these points would create data series without a normal distribution with the danger of leading to QSPRs that could give false estimates due to chance correlations.

A general comparison between the new and previously reported[14,17] QSPR models is given in Table 1. As one can see, the new models are statistically slightly better. Correlation coefficients are on average more than 0.01 units better in both cases. Also less descriptors are generally involved in models.

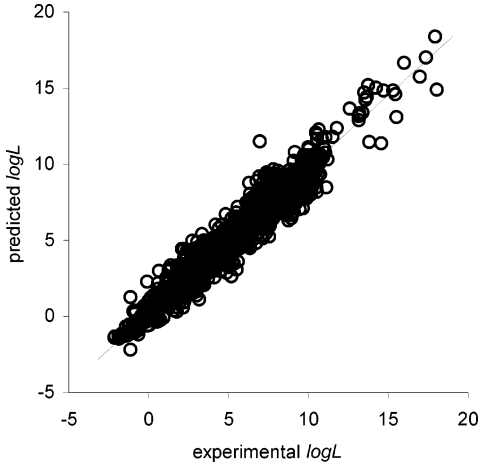


Figure 2. Predicted vs experimental solubility values for 87 solvent series: $y = 0.9559x + 0.1878$; $R^2 = 0.9574$ (4167 points).

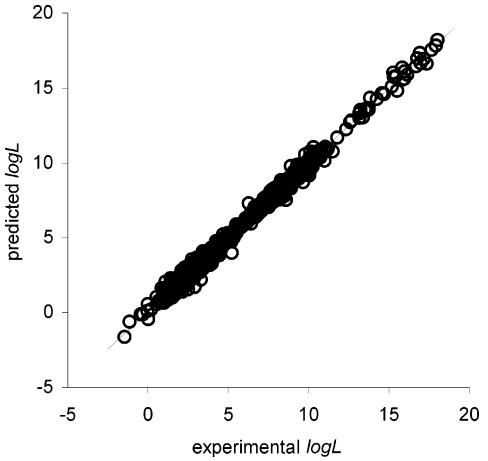


Figure 3. Predicted vs experimental solubility values for 91 solute series: $y = 0.9958x + 0.0203$; $R^2 = 0.9958$ (3394 points).

The predicted Ostwald solubility values are all plotted versus the corresponding experimental values for solvents and solutes in Figures 2 and 3, respectively. The squared correlation coefficient, $R^2 = 0.996$, shows a higher quality of prediction for solutes by comparison with the models of solvents ($R^2 = 0.957$). A comparison between those squared correlation coefficients (Figure 2 for solvents models and Figure 3 for solutes models) and the calculated average values by taking into account all QSPR models (see Supporting Information SI-G for solvents and SI-H for solutes models) shows similarity with the series for solvents ($R^2 = 0.957$ and 0.961) and a slight difference for the series of solutes (0.996 and 0.920, respectively). An explanation

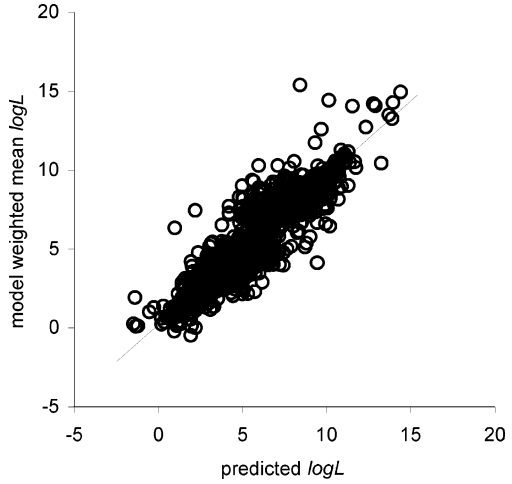


Figure 4. Model weighted mean vs predicted solubility values for 87 solvent series: $y = 0.9456x + 0.1832$; $R^2 = 0.8865$ (3011 points).

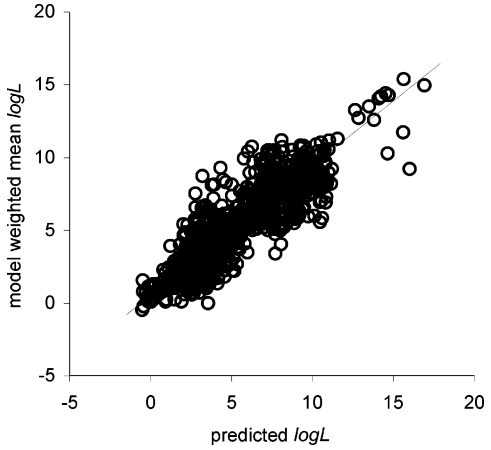


Figure 5. Model weighted mean vs predicted solubility values for 91 solutes series: $y = 0.8899x + 0.5237$; $R^2 = 0.8558$ (3011 points).

