A General Treatment of Solubility 4. Description and Analysis of a PCA Model for Ostwald Solubility Coefficients

Indrek Tulp,[⊥] Dimitar A. Dobchev,[‡,§] Alan R. Katritzky,[‡] William Acree, Jr.,[||] and Uko Maran*[,⊥]

Institute of Chemistry, University of Tartu, 14A Ravila, Tartu 50411, Estonia, Center for Heterocyclic Compounds, Department of Chemistry, University of Florida, Gainesville, Florida, 32611, Department of Chemistry, Tallinn University of Technology, Akadeemia tee 15, Tallinn 12618, Estonia, and Department of Chemistry, University of North Texas, Denton, Texas 76203-5070

Received March 1, 2010

Principal component analysis (PCA) of a large data matrix (153 solvents × 396 solutes) for Ostwald solubility coefficients (log *L*) resulted in a two-component model covering 98.6% of the variability. Analysis of the principal components exposed the structural characteristics of solutes and solvents that codify interactions which determine the behavior of a chemical in the surrounding media. The pattern revealed by PCA analysis distinguishes solutes according to the molecular size, functional groups, and electrostatic interactions, such as polarity and hydrogen-bonding donor and acceptor properties.

INTRODUCTION

Solubility is a key property in almost all fields which are related to chemistry and is crucial in the production of new material and substances, assessing environmental risk for sustainability of environment and health, detecting drug-likeness, etc. Extensive studies of solute−solvent interactions and diverse theories unfolding those interactions have formed the basis for understanding solubility, as is comprehensively reviewed by Reichardt.[1] Despite more than a century of studies directed toward examining the relationships between chemical structure and solubility, the challenge of improved experimental detection, precise computational prediction, and detailed understanding of interactions between chemicals and the surrounding medium still remains.[2] For instance, a recent prediction of the intrinsic solubilities of 32 crystalline drug-like molecules in water using a data set of accurately determined solubilities of 100 compounds was challenged[3] and resulted in about 100 contributions.[4] Only a few of the top 10 successful results have been published,[5,6] which present simple and straightforward models and reveal problems in modeling of solubility as well.

The current study continues to analyze Ostwald solubility coefficients (log *L*; Here and throughout the text, a decadic logarithm is denoted by log) and is the fourth part of our series entitled "A General Treatment of Solubility". The first two parts focused on the theoretical foundations, data gathering, and multilinear quantitative structure−property relationship (QSPR) modeling of a series of solvents[7] and solutes.[8] The third part utilized the derived QSPR models and provided a systematic approach to predict missing data points using a combination of QSPR and principal component analysis (PCA) methods.[9] Using this combined approach, several regions of the data matrix were filled by predictions from QSPR models. Where simple QSPR predictions were not possible, the data gaps were filled by a combination of PCA and QSPR. The detailed scheme explaining each step of the analysis is thoroughly described in Part 3 of ref 9, Figure 1 therein. The first four steps have been completed in the previous three parts. The current manuscript is the last in the series and follows the final step (no. 5), which is the analysis and discussion of principal components (PCs) and their scores and loadings.

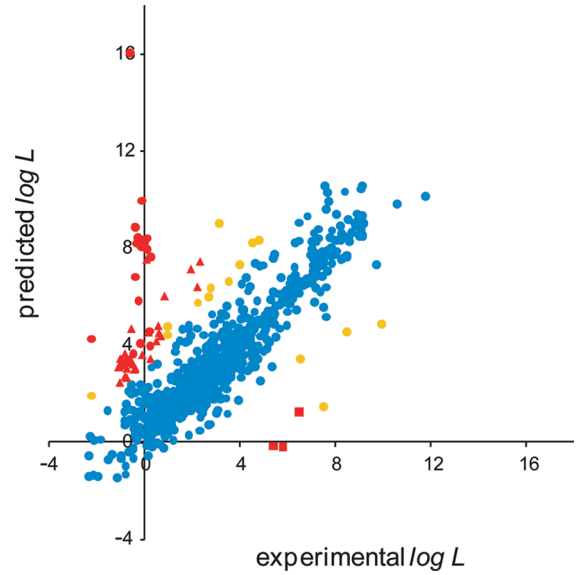

Figure 1. Validation of previously estimated values.

The definition of the Ostwald solubility coefficient (log *L*) expresses the distribution coefficient of a solute distributed between a liquid solvent and gas phase, and it is related to the solute's free energy of solvation according to eq 1:[7]

$$\Delta G_S = -2.3RT \log L = -2.3RT \log \left(\frac{c_l}{c_g}\right) \quad (1)$$

* Corresponding author. E-mail: uko.maran@ut.ee.
[⊥] University of Tartu.
[‡] University of Florida.
[§] Tallinn University of Technology.
[||] University of North Texas.

10.1021/ci1000828
Published on Web 07/01/2010

where c_l and c_g are the solute's concentrations in the liquid and gas phases, respectively. This relationship is valid for standard states of unit concentration in the gas phase and in solution, and the dependence is linear with respect to log L at a constant temperature.

$$\Delta G_S = \Delta G_{cavity} + \Delta G_{disp} + \Delta G_{el} + \Delta G_{HB} \quad (2)$$

The free energy of solvation is considered to consist of four main components (eq 2):[1,10,11] the cavity formation term (ΔG_{cavity}), dispersion interactions (ΔG_{disp}), free energy of electrostatic interactions (ΔG_{el}), and a term which takes into consideration the formation and reorganization of hydrogen bonds (ΔG_{HB}). The first two terms in eq 2 are related to the bulk characteristics of the solute, and together they are the major energy contributors to solvation free energy. This also holds for systems which are normally known to be very polar and strongly hydrogen bonded (HB).[12] Both terms (ΔG_{cavity}, ΔG_{disp}) can be regarded as characteristics of nonspecific interactions. The term for electrostatic interactions (ΔG_{el}) involves, in addition to the pure electrostatic Coulomb interactions, other interaction forces such as ion−dipole, strong dipole−dipole, and ion−pair formation, etc.[1,13,14] The HB forces are also electrostatic by nature.[15] Hence, it can be concluded that the last two terms comprise the electrostatic-specific interactions.

As discussed in our earlier work (ref 7, Figure 1 and discussion), the molecular descriptors closely reflect the terms of the free energy of solvation. The cavity formation term can be satisfactorily modeled with the use of topological and geometrical descriptors, semiempirically derived molecular polarizability, and entropy. Electrostatic and quantum chemical descriptors contribute significantly to both nonspecific and specific solvation either through atomic charges, charged surface areas, dipole moments, reactivity indices, or other similar structural parameters. Descriptors designed for HB include molecular surface areas that were confined by hydrogen-bond donor or acceptor sites as well as those that merely count such sites derived from atomic charge considerations. This provides a framework that enables construction of QSPR models for significant scores and loadings of the PCA model. Such an approach facilitates the discussion of the main structural characteristics influencing intermolecular interactions that determine the distribution of solutes between the solvent and its gas phases. Pattern analysis of score plots provides the second framework for the study and description of the PCA model. Description of the pattern according to chemical constitution enables validation of the solute and solvent classification. Such an analysis, based on a large and structurally wide-ranging data set, provides insight into important structural characteristics influencing the interactions and determining the solubility.

