Solubility Parameters of Nonelectrolyte Organic Compounds: Determination Using Quantitative Structure−Property Relationship Strategy

ARTICLE *in* INDUSTRIAL & ENGINEERING CHEMISTRY RESEARCH · SEPTEMBER 2011

Impact Factor: 2.59 · DOI: 10.1021/ie200962w

5 AUTHORS, INCLUDING:

Farhad Gharagheizi
Texas Tech University
168 PUBLICATIONS **2,915** CITATIONS
SEE PROFILE

Ali Eslamimanesh
OLI Systems
106 PUBLICATIONS **1,755** CITATIONS
SEE PROFILE

Amir H. Mohammadi
557 PUBLICATIONS **4,820** CITATIONS
SEE PROFILE

Dominique Richon
Aalto University
533 PUBLICATIONS **6,601** CITATIONS
SEE PROFILE

I&EC research
Industrial & Engineering Chemistry Research

ARTICLE

pubs.acs.org/IECR

Solubility Parameters of Nonelectrolyte Organic Compounds: Determination Using Quantitative Structure−Property Relationship Strategy

Farhad Gharagheizi,[†] Ali Eslamimanesh,[‡] Farhad Farjood,[§] Amir H. Mohammadi,[*,‡,∥] and Dominique Richon[‡]

[†]Saman Energy Giti Company, Postal Code 3331619636 Tehran, Iran

[‡]MINES ParisTech, CEP/TEP - Centre Energétique et Procédés, 35 Rue Saint Honoré, 77305 Fontainebleau, France

[§]School of Chemical Engineering, University of Birmingham, Birmingham B15 2TT, United Kingdom

[∥]Thermodynamics Research Unit, School of Chemical Engineering, University of KwaZulu-Natal, Howard College Campus, King George V Avenue, Durban 4041, South Africa

Supporting Information

ABSTRACT: The solubility parameter is considered to be a significant parameter for the chemical industry. In this study, the quantitative structure−property relationship (QSPR) method is applied to develop three models for determination of the solubility parameters of pure nonelectrolyte organic compounds at 298.15 K and atmospheric pressure. To propose comprehensive, reliable, and predictive models, about 1400 data belonging to experimental solubility parameter values of various nonelectrolyte organic compounds are studied. The genetic function approximation (GFA) mathematical approach is applied for selection of proper model parameters (molecular descriptors) and to develop a linear QSPR model. To study the nonlinear relations between the selected molecular descriptors and the solubility parameter, two approaches are pursued: the three-layer feed forward artificial neural networks (3FFANN) and the least square support vector machine (LSSVM). Furthermore, the Levenberg−Marquardt (LM) and genetic algorithm (GA) optimization methods are respectively implemented to optimize the 3FFANN and LSSVM models. Consequently, we obtain three predictive models with satisfactory results quantified by the following statistical parameters: absolute average relative deviation (AARD) of the represented/predicted properties from existing experimental values by the GFA linear equation of 4.6% and squared correlation coefficient of 0.896; AARD of the QSPR-ANN model of 3.4% and squared correlation coefficient of 0.941; and AARD of 3.1% and squared correlation coefficient of 0.947 evaluated by the QSPR-LSSVM model.

1. INTRODUCTION

Searching for the most appropriate solvents for particular industrial purposes is of utmost interest for the engineers and scientists dealing with the different chemical, pharmaceutical, petroleum, and polymer engineering processes.[1] In 1931, Scatchard[2] introduced a physicochemical parameter defining a solvent affinity for solving a definite kind of solute. This parameter was later described in detail by Hildebrand and Scott,[3] who named it the "solubility parameter". Hansen[4,5] continued to investigate and improve the concept of the solubility parameter along with stating its various applications in coating and paint technologies, complex extraction operations, polymer processes, etc.[1,6] On the other hand, many of the presented thermodynamic models for prediction of the amounts/conditions of precipitations/depositions of heavy petroleum fractions such as asphaltene and wax have been generally developed based on the regular solution theory,[7] which is based on the difference between the solubility parameters of the solute (asphaltene/wax) and the related solvent (maltene/oil).[8−16]

The solubility parameter definition is generally written as follows:[1−5]

$$\delta = \left(\frac{\Delta E_v}{v}\right)^{1/2} = \left(\frac{\Delta U_{vap}}{v}\right)^{1/2} = \left(\frac{\Delta H_{vap} - RT}{v}\right)^{1/2} \quad (1)$$

where δ is the Hildebrand one-component solubility parameter, ΔE_v denotes the cohesive energy, which is introduced as the energy required for separating a molecule from its surrounding neighbors,[1−5] v is the molar volume, ΔU_{vap} represents the energy change upon isothermal vaporization of the saturated liquid to the ideal gas state (energy of a complete vaporization),[17] and ΔH_{vap} stands for the enthalpy of vaporization. The solubility parameter can also be related to the internal pressure, which stands for the change in internal energy of a liquid upon a small isothermal expansion as follows:[1,18]

$$P_i = T\left(\frac{\partial P}{\partial T}\right)_v - P = \delta^2 \quad (2)$$

where P_i is the internal pressure and T is temperature.

The solubility parameter defined by eqs 1 and 2 is based on the famous statement: "like dissolves like".[1] However, the interactions between the solvents and solutes are also originated from their electron pairs, donor−acceptors, and hydrogen-bonding interactions; i.e., the aforementioned Hildebrand parameter does

Received: May 4, 2011
Accepted: August 8, 2011
Revised: August 2, 2011
Published: August 08, 2011

not account for these interactions and considers only one part of the molecular forces, which is the dispersion.[1,6] Therefore, the value of the Hildebrand solubility parameter is applicable for the systems including weakly interacting species. Consequently, Hildebrand's theory was modified by several authors to derive the two-component solubility parameter as follows:[19−25]

$$\delta = (\delta_\lambda^2 + \delta_\tau^2)^{1/2} \quad (3)$$

where subscripts "λ" and "τ" stand for nonpolar and polar solubility parameters, respectively.