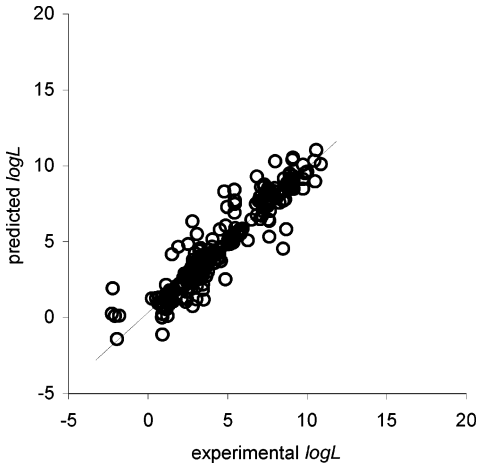


Figure 6. External validation set: predicted vs experimental solubility values: $y = 0.9532x + 0.3159$; $R^2 = 0.8815$ (289 points).

can be that the ranges of experimental values (i.e. ranges where predictions are most accurate) for the solutes are smaller in comparison with those for the solvents. The average range (ΔRange in Suppporting Information SI-I) for the solute series is 2.1 solubility units and for the solvent series 8.1 (Supporting Information SI-J). Obviously, due to the narrower range, the QSPRs for solute series has much smaller standard deviations (s^2). The average s^2 for solutes is 0.032 and for solvents 0.2 (see Supporting Information SI-I and SI-J). Although Figure 2 shows very good agreement between predicted and experimental values, it includes 1 outlier. The deviating point is dimethyl-4-nitrophenyl thiophosphate (536) predicted with the QSPR model developed for chloroform (92) solute.

The missing *logL* values in the small matrix SM0 were predicted (interpolated) (i) horizontally using QSPR models for solutes resulting in the matrix $SM0_1$ and (ii) vertically using QSPR models for solvents resulting in the matrix $SM0_2$ (see Figure 1).

Filling the Small Matrix − SM1 (Step 2). The matrices $SM0_1$ and $SM0_2$ were merged into SM1 according to the designated rules (see Methodology section, step 2). The final filled SM1 matrix, with dimensions of 87 solvents and 91 solutes, consists of the following data points: (i) 3074 experimental (38.8%), (ii) 3011 model-weighted average (eq 2) of the two models (38.0%), (iii) 1134 interpolated horizontally (14.3%), (iv) 482 interpolated vertically (6.1%), and (v) 216 polarity scale (E_{T30}) ordered and calculated by "left-right three neighbor weighted average" (eq 5) of solubility values (2.7%).

Figures 4 and 5 provide comparisons between the model-weighted averages and predicted values of solvent/solute series. The squared correlation coefficient for the predictions from QSPRs of solvents series is 0.886 (see Figure 4), and the squared correlation coefficient for the predictions of solutes series is 0.856 (see Figure 5). As it can be seen, the difference between the squared correlation coefficients is quite small. The R^2 was expected to be better in the case of the solvents. According to the eq 2 for the model-weighted average, the defined model statistical coefficient (k) is very highly affected by the correlation coefficient (R^2) and by the number of descriptors present in model (N) and therefore the mean supposed to be closer to the predicted values that are obtained from more significant model. In general, the models for solvents show higher quality of statistical characteristics (see Supporting Information SI-I and SI-J). The models for solvents have averaged squared correlation coefficient ($\overline{R^2}$) 0.961, while for the solutes $\overline{R^2}$ is 0.920. In the solvent models, the average number of descriptors ($\bar{N}$) is 3.0, while in the models for solutes it is 4.4. In fact, the linear correlations equations show clearly that solvent values are closer to the mean values, the slope is almost 1, and the intercept is very small. Figure 4 shows that the values for solvents are normally distributed. At the same time the solutes have a narrow range between 4.5 and 7 logarithmic units of solubility (Figure 5). During the preparation of the present manuscript, 289 additional experimental solubility values were collected, and these have been used as an external validation data set for SM1. In Figure 6 these experimental values are plotted versus respective estimated values with the corresponding $R^2 = 0.882$. The estimated values for the validation set of 289 compounds are acceptable. The 79 estimated values differ from the experimental values between 0.5 and 1.5 *logL* units, 27 estimated values differ more than 1.5 *logL* units. Seventeen out of those 27 values belong to the water series where seven estimates come directly from the QSPR model for water, and 10 of them

Table 2. 10 First Principal Components and the Percentage of the Variance Covered for the Normal Matrix (Solutes × Solvents)