The present study consists of the following parts: (i) external validation of previous predictions with new experimental data; (ii) PCA model development and detailed outlier analysis, (iii) analysis of scores and loadings using QSPR models and molecular descriptors therein; and (iv) pattern analysis of common chemical spaces of the solutes in terms of chemical constitution.

DATA AND METHODS

Solubility Data. The data matrix used in this work was adapted from a previous publication in this series.[9] One duplicate, 1-nitropropane (compound ID - cID123), was found and removed. As a result, the matrix of the logarithm of Ostwald solubility coefficients (log L) consisting of 153 solvents × 396 solutes was formed. The total matrix now comprises 60 588 data points of which 4540 were experimental. The remaining data points were calculated according to the combined QSPR and PCA methodology as described in detail.[9] In addition, the following improvements were made: (i) 1285 new experimental data points were included into the data matrix replacing the previous estimated values[9] and (ii) 24 previous experimental values were replaced with new ones due to improved measurements (see Supporting Information for detailed description). After this update, the final data matrix comprises 5825 experimental log L values. The whole data set of compounds used in the work is rather diverse. For example, the data set includes 17 unique chemical classes defined by their functional groups (see Supporting Information, Table S1).

Principal Component Analysis. The dimensionality of the original data matrix was reduced by the multivariate PCA technique into a small number of orthogonal principal components (PC).[16,17] The data matrix was expanded as a sum of the principal components defined by scores and loadings:

$$\mathbf{D} = \mathbf{T} \cdot \mathbf{P} = \sum_{n=1}^{k} \mathbf{t}_n \mathbf{p}_n = \sum_{n=1}^{k} t_{i,n} p_{n,j} \quad (3)$$

In eq 3, **D** is the data matrix, **T** and **P** are the score and loading matrices, respectively, and $\mathbf{t}_n$ and $\mathbf{p}_n$ are the score and loading vectors for a given component which are expanded to their elements $t_{i,n}$ and $p_{n,j}$, respectively. The indices i and j correspond to observations (solutes) and variables (solvents), respectively, and n is the number of principal components. In the current study, the PCA implemented by the SIMCA-P software[18] was used to analyze the total data matrix of log L values. The obtained scores and loadings enable contribution analysis of the variance based on solutes and solvents, respectively. The graphical plots of the score and loading vectors also reveal relationships between the objects and variables. In our case, the score plots summarized a pattern among the solutes (observations), and the loading plots summarized a pattern for the solvents (variables). The loading plot also enables interpretation of the pattern seen in the score plot. Hence, the patterns of these two presentations aid in the analysis of regularities encoded by the chemical structure.

A vital issue for the PCA model is the identification of strong and moderate outliers, which could skew the model. Strong outliers can be traced in plots of PC scores, while moderate outliers can be found by inspecting the model residuals. Generally, the strong outliers tend to significantly shift (rotate) the PCA model toward them. An appropriate statistical method for identifying such outliers is Hotelling's T^2,[19] a generalization of the Student's t-statistic. T^2 is graphically presented as an ellipsoid of T^2 range on score plots and indicates deviations far from defined confidence intervals (95 or 99%). Strong outliers can also be spotted by the distance to the model X (DModX). Observations with a DModX over a critical value (D-Crit) are outliers to the PCA model. The probability of these observations belonging to the model (PModX) is less than 5%.[20]

PCA is a maximum variance projection method that is usually associated with a large number of variables. The data is generally preprocessed to provide all scales with equal weight, usually via the unit variance scaling method, where the data is standardized, centralized, and normalized using a sample standard deviation, variance, and mean. In our case, the nature of the experimental data did not require preprocessing of the log *L* values, because all data was measured at the same dimensional scale and the measurement error is nearly uniform. In addition, such preprocessing of the data could result in a loss of information[20−22] as well as decrease the sensitivity of the PCA toward some important characteristics of the chemical objects.

In a typical PCA investigation, the number of principal components sufficient for the model should be detected. There are several standard guidelines available[20,23] for detection of the number of optimal PCs. In the present analysis, the *scree* test was used, which allows one to plot the eigenvalues graphically with respect to the number of PCs. In principle, the *scree* test suggests that the optimum number of PCs is located at the point where the smooth decrease of eigenvalues appears to level off to the right of the plot.[22]

Multilinear Regression Analysis. The identification of the most significant and mutually orthogonal molecular descriptors related to the scores and loadings should indicate the structural features that determine solubility. For this purpose, multilinear regression (MLR) analysis can be employed to build QSPR models where scores and loadings are used as dependent variables. For the characterization of chemical structures of the solutes and solvents, more than a thousand molecular descriptors were calculated (for detailed descriptions, please refer to Parts 1−3)[7−9] using Codessa Pro.[24] Each score vector can be regarded as a solute property; similarly, each loading can be regarded as a solvent property. Elucidation of the most significant multilinear relationships for scores and loadings from the large number of descriptors requires a robust method. We chose the best multilinear regression (BMLR) approach,[25] which utilizes forward selection of the best few descriptors related to the dependent variable and has proven to be a reliable method for QSPR model development.[26−28] At the beginning of the BMLR procedure, the descriptors with missing values were removed. This was followed by construction of the best two-parameter regression and then the best three-parameter regression, etc., based on the statistical significance, orthogonality ($R^2 < 0.1$), and noncollinearity criteria ($R^2 < 0.6$) of the selected descriptors. The descriptor scales were normalized and centered automatically, with the final result given in natural scales. The final model has the best representation of the activity in the given descriptor pool within the given number of parameters. The quality of the models was assessed by the coefficient of determination (R^2), leave-one-out cross-validated coefficient of determination (R^2_{CV}), leave-many-out cross-validated coefficient of determination (R^2_{CVMO}), *Y* scrambled (10 000 randomization steps) coefficient of determination (R^2_{SCR}), Fisher's criterion (F), and the squared standard error of the regression (S^2).

RESULTS AND DISCUSSION

Validation of Predicted Ostwald Partition Coefficients. The 1285 new experimental data points (Supporting Information, SI-A) obtained during the preparation of this manuscript were used to test the previously calculated log *L* values.[9] This external validation shown graphically in Figure 1 had the correlation coefficient $R^2_{val} = 0.585$, which is reasonably significant considering the large number and diversity of data points.