Later, Hansen[5] proposed the three-component (Hansen) solubility parameter considering the effects of all of the cohesive bonds including the atomic dispersion forces, the molecular permanent dipole−permanent dipole forces, and the molecular hydrogen bonding on the solubility parameter value as follows:

$$\delta_{HSP} = (\delta_D^2 + \delta_P^2 + \delta_H^2)^{1/2} \quad (4)$$

where the subscripts "D", "P", and "H" stand for the dispersion, polar, and hydrogen-bonding effects, respectively, and the subscript "HSP" indicates the total Hansen solubility parameter. The values of the one-component and total three-component solubility parameters would be almost the same for the substances with nonpolar and non-hydrogen-bonding effects such as the light hydrocarbons.

Considering the significance of the solubility parameter issue, many attempts have been made for its calculation/estimation. Fedors[26] correlated the solubility parameters and molar volumes of several chemical liquids. The Carnahan−Starling[27] equation of state was used by Lozada et al.[28] to estimate the Hildebrand one-component solubility parameter of several chemical compounds in their supercritical states. Group contribution methods have been also developed for the calculation/estimation of the partial solubility parameters, i.e., δ_D, δ_P, and δ_H, for a few chemical compounds.[29−31]

Allada[32] determined the Hildebrand solubility parameters for several pure compounds, applying their internal energies in supercritical state. Lattice fluid theory (LFT) was employed by Panayiotou[33] in order to estimate the hydrogen-bonding component (δ_H) of the three-component (Hansen) solubility parameter. Later, Williams et al.[34] proposed a correlation to represent the HSP of CO_2 in wide ranges of pressures and temperatures including the supercritical region.

In 2004, Bozdogan[35] concluded that the partial molar entropy change of a polymer for mixing at a given temperature is proportional to the hydrodynamic volume or segment number of the polymer. Using this theory, he calculated the solubility parameter of a high molar mass polymer at a specified temperature by extrapolating solubility parameter values of polymer fractions to high molar mass, applying the solubility parameter−segment number relation of the polymer fraction. Another approach was pursued by Utracki and Simha,[36] who applied the Simha and Somcynsky (S−S) lattice−hole theory to successfully represent the pressure−volume−temperature (PVT) surface of chain molecular melts and consequently their solubility parameters.

The calculation procedure of the HSP using the Hansen sphere model[4,5] was modified by Gharagheizi and co-workers[6] to reduce the deviations of the model predictions from experimental values, investigating the effects of the presented strategy on the values of the obtained solubility parameter. A model based on an equation of state was developed using a configurational partition function for calculation of the partial solubility parameters of several chemical compounds and polymers by Stefanis et al.,[37] who achieved satisfactory results. Moreover, the perturbed-chain statistical associating fluid theory (PC SAFT) was applied by Zeng et al.[38] for calculation of the HSPs of *n*-alkanes and 1-alcohols.

Code et al.[39] developed a quantitative structure−property relationship (QSPR) based model to predict Hildebrand solubility parameters of small organic molecules. They reported a squared correlation coefficient of 0.97 for the predicted solubility parameters of about 20 various compounds. The application of the "one-third" rule[40] for calculations of solubility parameters of hydrocarbons and crude oil systems was investigated by Vargas and Chapman.[40] They presented a correlation based on the molar density of the hydrocarbon fractions for this purpose. Recently, Eslamimanesh and Esmaeilzadeh[1] have successfully developed a model based on the modified Esmaeilzadeh and Roshanfekr (m-ER)[41,42] three-parameter cubic equation of state (CEoS) for calculation of the solubility parameters of different chemical compounds categorized in 13 chemical groups.

Regarding the preceding methods, there is still a need for developing more reliable methods for determination of the solubility parameters of large groups of chemical compounds. In this study, we propose novel approaches based on the quantitative structure−property relationship (QSPR) strategy to represent/predict the Hildebrand[3] solubility parameters of around 1400 nonelectrolyte organic compounds at 298.15 K and atmospheric pressure.

2. EXPERIMENTAL DATA AND MATHEMATICAL METHODS

2.1. Experimental Database. The accuracy and reliability of models for representation/prediction of physical properties, especially those dealing with a large number of experimental data, normally depend on the validity of the employed data set for their development.[43−77] A comprehensive data set consists of various chemical families and a high number of available pure compounds. In this work, the DIPPR 801 database,[78] which is one of the most reliable sources of physical property data for pure compounds, has been applied. The solubility parameter values of 1438 chemical species from various chemical families at 298.15 K and atmospheric pressure have been treated to develop and validate the models.