PC	eigenvalue	% total	cumulative eigenvalue	cumulative %	Q^2 cumulative
			SM1		
1	80.679	92.735	80.679	92.735	0.924
2	2.172	2.496	82.851	95.231	0.946
3	0.692	0.795	83.543	96.026	0.952
4	0.440	0.505	83.982	96.532	0.952
5	0.330	0.380	84.313	96.911	0.950
6	0.297	0.341	84.609	97.252	0.951
7	0.238	0.274	84.848	97.526	0.949
8	0.215	0.247	85.062	97.773	0.948
9	0.202	0.232	85.264	98.005	0.948
10	0.149	0.171	85.413	98.176	0.945
			HM1		
1	134.513	87.346	134.513	87.346	0.871
2	4.884	3.171	139.397	90.517	0.899
3	3.125	2.030	142.522	92.547	0.917
4	2.521	1.637	145.044	94.184	0.932
5	1.139	0.739	146.182	94.923	0.937
6	0.789	0.513	146.972	95.436	0.940
7	0.658	0.427	147.629	95.863	0.943
8	0.520	0.338	148.149	96.201	0.943
9	0.465	0.302	148.614	96.502	0.943
10	0.411	0.267	149.026	96.770	0.944

Table 3. 10 First Principal Components and the Percentage of the Variance Covered for the Transposed Matrix (Solvents × Solutes)

PC	eigenvalue	% total	cumulative eigenvalue	cumulative %	Q^2 cumulative
			SM2		
1	23.432	26.933	23.432	26.933	0.233
2	20.358	23.400	43.789	50.332	0.390
3	8.716	10.019	52.505	60.351	0.486
4	4.870	5.598	57.376	65.949	0.526
5	3.101	3.565	60.477	69.514	0.526
6	2.403	2.763	62.880	72.276	0.517
7	2.175	2.499	65.055	74.776	0.507
8	1.870	2.149	66.925	76.925	0.503
9	1.706	1.961	68.631	78.886	0.494
10	1.512	1.738	70.142	80.624	0.478
			HM2		
1	41.744	27.106	41.744	27.106	0.205
2	29.333	19.047	71.077	46.154	0.396
3	13.848	8.992	84.924	55.146	0.470
4	12.040	7.818	96.964	62.964	0.526
5	7.108	4.616	104.072	67.580	0.557
6	6.150	3.993	110.222	71.573	0.596
7	4.063	2.638	114.285	74.211	0.607
8	3.702	2.404	117.987	76.615	0.606
9	3.307	2.147	121.294	78.762	0.621
10	2.506	1.628	123.800	80.390	0.632

are calculated using model-weighted average (eq 2). This clearly shows the complexity of the solubility process in water.

Big differences of the estimated values of more than 3 *logL* units were noticed for the solubilities of diphenyl sulfone (13), 4-hydroxybenzoic acid (102), 2-hydroxybenzoic acid (145), and acetylsalicylic acid (543) between the solvents 1-propanenitrile (84) and 1-butanenitrile (85). The solubility of any single solute would be expected to be quite similar in analogous solvents. All these estimated solubility values are calculated as model-weighted averages (eq 2). The differences arise from the specific solvent models. Thus the QSPR model for 1-butanenitrile gives low estimates for the diphenyl sulfone, and the 1-propanenitrile models gives low estimates for the three acids listed above. This is because the 1-propanenitrile and 1-butanenitrile QSPR models are based on a different solute set and thus involve different descriptors. The 1-propanenitrile model includes topological and electrostatic descriptors, and the 1-butanenitrile model includes topological, quantum-chemical, and thermodynamic descriptors (see Supporting Information SI-G).

The first principal component analysis (I, Figure 1) was performed on both the normal (SM1) and transposed (SM2) matrices. For the SM1 matrix (normal mode, 91 solutes × 87 solvents) the first three principal components cover 96.03% of the total variance. The contributions of the next PCs are negligible as indicated by the measure of the quality of prediction, Q^2, which shows no improvement with additional components. Information on the first 10 principal components is given in Table 2. The PCA of SM2 matrix (transposed mode, 87 solvents × 91 solutes) gives moderate results. In this case the two first principal components cover only 50.33% of the cumulative variance. The third and fourth PCs have the contribution: 10.02% and 5.60%, respectively. As one can see from Table 3, the contribution of the factors becomes very small starting with the fifth PC and also the Q^2 value shows no improvement after the fourth component.

Calculating of *logL* Using "Backward Procedure" (Step 3). The "backward procedure" for the calculation of *logL* using the results obtained from PCA treatment (I, Figure 1) as described in the previous chapter was applied. The matrices of the values for the scores and loadings of each considered principal component, and the standard deviations and means, are given as Supporting Information in SI-K and SI-L.