This validation discloses several large outliers with prediction residuals higher than 3.05 log *L* units (that is, two times the standard deviation) and highlights data points outside of the 95% confidence region. The following groups of compounds lie outside of the domain of most organic solutes: (i) small inorganic gaseous solutes and (ii) phosphates (cID469−471) in water. Group (i) is overestimated, where SF_6 (cID194) has the biggest deviation (red circles), others are CF_4 (cID221), CO_2 (cID212), CS_2 (cID161), and SO_2 (cID198) (red triangles); group (ii) is underestimated (red boxes). Those groups of compounds were not well covered by the QSPR models used for the predictions because they had seven or fewer experimental log *L* data points and therefore, their log *L* values were not adequately estimated in Part 3.[9]

Within the domain of organic solutes, outliers underestimated by more than 3.05 units (colored in orange circles) consist of dimethyl sulfoxide (cID63) and acetylsalicylic acid (cID543) in water (cID116); dimethyl sulfoxide (cID63) in chloroform (cID92); and eicosane (cID313) in *N*,*N*-dimethylformamide (cID88). Those organic solutes overestimated by more than 3.05 units (colored in orange circles) include ammonia (cID321), 2,5-dimethylhexane (cID187), *trans*-stilbene (cID2), and 4-chlorobenzoic acid (cID340) in water (cID116); methyl iodide (cID203) and ethyl iodide (cID184) in aniline (cID90); difluorodichloromethane (cID530) in *N*,*N*-dimethylformamide (cID88); difluorodichloromethane (cID530) in *N*-methyl-2-pyrrolidone (cID405); and 2,2,2-trifluoroethanol (cID72) in 2,2,2-trifluoroethanol (cID72).

Excluding all 75 outliers indicated above, the squared correlation coefficient between the new 1210 experimental values and the previously estimated log *L* produced the highly significant $R^2_{val} = 0.881$. Therefore, considering the amount and diversity of the compounds in the external validation set, one can presume our previous predictions[7−9] are reliable. This demonstrates potential for the approach that encouraged analysis of the results and further exploration of the log *L* matrix.

Principal Component Model. In contrast with Part 3,[9] we did not use preprocessed (standardized, centralized, and normalized) data. The PCA model (M1) for the full raw data matrix (153 solvents × 396 solutes) led to surprisingly good results in comparison with a previously reported PCA model.[9] As can be seen from Table 1, the first PC alone covered 97.1% of the data variability. The addition of the second PC improved the model to include 97.9% of the total variance in terms of R^2X(cum). Further, each consecutive component provided less than 0.5% improvement. Following the *scree* test discussed above, the remaining components did not cover enough variance to be considered significant. Thus, only the first two PC's of model M1 were used in the next step. For completeness, the first 10 components and their eigenvalues are provided in Table 1. Figure S1 of the Supporting Information depicts a score plot for the first two scores of model M1.

The observations (solutes) of the PCA model were analyzed to eliminate outliers that can influence the analysis of the model. Strong and moderate outliers were eliminated

Table 1. First 10 Principal Components of PCA Models M1 and M2[a]

comp no.	eig	R^2X	R^2X(cum)	Q^2	Q^2(cum)	S^2X
			PCA model M1			
1	131.72	0.9708	0.971	0.9703	0.970	0.713
2	5.82	0.0079	0.979	0.2407	0.977	0.525
3	3.00	0.0041	0.983	0.1378	0.981	0.427
4	2.30	0.0031	0.986	0.1338	0.983	0.353
5	1.38	0.0021	0.988	0.0823	0.985	0.302
6	1.23	0.0017	0.990	0.0785	0.986	0.262
7	0.77	0.0010	0.991	0.0145	0.986	0.237
8	0.71	0.0010	0.992	0.0583	0.987	0.214
9	0.48	0.0006	0.992	0.0045	0.987	0.199
10	0.42	0.0006	0.993	0.0031	0.987	0.186
			PCA model M2			
1	129.81	0.9778	0.978	0.9774	0.977	0.400
2	8.14	0.0078	0.986	0.3189	0.985	0.262
3	2.94	0.0028	0.988	0.1613	0.987	0.214
4	1.76	0.0018	0.990	0.0521	0.988	0.182
5	1.68	0.0016	0.992	0.1267	0.989	0.154
6	1.18	0.0011	0.993	0.1018	0.990	0.134
7	0.78	0.0007	0.994	−0.0105	0.990	0.121
8	0.67	0.0006	0.994	0.0567	0.991	0.110
9	0.52	0.0005	0.995	0.0292	0.991	0.102
10	0.44	0.0004	0.995	0.0282	0.991	0.095

[a] Eig is eigenvalues; R^2X is fraction of sum of squares (SS) of the entire X explained by the current component; R^2X(cum) is cumulative SS of the entire X explained by all extracted components; Q^2 is fraction of the total variation of X that can be predicted by the current component; Q^2(cum) us cumulative Q^2 for all the *x*-variables for the extracted components; and S^2X is variance of the X matrix. For component number A, it is the residual variance of X after component A.

using raw residuals, such as distance to the model X (DModX) and Hotelling's T^2 statistics. Distance to the model X (Supporting Information, Figure S2) identifies both moderate and strong outliers. Moderate outliers according to the DModX (Supporting Information, Table S2) have values between D-Crit[2] (1.151) and two times the value of D-Crit[2]. Strong outliers are those that have more than twice the value of D-Crit[2] (Supporting Information, Table S2). The strong outliers were small gaseous compounds (SF_6 (cID194), CF_4 (cID221), and N_2O (cID370)), phosphates (cID469−471), and water (cID116). They were also the biggest outliers in the external validation (see previous section). According to DModX, a total of 67 strong and moderate outliers were removed from the final analysis.

For the PCA model M1, the T2Crit value within a 95% confidence interval is 6.05 within 2 PCs (Supporting Information, Figure S3). Thus, 22 compounds with a value higher than T2Crit were considered to be highly deviating points because they are too far away from the origin of the PCA model plane. The eliminated compounds include: (i) long aliphatic chains, such as eicosane (cID313), one ester (cID490), and long aliphatic alcohols, like 1-hexadecanol (cID311); (ii) bulky *para*-substituted benzenes like haloperidol (cID142); (iii) 3 phosphates (cID535−537); and (iv) water (cID116) (see Supporting Information, Table S3).

Seventy-six solutes were defined as outliers by these two criteria and eliminated. The remaining data matrix consisted of 320 solutes × 153 solvents. Next, a new PCA model (M2) based on the refined data was constructed (shown in Table 1), where PC_1 still describes 97.8% of the variability and PC_2 extends to 98.6%. PCA model M2 is used for further analysis.