2.2. Determination of Molecular Descriptors. Molecular descriptors are defined as numerical characteristics associated with chemical structures.[49,65−69,72−74,76,77] They are basic molecular properties of a compound and are determined from the chemical structure. Each type of molecular descriptor is related to a specific type of interaction between chemical groups in a particular molecule.[49,65−69,72−74,76,77] Several software packages are generally used for the computation of molecular descriptors of any desired chemical structure. A review of these software packages can be found in the work of Todeschini and Consonni.[79] In this study, one of the most widely used software packages, Dragon,[80] has been used. This software is able to calculate more than 3000 molecular descriptors for any desired chemical structure. So far, these molecular descriptors have been calculated for about 234 000 pure compounds with Dragon software, which are freely available[81] (many of these compounds have not been synthesized up to now). Since the values of many descriptors are related to the bond lengths, bond angles, etc., each chemical structure is optimized before its molecular descriptors

are calculated. For this purpose, chemical structures of all 1438 pure organic compounds have been sketched in Hyperchem software[82] and optimized using the MM+ (the classical molecular dynamics) molecular mechanics force field. Finally, the molecular descriptors have been determined with the Dragon software.[80]

2.3. Developing the Models. In QSPR strategies, the next step after evaluation of molecular descriptors is to find a relationship between an optimal subset of variables (molecular descriptors) and the property under consideration (here the solubility parameter).[49,65−69,72−74,76,77] Different mathematical methods have been proposed in the open literature in order to select the optimal subset of molecular descriptors from the pool (a mathematical domain containing all of the possible answers) containing all of the molecular descriptors. In the majority of these methods, it is assumed that the optimal subset of molecular descriptors can be well represented using a linear correlation, based on the desired property. Later, the optimal subset of descriptors is selected using various strategies. This procedure enables us to detect those molecular descriptors that have the most statistical influences on the value of the considered property.[79,83] In this work, the genetic function approximation (GFA)[79,84] method has been applied for the selection of the appropriate molecular descriptors. The GFA[79,84] is a genetics-based algorithm of variable selection, which combines the traditional genetic algorithm (GA)[85] with Friedman's multivariate adaptive regression splines (MARS).[84,86−90] As a matter of fact, the GFA[79,84] evolves the population of equations that best fit the training set data.[89] It provides an error measure, called the Friedman's lack of fit (LoF) score, that automatically penalizes models with too many features.[90] The calculation steps of this algorithm are as follows:[89,90]

1. An initial population of equations is produced by the random selection of descriptors.
2. Parent pairs are selected from the population of equations at random, the crossovers are performed, and progeny equations are generated.
3. The fitness of each progeny equation is assessed by the LoF score.
4. If the fitness of new progeny equation is better than that of the previous one, then it is preserved.

The model with proper balance of all statistical terms will be used to explain the variance in the biological (genetic-based) activity.[89,90] One of the characteristics of the GFA[79,84] method is that it generates a population of models (e.g., 100), instead of generating a single model.[89,90] The range of variations in this population provides additional information on the quality of fitness and importance of the descriptors.[89,90]

The goodness criteria of each progeny equation is assessed by the lack of fit (LoF) score, which is obtained as follows:[87,89,90]

$$\mathrm{LoF} = \mathrm{LSE}/(1 - (C + dp)/m)^2 \quad (5)$$

where LSE is the least-squares error, C denotes the number of basis functions in the model, d stands for the smoothing parameter, p is the number of descriptors, and m is the number of observations in the training set used to train the mathematical model. In this work, the smoothing parameter has been set to 0.5. The best equation out of the considered population equations in the GFA approach is normally determined by observing the different statistical parameters such as regression coefficient, adjusted regression coefficient, and regression coefficient of cross validation.[87,89,90] For testing the validity of the obtained linear GFA[79,84] model, which is one of the significant steps in presenting QSPR based models, several validation techniques have been applied, which will be later explained in detail.

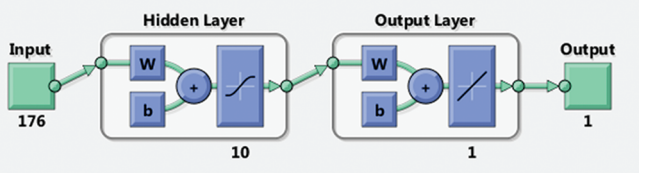

Figure 1. Schematic structure of the three-layer feed forward artificial neural network used in this study. *W*, weight; *b*, bias.

With the most proper molecular descriptors selected and consequently a linear correlation between these descriptors and solubility parameter values derived, they are further treated as the input variables of mathematical strategies for developing the nonlinear models, i.e., the nonlinear relations between the desired property (here, it is the solubility parameter) and the selected molecular descriptors.

2.4. Definition of Sub Data Sets. The database is generally divided into three sub data sets including the "training" set, the "validation (optimization)" set, and the "test (prediction)" set to begin the computational steps of the numerical modeling. In this paper, the training set is used to generate the structure of the models, the validation set is applied for optimization of the models, and the test set is used to investigate the prediction capability and validity of the obtained models. We have divided the main database into three sub data sets randomly. For this purpose, about 80, 10, and 10% of the main data set are randomly selected for the training set (1152 solubility parameter data), the validation set (143 solubility parameter data), and the test set (143 solubility parameter data). The effect of the percent allocation of the three sub data sets from the database on the accuracy of the developed model has been studied before.[91]

To present the final models, we have investigated two mathematical approaches in sections 2.4.1 and 2.4.2.

2.4.1. Artificial Neural Networks. Artificial neural networks are extensively used in various scientific and engineering problems,[46,48−79,87] e.g., calculations/estimations of physical and chemical properties of different pure compounds/mixtures[43−50] and phase behavior predictions of complex clathrate/semiclathrate hydrates.[54,55,57,60,62] Detailed descriptions of neural networks have been already well-established.[46,48−79,87] With the use of the artificial neural network toolbox of the MATLAB software (Mathworks Inc.), a three-layer feed forward artificial neural network (FFANN) has been developed for the problem. The typical structure of a three-layer FFANN is schematically presented in Figure 1.

All of the molecular descriptors and also the property values of the investigated pure compounds have been normalized between −1 and +1 to prevent truncation errors because we face a range of solubility parameter values for different organic compounds. This can be performed using maximum and minimum values of each molecular descriptor for input data and using maximum and minimum values of solubility parameters for output parameters. Moreover, this procedure, which is normally done in an optimization process, has been performed to obtain the parameters of the neural networks (W and b for the hidden layer and for the output layer, as shown in Figure 1) and it has no effects on the model results. Later, these values are again changed to the original solubility parameter values, which are finally reported as outputs of the developed model.