For the SM1 matrix, only the first, second, and third principal components are considered to contribute significantly to the solubility. Consequently, eight QSPR models (Table 4 and Supporting Information SI-K−SI-M) were built: 3 for the factors scores, S_i, 3 for the factors loadings, L_i, 1 for standard deviation, *SD*, and 1 for the mean, $\bar{M}$, where i is the number of principal components. In total 35 theoretical molecular descriptors were involved in the eight models derived and they belong to the following classes of descriptors: constitutional (1), geometrical (2), topological (7), electrostatic (15), thermodynamic (3), and quantum-chemical (7) (see Table 5).

The best QSPR model (see Table 4) was obtained for the first score (S_1) with $R^2 = 0.95$ and contains two descriptors. The respective model for the first loading (L_1) with $R^2 = 0.69$ includes six descriptors. The values of the first loading (L_1) vary in a very small range, from −0.99 to −0.87; only water has −0.75 (see Table 4 and Supporting Information SI-L), which shows that the influence of L_1 to the final results is almost negligible. According to the percentage of cumulative eigenvalues (see Table 2), the first principal component covers 92.7% of total variance of solubility, with the first score being by far the most important. The most significant descriptor (i.e. that with the highest *t*-test value) in the two-parameter model for S_1 (see Table 4) is the *gravitation index for all bonds* (D17), defined by eq 7

$$\mathrm{D17} = \sum_{i<j}^{N_b} \frac{m_i m_j}{r_{ij}^2} \tag{7}$$

where m_i and m_j are the atomic masses of atoms i and j, r_{ij}

Table 4. QSPR Models for the Scores and Loadings of First Three PCs and Standard Deviation and Mean Value for SM1 Matrix[a]

eq	QSPR models for principal components	N	n	R^2	R^2_{cv}	s^2	F	range
1	S_1=1.73(±0.0479)−0.00194(±0.0000502)D17−0.163(±0.0171)D7	91	2	0.952	0.948	0.0492	871	−3.86−1.55
2	S_2=2.06(±0.143)−1.70(±0.126)D30+0.0229 (±0.00209)D32−115(±10.9)D9−0.629(±0.118)D23	91	4	0.885	0.864	0.1201	166	−2.85−1.91
3	S_3=−2.19(±0.152)+0.198(±0.0122)D1+139(±13.9)D9−0.123(±0.0135)D3−0.432(±0.0556)D33−0.0590(±0.0133)D4+0.000661(±0.000269)D2	91	6	0.794	0.761	0.2209	53.9	−1.47−3.02
4	L_1=−0.605(±0.0356)−0.209(±0.0194)D20+0.000806(±0.000125)D11−0.107(±0.0175)D15+0.102 (±0.0251)D35−0.0155(±0.00399)D24−0.207 (±0.0679)D6	87	6	0.691	0.655	0.0004	29.9	−0.99 − −0.75
5	L_2=−1.23(±0.247)−0.134(±0.0150)D23−0.0106(±0.00147)D5−0.144(±0.0203)D30+0.0405(±0.00603)D28−6.92(±1.35)D9	87	5	0.902	0.883	0.0026	150	−0.54−0.29
6	L_3=−3.37(±0.542)+3.55(±0.564)D21+0.408(±0.0691) D12−0.0421(±0.00902)D34+0.0101(±0.00306)D31+ 0.00350(±0.00119)D18−0.0656(±0.0270)D30	87	6	0.674	0.592	0.0028	27.6	−0.14−0.28
7	SD=3.14(±0.142)−0.211(±0.0216)D16+33.2(±4.00) D13+0.0388(±0.00516)D27−0.148(±0.0231)D29+ 0.0648(±0.0136)D22−0.00778(±0.00164)D14−0.490(±0.113)D8+0.0411(±0.0109)D25−0.00914 (±0.00338)D3	87	9	0.763	0.701	0.0148	27.6	2.23−3.35
8	$\bar{M}$=8.75(±1.00)−61.0(±6.36)D9+2.94(±0.356)D12+ 0.00164(±0.000367)D26−0.0539(±0.0130)D19+ 0.00921(±0.00226)D10−0.0836(±0.0240)D28	87	6	0.746	0.646	0.0419	39.2	3.12−5.66

[a] Where N is the number of data points, n is the number of parameters in the model, R^2 and R^2_{cv} are the square of the correlation coefficient, and cross-validation correlation coefficient, respectively, s^2 represent the standard deviation, and F is the Fisher's criterion.