Table 2. Highest Leverages of Observations for PCA Model M2

solute	M2.OlevX[2]	influence
n-decane	0.029	1.45%
indecane	0.029	1.45%
4-nitrophenol	0.024	1.2%
tert-butylcyclohexane	0.022	1.1%
Nonane	0.020	1.0%

Some moderate outliers remained in PCA model M2. Their influence on the model was assessed using leverages as an additional criterion. A leverage is defined as a measure of the influence of a point (observation) on the PCA model in X-space (OLevX). The sum of leverages is equal to the number of principal components in the model (i.e., the sum of leverages of a one-component model is 1, for a two-component model the total sum is 2, etc.) Observations with high leverage are in the periphery of a data set and significantly influence the PCA model. Table 2 shows the five biggest leverages along PCA model M2, each with a value over 0.02. It can be seen that larger compounds (e.g., decane and undecane) have higher leverage, as they tend to have different Ostwald solubility coefficients than the rest of the compounds. The full set of these long aliphatic hydrocarbons formed a "line" on the score plot, and their DModX is small, indicating that their high leverage is acceptable. The influence of these five compounds on model M2 is between 1−1.5%. Thus, they do not affect the model enough to be considered outliers and should not be removed. For all remaining solutes, the influence is less than 1%, which means that the variance among the observations (solutes) is equally distributed. For completeness, leverages of all solutes are provided in the (Supporting Information, Table S4).

The contributions of the variables (solvents) can be evaluated by analyzing the cumulative fraction of the variations explained by the selected component ($R^2VX_{(cum)}$). This reveals only 7 solvents with $R^2VX_{(cum)}$ less than 0.95. Only two of the seven solvents, perfluorooctane (cID565) and water (cID116), have $R^2VX_{(cum)}$ values are less than 0.9 (0.82 and 0.88, respectively). Thus, the solvent series is well described, and each contributes significantly to the model. Statistics for all the variables (solvents) are provided in the Supporting Information (Table S5).

Analysis of Scores and Loadings. As discussed above, the first two PCs account for 98.6% of the total data variance. The physical meaning of the respective scores and loadings was analyzed via construction of QSPR models. The optimal model for the score vector of the first PC consisted of two molecular descriptors: gravitation index (all bonds) and HA dependent HDCA-1. Based on all 320 solutes, this model had an excellent squared correlation coefficient $R^2 = 0.96$. The graphical representation of the predicted and actual score values is depicted in Figure S4 of the Supporting Information, and the correlation equation is given in Table 3 (model 1).

The two descriptors in the QSPR equation have justified physicochemical meanings. The gravitation index (all bonds),

Table 3. QSPR Models for Scores and Loadings of PCA Model M2

no.	X	$\pm X$	t test	descriptor	ΔG_S components
			QSPR Model 1: First Score, Observed Range: 3.924−110.459[a]		
0	1.554	0.559	2.780	intercept	
1	0.0698	0.000820	85.050	gravitation index (all bonds)	cavity
2	6.166	0.205	30.013	HA dependent HDCA-1 (Zefirov PC) (all)	HB
			QSPR Model 2: Second Score, Observed Range: −10.683−13.261[b]		
0	12.345	0.516	23.908	intercept	
1	0.0838	0.00284	29.483	complementary information content (order 0)	cavity
2	−2.544	0.106	−24.090	total molecular electrostatic interaction	electrostatic
3	−4.118	0.206	−19.978	total hybridization component of the molecular dipole	electrostatic
4	−16.930	1.070	−15.827	polarity parameter (Zefirov)	electrostatic
			QSPR Model 3: First Loading, Observed Range: 0.0376−0.0963[c]		
0	0.0485	0.00577	8.400	intercept	
1	0.0470	0.00462	10.154	polarity parameter (Zefirov)	dispersion
2	0.0245	0.00267	9.159	maximum bonding contribution of one MO	HB
3	−0.00310	0.000357	−8.685	LUMO energy	HB
4	-7.72×10^{-6}	1.22×10^{-6}	−6.310	gravitation index (all bonds)	dispersion
5	0.00120	0.000283	4.223	total molecular 2-center resonance energy	electrostatic
6	−0.289	0.0900	−3.213	H-donors FCPSA (version 2)	HB
			QSPR Model 4: Second Loading, Observed Range: −0.325−0.278[d]		
0	0.222	0.0165	13.481	intercept	
1	−8.539	0.493	−17.330	FPSA3 fractional PPSA (PPSA-3/TMSA) (Zefirov PC)	electrostatic
2	0.0922	0.00850	10.839	difference (Pos − Neg) in charged partial surface area (Zefirov's PC)	electrostatic
3	−0.0126	0.00142	−8.839	maximum atomic force constant	electrostatic
4	−0.0509	0.00882	−5.769	total hybridization component of the molecular dipole	electrostatic
5	0.00304	0.000625	4.869	count of H-donors sites (Zefirov PC) (all)	HB

[a] $R^2 = 0.962$, $R^2_{CV} = 0.961$, $R^2_{CVMO} = 0.961$, $R^2_{SCR} = 0.006$, $F = 4020.79$, and $S^2 = 12.84$. [b] $R^2 = 0.910$, $R^2_{CV} = 0.907$, $R^2_{CVMO} = 0.906$, $R^2_{SCR} = 0.013$, $F = 794.86$, and $S^2 = 1.926$. [c] $R^2 = 0.684$, $R^2_{CV} = 0.632$, $R^2_{CVMO} = 0.624$, $R^2_{SCR} = 0.040$, $F = 52.652$, and $S^2 = 1.71 \times 10^{-5}$. [d] $R^2 = 0.781$, $R^2_{CV} = 0.755$, $R^2_{CVMO} = 0.754$, $R^2_{SCR} = 0.033$, $F = 104.65$, and $S^2 = 0.00149$.

G^2, provides a direct estimation of the mass distribution within the molecular space of the solute (eq 5).

$$G^2 = \sum_{i<j}^{N_b} \frac{m_i m_j}{r_{ij}^2} \tag{5}$$

where m_i and m_j are the atomic masses of atoms i and j, r_{ij} is the atomic distance between the bonded atoms i and j, and N_b is the number of chemical bonds in the molecule. Therefore, this descriptor characterizes size- and bulk-related properties of the solute molecules.

The HA dependent HDCA-1, $HDCA1^{HAdep}_{(Z,all)}$, computes the hydrogen-donor ability of the solute over its solvent-accessible surface area (eq 6).

$$HDCA1^{HAdep}_{(Z,all)} = \sum_D S_D \tag{6}$$

where S_D is the solvent-accessible surface area of hydrogen-bonding donor H atoms, identified by the threshold charge on a hydrogen atom. A comparison with eq 2 indicates that these two descriptors are each directly related to their respective terms; the bulk descriptor G^2 characterizes solute size and is related to cavity formation, while $HDCA1^{HAdep}_{(Z,all)}$ describes specific electrostatic interactions and, more importantly, the HB. The derived QSPR model utilizes the same descriptors as the models for boiling point[25] and vapor pressure[29] reported previously. Such similarity is not unexpected because all of these properties are interrelated. Boiling point is connected with vapor pressure through the Clausius−Clapeyron equation, and according to the ideal gas law, gas−liquid equilibria depend solely on vapor pressure, and gas concentration is a denominator in the Ostwald solubility coefficient. Thus, these properties depend quantitatively on similar structural features of a compound.