To generate an ANN model, the weight matrices and bias vectors should be determined.[46] As shown in Figure 1, there are two weight matrices and two bias vectors in a three-layer FFANN.[46] These parameters have been obtained through minimization of an objective function. The objective function used in this work is the sum of the squares of errors between the outputs of the ANN (represented/predicted properties) and the target values (experimental solubility parameters). This minimization has been performed by applying Levenberg−Marquardt (LM)[83] optimization strategy.

The main goal of developing an ANN model is to represent/predict the target values as accurately as possible. Generally and especially in three-layer FFANNs, it is more efficient that the number of neurons in the hidden layer is optimized according to the accuracy of the obtained FFANN.[46,48−79,87] Some factors should be taken into account in the evaluation of the optimum number of neurons. By increasing the number of neurons, the accuracy of the model, i.e., the squared correlation coefficient (R^2), is increased on the training set. However, the accuracy of the model on the test set is decreased gradually and the model may become unstable. Consequently, the overall R^2, which depends on the three data sets, fluctuates during changing of the numbers of neurons. The final (overall) R^2 value should be found through selecting the different numbers of neurons for a specified problem.

2.4.2. Least-Squares Support Vector Machine. In spite of the fact that ANNs have been generally proven to provide high accuracy for the models, they have the disadvantages of reproducibility of results, partly as a result of random initialization of the networks and variation of the stopping criteria during optimization.[92−94] The support vector machine (SVM) is a well-known strategy developed from the machine-learning community.[92,93] Some of the significant advantages of SVM methods in comparison with the traditional ANNs are as follows:[92,93,95]

1. There is a greater probability for convergence to the global optimum.
2. Normally a solution is found that can be quickly obtained by a standard algorithm (quadratic programming).
3. There is no need to determine the network topology in advance; it can be automatically determined as the training process ends.
4. There is generally less probability that the SVM strategy is encountered with an overfitting problem.

The SVM's outstanding performance makes it superior to the traditional empirical risk minimization principles. Furthermore, as a result of their specific formulation, sparse solutions can be found and both linear and nonlinear regressions can be performed.[92,93]

However, Suykens and Vandewalle[96] have presented a modification to the original SVM to overcome the difficulty of the previous algorithm in finding the final solution because it requires the solution of a set of nonlinear equations (quadratic programming). Their method, the least-squares SVM (LSSVM), encompasses advantages similar to those of SVM, but it requires solving a set of only linear equations (linear programming), which is much easier and more rapid compared to the traditional SVM method.[92,93]

In the LSSVM[96] approach, the regression error is defined as the difference between the represented/predicted property values and the experimental ones, which is considered as an addition to the constraint of the optimization problem. In the traditional SVM method, the value of the regression error is generally optimized during the calculations, while in the LSSVM[96] the error is mathematically defined.[92,97,98]

The cost function (penalized cost function) of the applied[96] method has been defined as follows:[92,97,98]

$$Q_{\mathrm{LSSVM}} = \frac{1}{2} w^{\mathrm{T}} w + \gamma \sum_{k=1}^{N} {e_k}^2 \tag{6}$$

subject to the following constraint:

$$y_k = w^{\mathrm{T}} \varphi(x_k) + b + e_k \qquad k = 1, 2, \ldots, N \tag{7}$$

In eqs 6and 7, x is the input vector of parameters of the model (molecular descriptors), y denotes the outputs (dependent parameter), b is the intercept of the linear regression in the modified SVM method (LSSVM),[96] w is the regression weight (slope of the linear regression), φ is the feature map, mapping the feasible region (input space) to a high dimensional feature space, in which the experimental hydrate dissociation data can be linearly separable by a hyperplane, e_k stands for the regression error for N training objects (the least-squares-error approach), γ indicates the relative weight of the summation of the regression errors compared to the regression weight (first term on the right-hand side of eq 6), and superscript "T" denotes the transpose matrix.

Using the Lagrange function,[92,97,98] the weight coefficient (w) is written as follows:[92,97,98]

$$w = \sum_{k=1}^{N} \alpha_k x_k \tag{8}$$

where

$$\alpha_k = 2\gamma e_k \tag{9}$$

Assuming the linear regression between the independent and dependent variables of the LSSVM[96] algorithm, eq 7 is rewritten as[92,97,98]

$$y = \sum_{k=1}^{N} \alpha_k {x_k}^{\mathrm{T}} x + b \tag{10}$$

Therefore, the Lagrange multipliers are calculated as[92,97,98]

$$\alpha_k = \frac{y_k - b}{{x_k}^{\mathrm{T}} x + (2\gamma)^{-1}} \tag{11}$$

The aforementioned linear regression can be well extended to a nonlinear one using the Kernel function as follows:[92,97,98]

$$f(x) = \sum_{k=1}^{N} \alpha_k K(x, x_k) + b \tag{12}$$

where $K(x,x_k)$ is the Kernel function calculated from the inner product of the two vectors x and x_k in the feasible region built by the inner product of the vectors $\Phi(x)$ and $\Phi(x_k)$ as follows:[92,97,98]

$$K(x, x_k) = \Phi(x)^{\mathrm{T}} \cdot \Phi(x_k) \tag{13}$$

In this work, the radial basis function (RBF) Kernel has been applied as below:[92,97,98]

$$K(x, x_k) = \exp\left(-\left\| x_k - x \right\|^2 / \sigma^2\right) \tag{14}$$

where σ is considered to be a decision variable, which is optimized by an external optimization algorithm during the calculations. The mean square error (MSE) of the results of the LSSVM[96] algorithm has been defined as

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n} (\delta_{\mathrm{rep/pred}_i} - \delta_{\mathrm{exp}_i})^2}{n} \quad (15)$$

where subscripts "rep/pred" and "exp" indicate the represented/predicted and experimental solubility parameter values, respectively, and n is the number of samples from the initial population. In this study, we have used the LSSVM[96] algorithm developed by Pelckmans et al.[97] and Suykens and co-workers[98] with some modifications, which will be discussed later.