Table 5. Descriptors and Their Occurrence Involved in the Models Presented in Table 4

ID	descriptor name	occurrence
	Constitutional	
D1	number of single bonds	1
	Electrostatic	
D2	1X BETA polarizability (DIP)	1
D3	count of H-donors sites (Zefirov PC) (all)	2
D4	difference (Pos − Neg) in charged part of charged surface area (Zefirov's PC)	1
D5	DPSA3 difference in CPSAs (PPSA3-PNSA3) (Zefirov PC)	1
D6	FPSA2 fractional PPSA (PPSA-2/TMSA) (Zefirov PC)	1
D7	HA dependent HDCA-1 (Zefirov PC) (all)	1
D8	HA dependent HDCA-2 (Zefirov PC) (all)	1
D9	H-donors FCPSA (version 2)	4
D10	H-donors PSA (version 2)	1
D11	PNSA2 total charge weighted PNSA (Zefirov PC)	1
D12	polarity parameter (Zefirov)	2
D13	positively charged part of partial charged surface area (Zefirov's PC)	1
D14	RNCS relative negative charged SA (SAMNEG*RNCG) (Zefirov PC)	1
D15	RPCG relative positive charge (QMPOS/QTPLUS) (Zefirov PC)	1
D16	WNSA3 weighted PNSA (PNSA3*TMSA/1000) (Zefirov PC)	1
	Geometrical	
D17	gravitation index (all bonds)	1
D18	shadow plane YZ	1
	Quantum-Chemical	
D19	HOMO − LUMO energy gap	1
D20	max bonding contribution of one MO	1
D21	max SIGMA-SIGMA bond order	1
D22	tot dipole of the molecule	1
D23	tot hybridization comp. of the molecular dipole	2
D24	tot molecular 2-center exchange energy	1
D25	tot molecular 2-center resonance energy	1
	Thermodynamic	
D26	thermodynamic heat of formation of the molecule at 300 K	1
D27	thermodynamic heat of formation of the molecule at 300 K/natoms	1
D28	translational entropy (300 K)	2
	Topological	
D29	average complementary information content (order 0)	1
D30	average information content (order 0)	3
D31	bonding information content (order 2)	1
D32	information content (order 1)	1
D33	Kier&Hall index (order 3)	1
D34	structural information content (order 0)	1
D35	topographic electronic index (all bonds)	1

is the interatomic distance between the atoms i and j, and N_b is the number of chemical bonds in the molecule. The gravitational index reflects the effective mass distribution in the molecule and reflects intermolecular dispersion forces in the bulk liquid media (i.e. D17 accounts simultaneously both for the atomic masses and for their distribution within the molecular space). The second descriptor is the hydrogen-bonding donor charged surface area, HDCA(1), defined as

$$D7 = \sum_D S_D \quad (8)$$

where S_D is the solvent-accessible surface area of H-bonding donor H atoms, selected by threshold charge on hydrogen atom. The summation in eq 8 is performed over all simultaneously possible hydrogen bonding donor and acceptor pairs per solute molecule.[24,27] The combination of the two descriptors (D17 and D7) evidently represents adequately the intermolecular forces that influence the solubility process. The gravitation index (D17) is related to the dispersion and cavity-formation effects in liquids. The HDCA-1 (D7) is related to the hydrogen-bonding ability of compounds.

The QSPR model for the first score (S_1) is similar to a previously reported good two-parameter boiling point (T_b) model ($R^2 = 0.95$), where the gravitation index over all pairs of atoms (G_P) is taken into account as its cube root. The second descriptor is related to the hydrogen bonding (HDSA-2).[28] We also noticed similarities with our previously reported two-parameter QSPR models of vapor pressure, where the *gravitation index over all bonded atoms* (G_I) and the *hydrogen-bonding donor charged surface area* (HDCA-2) gives a linear correlation with $R^2 = 0.88$.[29] Two-parameter QSPR models for liquid viscosity (*log η*) also included the same two descriptors (G_I and HDCA-2) giving correlation coefficient $R^2 = 0.79$ using 337[30] and $R^2 = 0.81$ using 361 diverse organic molecules, respectively.[27,31]

For the transposed small matrix, SM2 (where the solvents are observations and the solutes are variables), the first four factors together cover only 65.95% of the information. The measure of the quality of prediction, Q^2, decreases after the fourth PC that indicates that introducing a higher number of PC into the PCA model is not appropriate (see Table 3). Consequently, the second path (II, Figure 1) for filling the huge matrix cannot be followed effectively and was abandoned.

Filling the HM (Step 4). It is now demonstrated that the PCA results obtained for the small matrix (SM1) can be successfully extrapolated to the general matrix of solubility (HM0). For this, the general matrix (HM0) was divided into four virtual sectors: sector I, the upper left corner of the matrix (91 solutes × 87 solvents), i.e. the small matrix (SM0); sector II, the upper right corner (91 solutes × 67 solvents); sector III, the lower left corner (306 solutes × 87 solvents); and sector IV, the lower right corner (306 solutes × 67 solvents). The gaps in the sectors I−IV of the HM0 are filled as follows:

(i) Sector I (SM0) was filled as described in previous sections (step 2).

(ii) In sector II the missing values of the solubility were predicted using the QSPR models for 91 solutes. From these predicted values just 618 (10.1%) were out of the range of the QSPR models.