Based on the descriptors calculated for the 153 solvents, the QSPR equation containing six descriptors was derived for the loading of the first PC (see Table 3, model 3). A regression equation is obtained with moderate statistical parameters, $R^2 = 0.68$ and $F = 53$. Investigation of the equation shows that water is a strong outlier which strongly influences the correlation. The descriptors in the model are polarity parameter, maximum bonding contribution for one molecular orbital (MO), lowest unoccupied MO energy (LUMO) energy, G^2, total molecular 2-center resonance energy, and H-donor FCPSA (version2). The moderate correlation coefficient of the relationship is caused by the concealed connections between the loadings' and solvents' characteristics. In addition, the descriptors of model 3 can be grouped into terms as described by eq 2, namely, descriptors related to dispersion, hydrogen-bonding and electrostatic interactions are identified. However, the values of the first loading are in a narrow interval (0.0376−0.0963) compared to the scores (3.924−110.459). Consequently, the score variation is about 1800 times bigger than that for loadings, and thus the QSPR model for the first score is more relevant. This leads to the conclusion that the loading values for solvents are almost uniform, and indicates that Ostwald solubility coefficient magnitude is mainly influenced by the nature of the solute.

Similar QSPR equations were also developed for the loading and score vectors of the second PC (see Table 3, models 2 and 4). Here again, the descriptors in the respective models can be grouped into terms as in eq 2. The descriptors

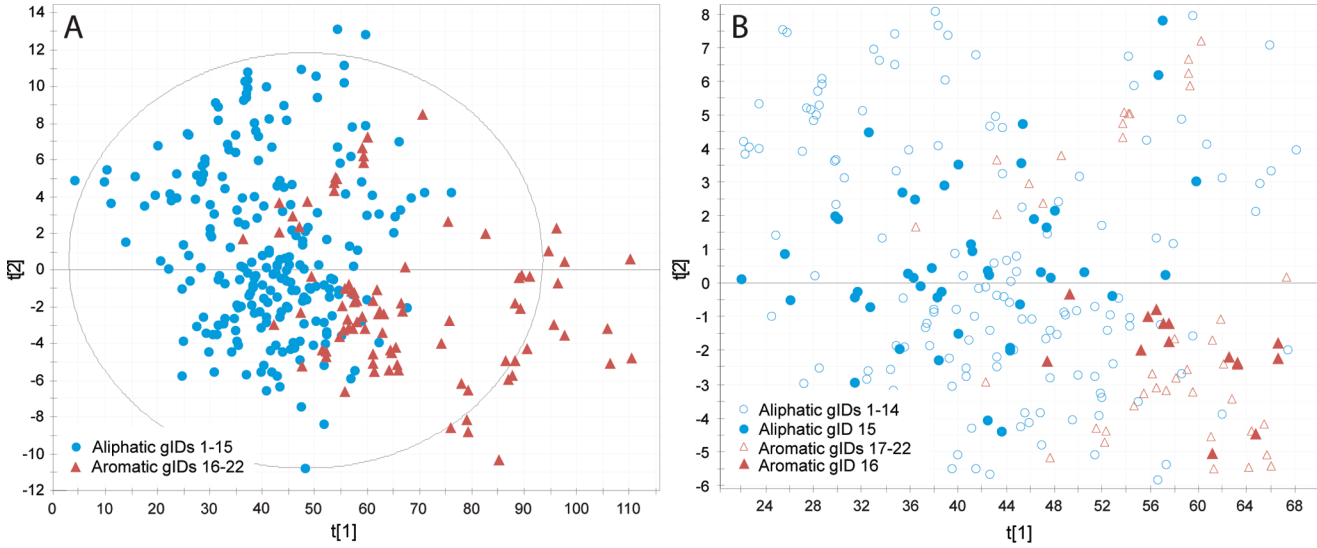

Figure 2. Score plots of model M2. (A) Aliphatic and aromatic compounds and (B) halo-hydrocarbons (IDs correspond to Table 4).

of the second score model are mostly related to electrostatic interactions, with one descriptor related to cavity formation (Table 3). The descriptors of the second loading model are related to electrostatic interactions, accompanied by one descriptor coding HB directly. Unlike the first loading, which has a very narrow variation, variation among the second loading is 10 times larger and shows an equal distribution of positive and negative loading values (Supporting Information, Figure S5). The second loading distinguishes the solvents according to their potential electrostatic interactions, which is supported and described through the relationship with hydrophobicity (Supporting Information, Figure S5). Therefore, the second PC accounts for the nature of solvents and describes the solutes' specific interactions with solvents.

Pattern Analysis of Score Plot. The distribution of solutes according to the functional groups listed in the Supporting Information, Table S1, does not exhibit a clear pattern on the score plot. Figure S1 in the Supporting Information shows strong overlap between different chemical groups, and although clustering is present, it is concealed by the complexity of the data. The next logical step is to simplify the pattern analysis of the score plot and to explain each component of the pattern.

The score plot of the first and second scores shows a separation between aliphatic and aromatic compounds (Figure 2A). A separate cluster of aromatic compounds in the aliphatic area contains exclusively benzenes with aliphatic substituents. Other groups of compounds in the overlapping area between aliphatic and aromatic compounds include benzenes with halogen substitutions, heteroaromatic compounds, and benzenes with a maximum of two HB donor−acceptor sites. Further analysis of the pattern of aliphatic and aromatic clusters incorporated structural (carbon) skeletons of the molecules as well as their saturation levels together with functional groups and provided the subsets given in Table 4. This analysis showed that the clear view of aliphatic and aromatic clusters is shadowed by the halo-hydrocarbons that are distributed over the PCA model plane. This is illustrated by Figure 2B, which shows the distribution of halo-hydrocarbons (Table 4: group IDs of gID15 and gID16). In Figure 3A and B, the halo-hydrocarbons have been hidden in order to observe the

Table 4. Two-Level Grouping of Solutes Based on Molecular Skeleton (Level 1) and Functional Groups (Level 2)[a]

group id	level 1	level 2
1	aliphatic:	saturated (aliphatic substitutions only)
2	aliphatic:	unsaturated (aliphatic substitutions only, includes also two cycles for solutes)
3	aliphatic:	saturated alicycles (aliphatic substitutions only)
4	aliphatic:	ethers (includes four cyclic ethers)
5	aliphatic:	amines (includes four cyclic amines)
6	aliphatic:	sulfides (includes one thiol for solutes)
7	aliphatic:	ketones (includes two cyclic ketones)
8	aliphatic:	nitriles
9	aliphatic:	aldehydes
10	aliphatic:	nitro compounds
11	aliphatic:	esters
12	aliphatic:	alcohols (includes two cyclic alcohols)
13	aliphatic:	amides (includes one cyclic amide)
14	aliphatic:	acids
15	aliphatic:	halogen substitutions only
16	aromatic:	benzenes (halogen substitutions only)
17	aromatic:	benzenes (aliphatic substitutions only)
18	aromatic:	benzenes (maximum two hydrogen-bond donor−acceptor sites)
19	aromatic:	benzenes (more than two hydrogen-bond donor−acceptor sites)
20	aromatic:	heteroaromatic (one ring only)
21	aromatic:	polyaromatic hydrocarbons (PAH)
22	aromatic:	heteropolyaromatic compounds

[a] Following the pattern on the scores plot.

patterns provided by the aliphatic (15 groups) and aromatic (7 groups) clusters (Table 4).