3. RESULTS AND DISCUSSION

First, the most accurate linear equation has been obtained by pursuing the aforementioned GFA[79,84] computational procedure. For obtaining this equation, the best linear two-molecular descriptor model has been determined.[49,65−69,72−74,76,77] This procedure has been repeated to develop the most accurate three, four, five, etc. linear molecular descriptors model. It has been demonstrated that the most accurate GFA[79,84] linear model contains 11 parameters because further increase in the number of molecular descriptors does not lead to any considerable effects on the accuracy of the obtained model quantified by the corresponding MSE and squared correlation coefficient parameters. The final equation is presented as follows:

$$\begin{aligned} Y = {} & 3.78154(\mathrm{HNar}) - 3.61628(\mathrm{CIC0}) \\ & + 0.913525(\mathrm{EEig02d}) \\ & + 2.07643(\mathrm{BEHe3}) - 0.71207(\mathrm{SEigZ}) \\ & + 1.13819(\mathrm{E1s}) + 2.74026(\mathrm{nROH}) \\ & + 1.24098(\mathrm{Hy}) + 0.12090(\mathrm{MLOGP2}) \\ & + 0.64084(\mathrm{BLTD48}) - 4.5868(\mathrm{B01[C-F]}) \\ & + 17.60974 \end{aligned} \quad (16)$$

The numbers of digits of the reported coefficient values of eq 16 are generally in agreement with conventional computer algorithms using the GFA[79,84] models. In eq 16, the variables are as follows:

Y = one-component solubility parameter; HNar is the Narumi harmonic topological index. It is calculated as the number of nonhydrogen atoms divided by the reciprocal vertex degree summation. It has a positive effect on the values of solubility parameter.

CIC0 denotes the complementary information content (neighborhood symmetry of zero order). It has a negative effect on solubility parameter values.

EEig02d is the second eigenvalue of the "edge adjacency" matrix weighted by dipole moments. It somehow demonstrates the molecular interaction between adjacent bonds in a molecule. It has a positive effect on solubility parameter values;

BEHe3 stands for the highest third eigenvalue of the Burden matrix weighted by atomic Sanderson electronegativities. It is a measure of bond strength and polarity in a molecule. It has a positive effect on the values of the solubility parameter.

SEigZ is the eigenvalues of the Barysz distance matrix. It is a measure of molecular size. It indicates that molecular size has a negative effect on solubility parameter values.

E1s is the first component accessibility directional WHIM index weighted by atomic electrotopological states. It is a measure of molecular shape and atom distribution. It has a positive effect on the values of the solubility parameter.

nROH is the number of hydroxyl groups in a molecule. It might be referred to as the hydrogen-bonding effect in a molecule. It has a positive effect on the values of the solubility parameter.

Hy is called the "hydrophilicity" descriptor. It is defined as

$$\mathrm{Hy} = \frac{(1 + N_{\mathrm{Hy}}) \log_2(1 + N_{\mathrm{Hy}}) + \mathrm{nC}\frac{1}{\mathrm{nSK}} \log_2\left(\frac{1}{\mathrm{nSK}}\right) + \sqrt{\frac{N_{\mathrm{Hy}}}{\mathrm{nSK}^2}}}{\log_2(1 + \mathrm{nSK})} \quad (17)$$

where N_{Hy} is the number of hydrophilic groups (OH, SH, and NH groups), nC is the number of carbon atoms, and nSK is the number of hydrogen-excluded atoms. It is in fact a measure of hydrogen bonding and polarity in a molecule. It has a positive effect on the values of the solubility parameter.

MLOGP denotes the Moriguchi descriptor. It is calculated as follows:

$$\begin{aligned} \mathrm{MLOGP}^2 = {} & \Big(-1.014 + 1.244(F_{\mathrm{CX}})^{0.6} - 1.017(N_{\mathrm{O}} + N_{\mathrm{N}})^{0.9} \\ & + 0.406F_{\mathrm{PRX}} - 0.145N_{\mathrm{UNS}}{}^{0.8} + 0.511I_{\mathrm{HB}} + 0.268N_{\mathrm{POL}} \\ & - 2.215F_{\mathrm{AMP}} + 0.912I_{\mathrm{ALK}} - 0.392I_{\mathrm{RNG}} - 3.6847F_{\mathrm{QN}} \\ & + 0.474N_{\mathrm{NO_2}} + 1.582F_{\mathrm{NCS}} + 0.773I_{\beta L}\Big)^2 \end{aligned} \quad (18)$$