(iii) Sector III was filled using 87 QSPR models for solvents. Here, 4207 (15.8%) of the predicted values were out of the range of the QSPR models.

(iv) The values out of range in sectors II and III were replaced using the so-called backward procedure as described in a previous section (step 3).

(v) Sector IV was completely filled using also the backward procedure. The predicted *logL* values were correlated with the respective experimental values to verify the correctness of prediction. The correlation between experimental and predicted values is with very good squared correlation coefficient ($R^2 = 0.997$) for sector II as plotted in Figure 7. The squared correlation coefficient is acceptable ($R^2 = 0.908$) for sector III (see Figure 8). Also the predicted values from the "backward procedure" and in the range

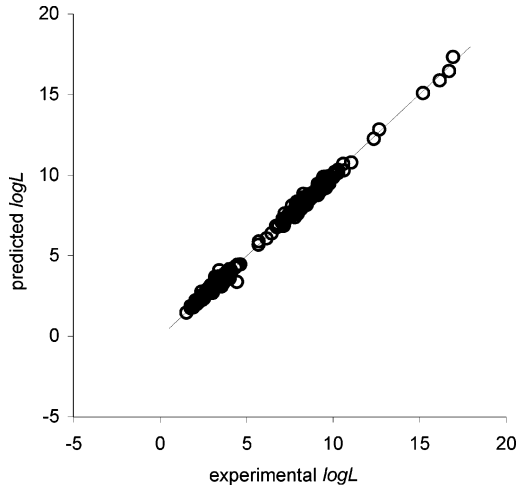


Figure 7. Predicted vs experimental solubility values for 91 solutes series in sector II: $y = 1.0041x - 0.0427$; $R^2 = 0.9968$ (357 points).

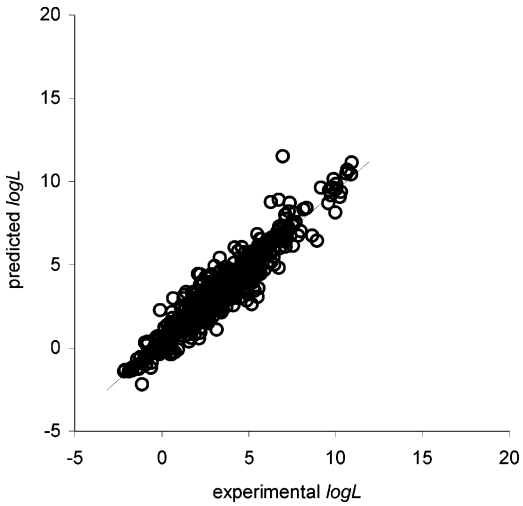


Figure 8. Predicted vs experimental solubility values for 87 solvents series in sector III: $y = 0.9092x + 0.3288$; $R^2 = 0.9082$ (1109 points).

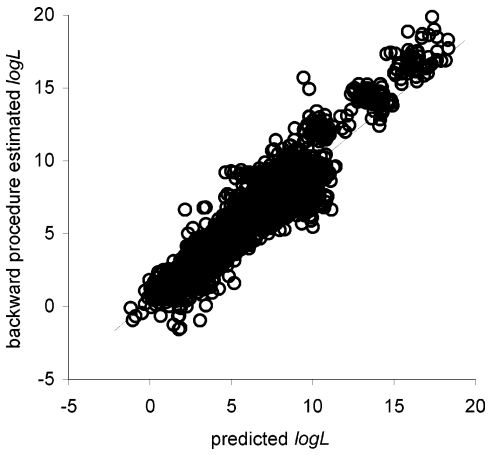


Figure 9. PCA backward estimated vs in the range predicted values for sector II: $y = 0.9235x + 0.3659$; $R^2 = 0.9002$ (5479 points).

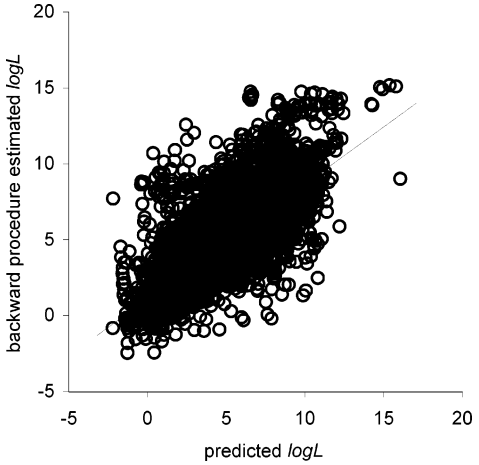


Figure 10. PCA backward estimated vs in the range predicted values for sector III: $y = 0.7557x + 1.0958$; $R^2 = 0.6414$ (22415 points).

predicted values from solute/solvent QSPR models were compared. For sector II this gives a very good correlation with $R^2 = 0.900$ and the corresponding plot is given in Figure 9. This comparison for sector III results in moderate correlation ($R^2 = 0.641$) and is plotted in Figure 10. The moderate result for sector III is due to the wide range of different solutes that also form the biggest part of the HM0.