Upon examination of the aliphatic compounds alone (Figure 3A), one does observe some clearly distinguishable groups as well as other groups that have strong overlap with each other. A more detailed analysis provides an explanation regarding the overlap and reveals the pattern. In the upper left corner, aliphatic compounds form a distinct group of hydrocarbons (gID1−3) consisting of saturated, unsaturated, and saturated alicyclic compounds with aliphatic substituents (Table 4). Hydrocarbons are considered to be nonpolar. The difference between the three groups can be attributed to the electronegativity order of hybridized carbon orbitals (sp > sp2 > sp3) and to the substitution pattern. In following discussion, one must also consider differences between

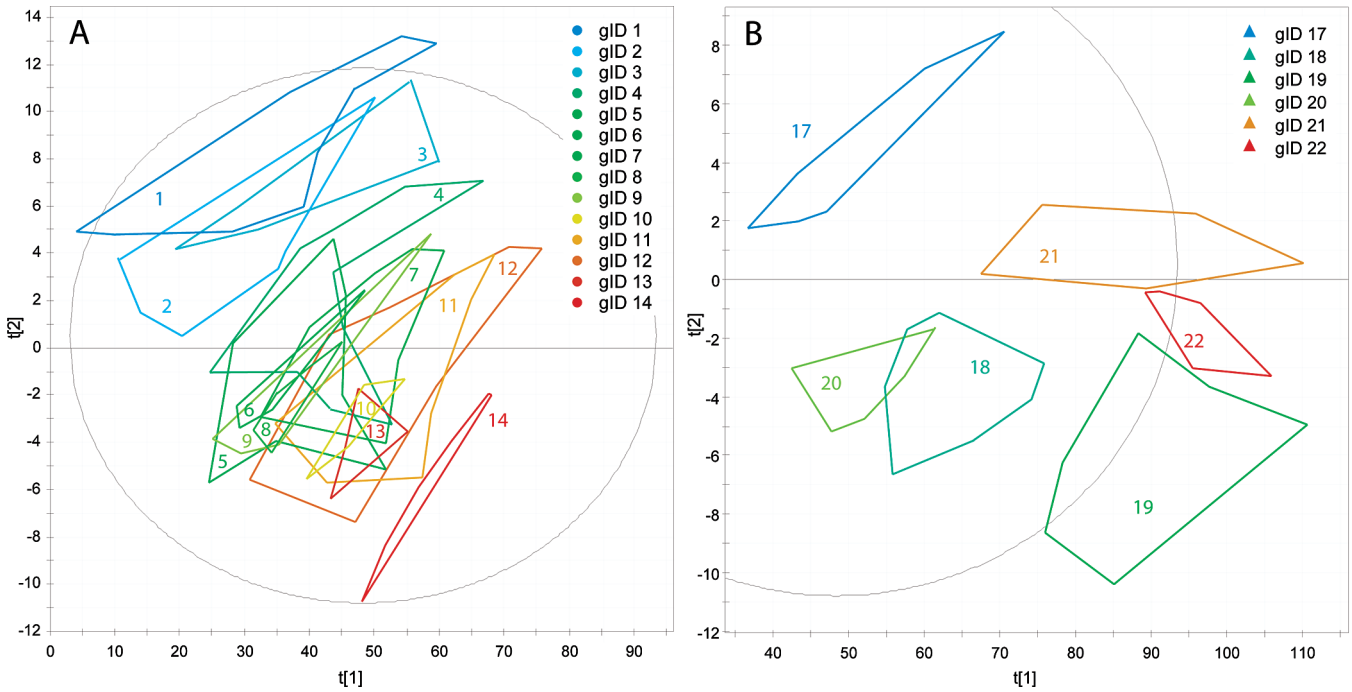

Figure 3. Score plots of model M2. (A) Aliphatic compounds and (B) aromatic compounds (IDs correspond to Table 4).

aprotic (ethers, ketones, nitriles, aldehydes, nitro compounds, and esters) and protic (amines, alcohols, and carboxylic acids) solutes along with the electronegativity of atoms and groups that cause the compound's polarity. Electrostatic interactions are dominant when analyzing the compound pattern along the diagonal from the upper left to the lower right corner. The other diagonal from the lower left to the upper right corners exclusively illustrates the size dependence. Groups 1−3 are followed by the ethers (gID4). Ethers are more polar than alkenes but not as polar as esters (gID11), alcohols (gID12), or amides (gID13) of comparable structure. This positions ethers in between those groupings. There are two points that severely deviate (Supporting Information, Figure S9) from the trend of ethers, namely 1,2- (cID458) and 1,4-dioxane (cID68). The deviation is probably due to the presence of two oxygens, which leads to differences in electronic structure and HB characteristics. Cyclic ethers are miscible in water and similar protic solvents because of the more exposed oxygen atom for hydrogen bonding as opposed to aliphatic ethers. The next group visible as one travels from the upper left to the lower right corner of the plot are amines (gID5), which is the first protic group and possesses properties of organic bases. The nitrogen in the amine is much less electronegative than oxygen in the respective alcohol. Therefore, the dipole on N−H is much weaker than the dipole on O−H. This discriminates amines from other protic solutes. The amine group contains all types of amines: primary, secondary, and tertiary. Two tertiary amines, triethylamine (cID198) and trimethylamine (cID493) are not HB donors and reside in the ether region, creating overlap between the two subsets. The amine group also contains a number of cyclic compounds containing a nitrogen atom as part of the ring system. One of the cyclic compounds, morpholine (cID460), deviates from the rest (Supporting Information, Figure S10) because of the additional electronegative atom (oxygen) in the compound, resulting in improved HB acceptor characteristics along with changes in electrostatic interactions, including polarity. The next small and overlapping aprotic groups are sulfides (gID6), nitriles (gID8), and aldehydes (gID9). The chemical properties of sulfides (or thioethers) are similar to those of the corresponding ethers. Thioethers are positioned slightly further apart because their dipole moments are somewhat larger in comparison with those of ethers. The sulfide group also includes two thiols located near the sulfides (Supporting Information, Figure S11). The low polarity of the S−H bond makes thiols comparable to the isomeric sulfides. Nitriles (gID8) and aldehydes (gID9) form rows of homologues (Supporting Information, Figures S13 and S14). The aldehyde group is aprotic and polar due to the electronegative double-bonded oxygen attached to the carbon, making them more polar than aliphatic ethers (gID4). Ketones (gID7) form a large group of aprotic solutes covering a wider area (Supporting Information, Figure S12) than aldehydes because of the inclusion of cyclic ketones (cyclohexanone (cID77) and cyclopentanone (cID326)) and 2,4-pentanedione (cID426). A series of homologous nitro compounds (gID10) reside in the middle of two large groups: esters (gID11) and alcohols (gID12). The ester group (gID11) consists of 20 compounds following the diagonal from the smallest to the largest compound (Supporting Information, Figure S16). Dimethyl carbonate (cID392) is located separately, close to the carboxylic acids (gID14), which may be due to the additional HB acceptor site.