where F_{CX} is the summation of number of carbon and halogen atoms weighted by C = 1.0, F = 0.5, Cl = 0, Br = 1.5, and I = 2.0. $N_{\mathrm{O}} + N_{\mathrm{N}}$ is the total number of nitrogen and oxygen atoms. F_{PRX} is the proximity effect of N/O: 2 for X−Y and 1 for X−A−Y (X, Y = N and/or O; A = C, S, or P; − is saturated or unsaturated bond) with a correction (−1) for −CON< and $-SO_2N<$. N_{UNS} is the total number of unsaturated bonds (not those in NO_2). I_{HB} is the dummy variable for the presence of an intermolecular hydrogen bond such as *o*-OH and −CO−R, −OH and $-NH_2$, $-NH_2$ and −COOH, or 8-OH/NH_2 in quinolines, 5- or 8-OH/NH_2 in quinoxalines, etc. N_{POL} is the number of aromatic polar substituents (aromatic substituents excluding Ar−C(X)(Y)− and Ar−C(X)═C; X, Y = C and/or H). Upper limit = 4. F_{AMP} is the amphoteric property as follows: α-amino acid = 1, aminobenzoic acid = 0.5, and pyridinecarboxylic acid = 0.5. I_{ALK} is the dummy variable for alkane, alkene, cycloalkane, cycloalkene (hydrocarbons with 0 or 1 double bond), or hydrocarbon chains with at least seven carbon atoms. I_{RNG} is the dummy variable for the presence of ring structures except benzene and its condensed rings (aromatic, heteroaromatic, and hydrocarbon rings). F_{QN} is 1 for the quaternary nitrogen $>N^+<$ and is 0.5 for *N*-oxide. $N_{\mathrm{NO_2}}$ is the number of nitro groups, F_{NCS} is a parameter that its value for isothiocyanate (−N═C═S) is 1.0 and for thiocyanate (−S−C≡N) is 0.5. $I_{\beta\mathrm{L}}$ is the dummy variable for the presence of β-lactam. The described descriptor has a positive effect on solubility parameter values.

Also in eq 16, BLTD48 is defined as the Verhaar model of *Daphnia* baseline toxicity for *Daphnia* (48 h) from MLOGP (mmol/L). This descriptor and also the MLOGP parameter have a positive effect on solubility parameter values.

B01[C−F] indicates a binary descriptor that accounts for the presence or absence of C−F groups at topological distance of 1. It has a positive effect on the values of the solubility parameter.

The traditional statistical parameters of the developed linear model are as follows:

$$n_{\text{training}} = 1151, \quad n_{\text{test}} = 287, \quad R^2_{\text{training}} = 0.894, \quad R^2_{\text{test}} = 0.900, \quad \text{rms} = 1.2, \quad F = 875.9, \quad \text{LoF} = 1.38$$

where n_{training} and n_{test} are the numbers of compounds available in the training set and test set, respectively, and R^2_{training} and R^2_{test} = 0.900 are the squared correlation coefficients of the training set and test results, respectively; rms is the root-mean-square error of the model results in comparison with the experimental values;[78] and F is the F-ratio of the obtained GFA[79,84] linear equation (eq 16), which is defined as the ratio between the model summation of squares (MSS) and the residual summation of squares (RSS):[99]

$$F = \frac{\text{MSS}/\text{df}_\text{M}}{\text{RSS}/\text{df}_\text{E}} \tag{19}$$

where df_M and df_E refer to the degrees of freedom of the model and the overall error, respectively. It is a comparison between the model explained variance and the residual variance. It should be noted that high values of the F-ratio test indicate a high reliability of the models.

For internal validation of the model, the leave-one-out cross-validation technique has been initially used. Its corresponding parameter is normally calculated as follows:[49,65−69,72−74,76,77]

$$Q^2_{\text{LOO}} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_{ic})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{20}$$

where y_i is the solubility parameter for the ith compound, $\bar{y}$ is the mean value of the solubility parameter for all of the investigated compounds, and $\hat{y}_{ic}$ is the response of ith object represented/predicted by the obtained model ignoring the value of the related object (ith experimental solubility parameter). With the smallest absolute difference between this value and the R^2 parameter, the highest reliability of the model is expected. The evaluated leave-one-out cross-validation parameter of the obtained linear model is 0.891.

Another statistical parameter for internal validation of the QSPR linear model is the adjusted-R^2 parameter, which is defined as follows:[49,65−69,72−74,76,77]

$$R^2_{\text{adj}} = 1 - (1 - R^2)\left(\frac{n-1}{n-p'}\right) \tag{21}$$

where n is the number of experimental values and p' is the number of model parameters. With the smallest absolute difference between this value and the R^2 parameter, the highest reliability of the model is expected. The evaluated adjusted-R^2 parameter of the obtained linear model is 0.893.

As mentioned before, for testing the validity of the first developed model, we have used several validation techniques including the RQK constraint, the bootstrap technique, y-scrambling, and external validation techniques.[49,65−69,72−74,76,77]

Todeschini et al.[100] proposed that the following RQK constraints should be satisfied to avoid achieving chance correlations and to enhance the prediction capability of the proposed model:[49,65−69,72−74,76,77]

1. $\Delta K = K_{XY} - K_X > 0$ (quick rule)
2. $\Delta Q = Q^2_{\text{LOO}} - Q^2_{\text{ASYM}} > 0$ (asymptotic Q^2 rule)
3. $R^P > 0$ (redundancy RP rule)
4. $R^N > 0$ (overfitting RN rule)