Further the HM1 was transposed into the corresponding transposed matrix, HM2 (154 solvents × 397 solutes), and the second PCA treatment (II) was applied for both matrices. The corresponding eigenvalues of the 10 first principal components for these two matrices are given in Tables 2 and 3. As it can be expectedly seen, increasement of the matrix dimensionality will reduce the PCA revenue. The cumulative percentage of the eigenvalue drops slightly from 96.03% to 92.55% for normal matrix and from 65.95% to 62.96% in case of transposed matrix. Results of the PCA will be discussed in subsequent papers.

CONCLUSIONS

Developments in chemistry, technology, and drug design require extensive analysis of existing data and frequently estimations of values experimentally unavailable or unmeasured. Examples include the design and screening of real and virtual combinatorial libraries, the analysis of ADME/Tox (absorption, distribution, metabolism, elimination, and toxicity) profiles in drug discovery process, and the optimization of process control in (chemical) technology. We believe that the methodology and computational procedures designed in this work are of general interest and applicable to various large-scale quantitative structure−activity relationship/quantitative structure−property relationship and data mining problems in relevant areas.

The quantitative−structure property relationships and principal component analysis combined into one methodology have been used successfully to predict a large number of solubility values. A total of 4540 experimental data points was analyzed. The 178 QSPRs redeveloped for the densest area of the data matrix (91 solutes × 87 solvents) with covering a total of 3074 experimental values were successful in the prediction of the remaining 4843 solubility values. The PCA on the densest area of the data matrix, combined with the 178 QSPR equations, were further used successfully in filling the remaining of the 397 × 154 data matrix. The prediction procedure was validated with an external test set of 289 experimental data points.

The proposed methodology, with its combination of QSPRs and PCA, shows potential for the prediction of numerous solubility values. The three principal components from the fully filled data matrix where solutes are observations and solvents are variables describe 92.55% of the variability. The PCA on the corresponding transposed matrix results in only a moderate description of the variability. The principal components that describe the variability in the data matrix capture the contributions of the intermolecular dispersion forces, cavity formation forces, electrostatic forces, and hydrogen bonding to the solvation free energy.

ACKNOWLEDGMENT

We thank Prof. Mati Karelson, Dr. Alexander A. Oliferenko, Ms. Polina V. Oliferenko, Dr. Andre Lomaka, and Dr. Ruslan Petrukhin for their collaboration and for useful comments on this manuscript. This research was supported in part by the Estonian Science Foundation (Grant #5805).

Supporting Information Available: Excluded/added compounds (SI-A); the matrices (SI-B); corrected structures (SI-C); replaced solubility values (SI-D); final descriptor pool (SI-E); solute series where water was excluded (SI-F); solvent QSPR models (SI-G); solute QSPR models (SI-H); solute QSPR models statistical characteristics (SI-I); solvent QSPR models statistical characteristics (SI-J); first three scores (SI-K); first three loadings and descriptives (SI-L); QSPR models of principal components (SI-M); descriptors involved in 178 solvent/solute QSPR models (SI-N); added experimental solubility values of single solvents (SI-O); and added experimental solubility values of single solutes (SI-P). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys*. **1935**, 300−313.
(2) Sadus, R. J. *Molecular simulation of fluids: theory, algorithms, and object-orientation*; Elsevier: New York, 1999.
(3) Reichardt, C. *Solvents and solvent effects in organic chemistry*; Wiley-VCH: Weinheim, 2003.
(4) Katritzky, A. R.; Petrukhin, R.; Tatham, D.; Basak, S.; Benfenati, E.; Karelson, M.; Maran, U. Interpretation of Quantitative Structure−Property and −Activity Relationships. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 679−685.
(5) Katritzky, A. R.; Karelson, M.; Sild, S.; Krygowski, T. M.; Jug, K. Aromaticity as a quantitative concept. 7. Aromaticity reaffirmed as a multidimensional characteristics. *J. Org. Chem*. **1998**, *63*, 5228−5231.
(6) Katritzky, A. R.; Barczynski, P.; Musumarra, G.; Pisano, D.; Szafran, M. Aromaticity as a Quantitative Concept. 1. A Statistical Demonstration of the Orthogonality of "Classical" and "Magnetic" Aromaticity in Five- and Six-Membered Heterocycles *J. Am. Chem. Soc.* **1989**, *111*, 7−15.
(7) Katritzky, A. R.; Tamm, T.; Wang, Y.; Karelson, M. A Unified Treatment of Solvent Properties. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 692−698.
(8) Krygowsky, T. M.; Fawcett, W. R. A Characteristic Vector Analysis of Solvent Effects for Thermodymamic Data. *Can. J. Chem.* **1976**, *54*, 3283−3292.
(9) Cramer, R. D. BC(DEF) Parameters. 1. The intrinsic dimensionality of intermolecular interactions in the liquid state. *J. Am. Chem. Soc*. **1980**, *102*, 1837−1849.
(10) Chastrette, M.; Raizmann, M.; Chanon, M.; Purcell, K. F. Approach to a general classification of solvents using a multivariate statistical treatment of quantitative solvent parameters. *J. Am. Chem. Soc*. **1985**, *107*, 1−11.
(11) Dunn, W. J., III; Koehler, M. G.; Grigoras, S. The role of solvent-accessible surface area in determining partition coefficients. *J. Med. Chem*. **1987**, *30*, 1121−1126.