Alcohols (gID12) form the largest single group and contain 26 compounds. Their polarity is comparable with esters and they also possess HB donor properties. The diverse substitution pattern of alcohols results in a wide distribution in the plot area. Amides (gID13) form a small separate group with four compounds, and their polar properties resembling esters and alcohols place them in the same area of the scores plot. Although amides are considered to be more polar than acids (gID14), they are positioned before acids in the pattern, which is probably due to the HB donor capabilities of acids in comparison with amides. A row of homologous carboxylic

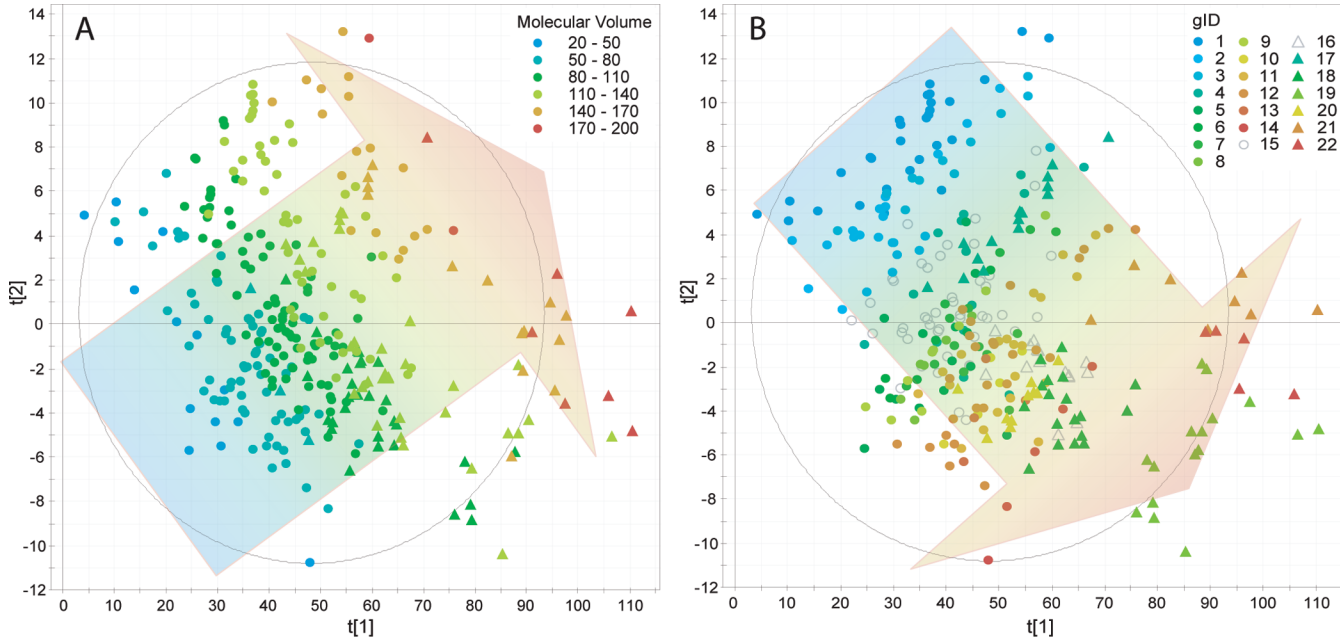

Figure 4. Score plots of model M2. (A) An arrow indicates the growth of molecule size. Data points are colored according to the molecular volume. (B) An arrow indicates the growth of polarity, electronegativity, and strength of electrostatic interactions. Data points are colored according to the grouping of solutes given in Table 4.

acids (gID14, Supporting Information, Figure S14) is a group that can be easily distinguished in the pattern.

When observing aromatic compounds separately (Figure 3B), one can identify benzenes with aliphatic substituents (gID17) as a distinct group in the upper left corner, ordered along the diagonal from benzene (cID29) to hexamethylbenzene (cID430), with all combinations of substitution in between (Supporting Information, Figure S22). Single-ring heteroaromatic compounds (gID20) and substituted benzenes having a maximum of two HB donor−acceptor sites (gID18) are adjacent and slightly overlap each other. Of the heteroaromatic compounds, only furan (cID298) deviates from the general trend, perhaps because it is the only compound in the group that contains an oxygen atom in the ring. All other heteroaromatic compounds contain nitrogen (Supporting Information, Figure S25). A group of substituted benzenes with a maximum of two HB donor−acceptor sites (gID18) consists of benzenes containing a single-electronegative aprotic or protic substituent, and disubstituted benzenes where the second substituent is a methyl group (Supporting Information, Figure S23). The next group consists of benzenes with more than two HB donor−acceptor sites (gID19) in the lower right corner of the score plot. The group contains two monosubstituted benzenes, benzoic acid (cID118) and fenuron (cID533), and the remaining solutes are mostly di- and para-substituted derivatives of benzoic acid or benzaldehyde. Polyaromatic hydrocarbons (gID21) form a separate group, followed by the heteropolyaromatic compounds (gID22). This group includes one deviating secondary amine, diphenylamine (cID433), placed here due to its electrostatic interactions (two π-systems and hydrogen-donor site).

CONCLUSIONS

Construction and analysis of our two-component principal component analysis (PCA) model for Ostwald solubility coefficients produced an explained variance of 98.6%. Most of the variance (97.8%) is covered by the first component. Quantitative structure−property relationship (QSPR) analysis indicates that the principal components describe multiple solubility interactions rather than a single solute−solvent interaction. The first component represents cavity formation and hydrogen-bond (HB) interactions which can be codified by the gravitational index and the hydrogen donor charged surface area molecular descriptors. The second component covers weaker and more specific electrostatic interaction types. Analysis of the pattern observed in the score plot provides a detailed explanation for each chemical group.

Despite the multi-effect nature of scores and loadings, in addition to the explained pattern of the score plots, two general types of interactions graphically represented in the score plot can be discerned. First, the diagonal from the lower left to the upper right corner in Figure 4A describes the size of the molecule, i.e., the nonspecific interactions which contribute to the cavity formation and the dispersion force energy terms in eq 2. In Figure 4A, this trend is illustrated with the molecular volume. Second, the diagonal from the upper left to the lower right corner of Figure 4B indicates specific interactions that are related to the polarity of functional groups (electronegativity of atoms) which contribute to the HB, the electron pair donor/acceptor, and other specific interaction forces. The trend is illustrated according to the groupings given in Table 4. However, we could not identify any single property or molecular descriptor which both quantified the electrostatic interactions and also showed the complex nature of specific interactions.