K is calculated using following the equation

$$K = \frac{\sum_j \left| \left(\lambda_j / \sum_j \lambda_j - (1/p) \right) \right|}{2(p-1)/p}$$
$$j = 1, \ldots, p \text{ and } 0 \le K \le 1 \tag{22}$$

where λ values are the eigenvalues obtained from the correlation matrix of the data set $X(n,p)$, where n is the number of experimental data and p is the number of model parameters. K_{XY} and K_X are calculated using the set of the selected variables and the selected variables in addition to the solubility parameter values, respectively. The statistical parameters Q^2_{ASYM} and R^P are defined as follows:

$$Q^2_{\text{ASYM}} = 1 - (1 - R^2)\left(\frac{n}{n-p'}\right) \tag{23}$$

and

$$R^P = \prod_{j=1}^{p^+} \left(1 - M_j \left(\frac{p}{p-1}\right)\right) \quad M_j > 0 \text{ and } 0 \le R^P \le 1 \tag{24}$$

where M_j is defined as

$$M_j = \frac{R_{jy}}{R} - \frac{1}{p} \qquad -\frac{1}{p} \le M_j \le 1 - \frac{1}{p} \tag{25}$$

where R_{jy} and p are the correlation coefficient of the jth solubility parameter value and the number of variables of the model. It should be noted that the product in R^P runs over the p^+ variables, giving a positive difference M_j, while the sum in R^N runs over the p^- variables, giving a negative difference. It should be stated that $p^+ + p^- = p$, where

$$R^N = \sum_{j=1}^{p^-} M_j \tag{26}$$

The values of the preceding criteria have been calculated as ΔK = 0.026, ΔQ = 0.000, R^P = 0.019, and R^N = 0.000. The values of four constraints of the model are equal to or greater than zero, which proves the validity of the presented linear model and the fact that it is not a chance correlation.

To perform the bootstrap validation technique, a parameter Q^2_{boot} is defined as the average value of PRESS parameters calculated using repetition of constructing the training set and test set as below:

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{i/i})^2 \tag{27}$$

where $\hat{y}_{i/i}$ denotes the response of the ith represented/predicted solubility parameter value using the obtained model ignoring the

use of the ith experimental solubility parameter. It is a kind of a measure of the average prediction power of the obtained model. The bootstrapping has been repeated 5000 times. Consequently, the value Q^2_{boot} parameter of the obtained model has been evaluated to be 0.889.

The y-scambling validation technique is normally performed to prevent the final linear model from containing the independent variables, which are randomly selected as the parameters of the final equation. In order to do this test, calculation of the squared correlation coefficient of the obtained equation which randomly improves the sequence of the y vector is required (in this work, y is the value of the solubility parameter). In other words, by allocation to each object, a response is randomly chosen from true responses. Each scrambling is characterized in terms of the correlations of the scrambled response with the unperturbed data ($R^2_{yy'}$). The y-scrambling parameter is the intercept of the following equation:

$$Q_k^2 = a + br_k(y, \tilde{y}_k) \quad (28)$$

where Q_k^2 is the variance of the model obtained using the same predictors and the kth y-scrambled vector ($\tilde{y}_k$) and r_k is the correlation between the true response vector and the kth y-scrambled vector. In the case where the value of the intercept (a) is close to zero, it can be concluded that the obtained model is not a chance correlation.

The y-scrambling should be repeated hundreds of times (in this work 300 times). A high value of the intercept (a) indicates that the model is unstable. The value of intercept a has been calculated as −0.009 for the developed linear model.

Moreover, the external validation of the model has been demonstrated using the Q^2_{ext} parameter as follows:

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{n_{test}} (\hat{y}_{i/i} - y_i)^2}{\sum_{i=1}^{n_{test}} (y_i - \bar{y}_{training})^2} \quad (29)$$

where $\bar{y}_{training}$ is the average value of the solubility parameters of the compounds present in training set; $\hat{y}_{i/i}$ is the response of the ith object in training set represented/predicted by the obtained model ignoring the value of the related object (ith experimental solubility parameter). The less difference there is between this value and the R^2 parameter, the more validity of the model is expected. The evaluated Q^2_{ext} parameter of the obtained linear model is 0.896.

Figure 2 depicts the represented/predicted solubility parameters values by eq 16 versus experimental data.[78] Also, the absolute deviations of these results in comparison with the experimental values[78] have been better interpreted in Figure 3. All of the calculated/estimated solubility parameter values by eq 16, and the corresponding deviations from experimental values,[78] accompanied by the detailed allocation of the molecular descriptors in each organic compound have been extensively presented as Supporting Information.

The selected molecular descriptors have been later treated using two mathematical nonlinear methods including the ANN and the LSSVM[96] to develop more accurate and reliable models. First, a three-layer feed forward artificial neural network has been obtained for representation/prediction of the solubility parameters of the investigated compounds following the calculation procedure, as already described. For this purpose, several 3FFANN

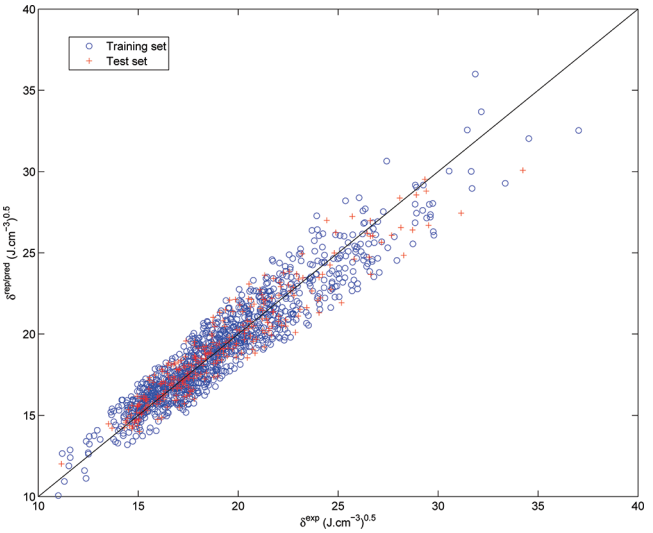

Figure 2. Comparison between represented/predicted results of the developed GFA model (eq 16) and experimental values[78] of solubility parameters of investigated chemical compounds.