(12) Castells, C. B.; Reta, M. R. Study of gas−liquid partitioning of alkane solutes in several organic solvents by using principal analysis and linear solvation energy relationships. *Anal. Chim. Acta* **2003**, *488*, 107−122.
(13) Kolar, P.; Shen, J.-W.; Tsuboi, A.; Ishikawa, T. Solvent selection for pharmaceuticals. *Fluid Phase Equil.* **2002**, *194*−*197*, 771−782.
(14) Katritzky, A. R.; Oliferenko A. A.; Oliferenko P. V.; Petrukhin R.; Tatham, D. B.; Maran U.; Lomaka A.; Acree, W. E., Jr. A General Treatment of Solubility. 1. The QSPR Correlation of Solvatation Free Energies of Single Solutes in Series of Solvents. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1794−1805.
(15) Fogg, P. G. T.; Bligh, S.-w. A.; Derrick, M. E.; Yampol skii, Y. P.; Clever, H. L.; Skrzecz, A.; Young, C. L. IUPAC−NIST Solubility Data Series. 76. Solubility of Ethyne in Liquids. *J. Phys. Chem. Ref. Data* **2002**, *30*, 1693−1875.
(16) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart J. J. P. AM1: A New General Purpose Quantum Mechanical Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(17) Katritzky, A. R.; Olifenko A. A.; Olifenko P. V.; Petrukhin R.; Tatham, D. B.; Maran U.; Lomaka A.; Acree, W. E., Jr. A General Treatment of Solubility. 2. QSPR Prediction of Free Energies of Solvation of Specified Solutes in Ranges of Solvents. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1806−1814.
(18) Malinowski, E. R.; Howery, D. G. *Factor Analysis in Chemistry*; Wiley-Interscience: 1980.
(19) Strouf, O. *Chemical Pattern Recognition*; Wiley: New York, 1986.
(20) Meloun, M.; Militky, M.; Forina, M. *Chemometrics in Analytical Chemistry*; Ellis Horwood: New York, 1992.
(21) www.umetrics.com.
(22) Peterangelo, S. C.; Seybold, P. G. Synergistic Interactions among QSAR Descriptors. *Int. J. Quantum Chem.* **2004**, *96*, 1−9.
(23) Katritzky, A. R.; Fara, D. C.; Kuanar, M.; Hur, M.; Karelson, M. A Three-Dimensional Classification of Solvents Based on QSPR Unpublished work.
(24) www.codessa-pro.com.
(25) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure. *Chem. Soc. Rev.* **1995**, *24*, 279−287.
(26) Draper, N. R., Smith, H. *Applied Regression Analysis;* Wiley: New York, 1981.
(27) Katritzky, A. R.; Chen, K.; Wang, Y.; Karelson, M.; Lučić, B.; Trinajstić, N.; Suzuki, T.; Schüürmann, G. J. Prediction of liquid viscosity for organic compounds by a quantitative structure−property relationship. *J. Phys. Org. Chem.* **2000**, *13*, 80−86.
(28) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and Test Set of 9 Simple Inorganics. *J. Phys. Chem.* **1996**, *100*, 10400−10407.
(29) Katritzky, A. R.; Wang, Y.; Sild, S.; Tamm, T.; Karelson, M. QSPR Studies on Vapor Pressure, Aqueous Solubility, and the Prediction of Water−Air Partition Coefficients. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 720−725.
(30) Ivanciuc, O.; Ivanciuc, T.; Filip, P. A.; Cabrol-Bass, D. Estimation of the Liquid Viscosity of Organic Compounds with a Quantitative Structure−Property Model. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 515−524.
(31) Lučić, B.; Bašic, I.; Nadramija, D.; Miličević, A.; Trinajstić, N.; Suzuki, T.; Petrukhin, R.; Karelson, M.; Katritzky, A. R. Correlation of liquid viscosity with molecular structure for organic compounds using different variable selection methods. *ARKIVOC* **2002**, *4*, 45−59.

CI0496189