External validation of the previously published predicted Ostwald solubility coefficients provides a significant correlation, $R^2_{val} = 0.881$, between the 1210 data points. This exemplifies the relevance of the proposed methodology of combining QSPR and PCA techniques for estimating solubility in this series of publications. In principle, this approach could be used for the prediction and analysis of other large sets of experimental values.

ACKNOWLEDGMENT

I.T. and U.M. are grateful for the financial support from the Estonian Ministry of Education and Research (grant SF0140031Bs09) and the Estonian Science Foundation (grant no. 7709). D.D. acknowledges funding by the European Union through the European Regional Development Fund through the Center of Excellence in Chemical Biology, Estonia. A.R.K. is grateful for support from the Kenan Foundation.

Supporting Information Available: Added experimental data (SI-A), data matrix (SI−B), supporting tables (SI−C), and supporting figures (SI-D). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Reichardt, C. *Solvents and Solvent Effects in Organic Chemistry*, 3rd ed.; Wiley-VCH: New York, 2003.
(2) Dearden, J. C. In Silico Prediction of aqueous solubility. *Expert Opin. Drug Discovery* **2006**, *1*, 31–52.
(3) Llinás, A.; Glen, R. C.; Goodman, J. M. Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements. *J. Chem. Inf. Model.* **2008**, *48*, 1289–1303.
(4) Hopfinger, A. J.; Esposito, E. X.; Llinás, A.; Glen, R. C.; Goodman, J. M. Findings of the Challenge To Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2009**, *49*, 1–5.
(5) Wang, J. M.; Hou, T. J.; Xu, X. J. Aqueous Solubility Prediction Based on Weighted Atom Type Counts and Solvent Accessible Surface Areas. *J. Chem. Inf. Model.* **2009**, *49*, 571–581.
(6) Hewitt, M.; Cronin, M. T. D.; Enoch, S. J.; Madden, J. C.; Roberts, D. W.; Dearden, J. C. In Silico Prediction of Aqueous Solubility: The Solubility Challenge. *J. Chem. Inf. Model.* **2009**, *49*, 2572–2587.
(7) Katritzky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, P.; Tatham, D. B.; Maran, U.; Lomaka, A.; Acree, W. E., Jr. A General Treatment of Solubility. Part 1. The QSPR Correlation of Solvation Free Energies of Single Solutes in Series Solvents. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1794–1805.
(8) Katritzky, A. R.; Oliferenko, A. A.; Oliferenko, P. V.; Petrukhin, P.; Tatham, D. B.; Maran, U.; Lomaka, A.; Acree, W. E., Jr. A General Treatment of Solubility. Part 2. QSPR Prediction of Free Energies of Solvation of Specified Solutes in Ranges of Solvents. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1806–1814.
(9) Katritzky, A. R.; Tulp, I.; Fara, D. C.; Lauria, A.; Maran, U.; Acree, W. E., Jr. General Treatment of Solubility. Part 3. Principal Component Analysis (PCA) of Solubilities of Diverse Solutes in Diverse Solvents. *J. Chem. Inf. Model.* **2005**, *45*, 913–923.
(10) Tomasi, J.; Persico, M. Molecular Interactions in Solution: An Overview of Methods Based on Continuous Distributions of the Solvent. *Chem. Rev.* **1994**, *94*, 2027–2094.
(11) Karelson, M. Quantum Chemical Treatment of Molecules in Condensed Disordered Media. *Adv. Quantum Chem.* **1997**, *28*, 141–157.
(12) Vitha, M.; Carr, P. W. The chemical interpretation and practice of linear solvation energy relationships in chromatography. *J. Chromatogr., A* **2006**, *1−2*, 143–194.
(13) Drago, R. S. Solvation. In *Applications of Electrostatic-Covalent Models in Chemistry*, Surfside Scientific Publishers: University of Florida, Gainesville, 1994; pp 183−218.
(14) Dutt, G. B. Molecular Rotation as a Tool for Exploring Specific Solute-Solvent Interactions. *Chem. Phys. Chem.* **2005**, *6*, 413–418.
(15) Desiraju, G. R. Hydrogen Bridges in Crystal Engineering: Interactions without Borders. *Acc. Chem. Res.* **2002**, *35*, 565–573.
(16) Kettaneh, N.; Berglund, A.; Wold, S. PCA and PLS with very large data sets. *Comput. Stat. Data Anal.* **2005**, *48*, 69–85.
(17) Eriksson, L.; Antti, H.; Gottfries, J.; Holmes, E.; Johansson, E.; Lindgren, F.; Long, I.; Lundstedt, T.; Trygg, J.; Wold, S. Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Anal. Bioanal. Chem.* **2004**, *380*, 419–429.
(18) *SIMCA-P*, version 11.5; Umetrics AB: Umeå, Sweden, 2008.
(19) Johnson, R. A.; Wichern, D. W. Inferences about a Mean Vector. *In Applied Multivariate Statistical Analysis*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1982; pp 177−225.
(20) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. PCA. In *Multi- and Megavariate Data Analysis*, Umetrics AB, Umeå, Sweden, 2001; pp 43−69.
(21) Gempeline, P. Principal Component Analysis. In *Practical Guide to Chemometrics*; 2nd ed.; Gempeline, P., Ed.; CRC Press, Taylor & Francis Group: Boca Raton, FL, 2006; pp 69−104.
(22) Hill, T.; Lewicki, P. *Electronic Statistics Textbook*; StatSoft, Inc.: Tulsa, OK, 2007; www.statsoft.com/textbook/stathome.html.
(23) Varmuza, K. Chemometrics: Multivariate View on Chemical Problems. In *Encyclopedia of Computational Chemistry.*; Schleyer, P.v.R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Schreiner, P. R., Eds.; John Wiley & Sons: Chichester, UK, 1998; Vol. 1, pp 346−366.
(24) *CODESSA PRO*, version 1.0 RC2; University of Florida: Gainesville, FL, 2002.
(25) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of boiling points with molecular structure 0.1. A training set of 298 diverse organics and a test set of 9 simple inorganics. *J. Phys. Chem.* **1996**, *100*, 10400–10407.
(26) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally Diverse QSPR Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1–18.
(27) Karelson, M.; Maran, U.; Wang, Y.; Katritzky, A. R. QSPR and QSAR Models Derived Using Large Molecular Descriptor Spaces. A Review of Codessa Applications. *Collect. Czech. Chem. Commun.* **1999**, *64*, 1551–1571.
(28) Katritzky, A. R.; Tatham, D. B.; Maran, U. The Correlation Of The Solubilities Of Gases And Vapors In Methanol And Ethanol With Their Molecular Structures. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 358–363.
(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.

CI1000828