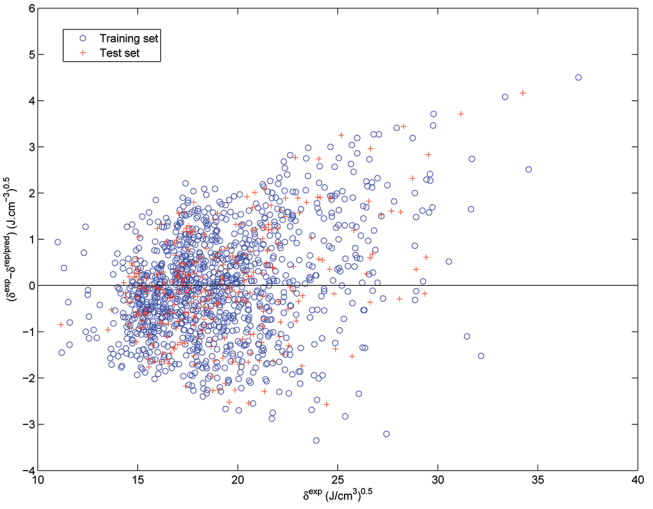

Figure 3. Deviations of the determined solubility parameter values of investigated chemical compounds by eq 16 from experimental values.[78]

modules have been generated assuming numbers 1−20 for n (number of neurons in the hidden layer). The most accurate results have been observed at n = 10 considering the probable problems of the ANNs including the overfitting and instability. Therefore, the developed three-layer FFANN has the structure of 11−10−1 (11 molecular descriptors have been regarded as the inputs of the algorithm).

Figures 4 and 5 show the values of determined solubility parameters and their absolute deviations using the developed QSPR-ANN model in comparison with the experimental values, respectively. The statistical parameters of this model have been reported in Table 1. As can be seen, the squared correlation coefficients, absolute average deviations (ADDs), standard deviation errors, and root-mean-square errors of the model over the training set, the optimization set, the test (prediction) set, and the main data set for the QSPR-ANN model are 0.940, 0.947,

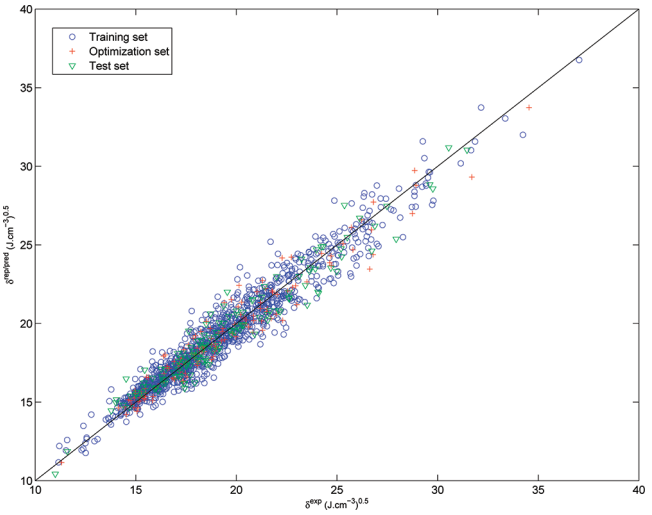

Figure 4. Comparison between represented/predicted results of the developed QSPR-FFANN model and experimental values[78] of solubility parameters of investigated chemical compounds.

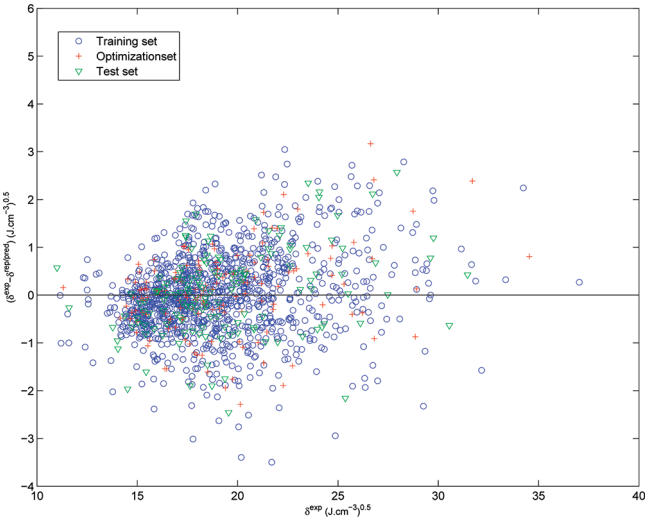

Figure 5. Deviations of the determined solubility parameter values of investigated chemical compounds by the QSPR-FFANN model from experimental values.[78]

0.947, 0.941, 3.3%, 3.3%, 3.6%, 3.4%, 3.4, 3.6, 3.6, 3.4, 0.9, 0.9, 0.9, and 0.9, respectively. It should be noted that, due to application of the large data set including 1438 experimental values, a number of model parameters have weak effects on the evaluated percent ADD of the results, effects that should be noted in reporting the deviations for the models developed based on few numbers of experimental data. All of the determined parameter values by the QSPR-ANN nonlinear model, the corresponding deviations from experimental values,[78] and the detailed allocation of the molecular descriptors in each organic compound have been extensively presented as Supporting Information.

To investigate the effects of reported advantages of the LSSVM[96] algorithms in comparison with the ANN method,[92−94] we have also applied the LSSVM[96] nonlinear computational procedure for developing the third model, as explained before. Similar to the FFANN model, the selected molecular descriptors

Table 1. Statistical Parameters of the QSPR-ANN Nonlinear Model[a]

statistical parameter	value
Training Set	
R^2	0.940
absolute average relative deviation,[b] %	3.3
standard deviation error	3.4
root-mean-square error	0.9
N[c]	1152
Validation Set	
R^2	0.947
absolute average relative deviation, %	3.3
standard deviation error	3.6
root-mean-square error	0.9
N	143
Test Set	
R^2	0.947
absolute average relative deviation, %	3.6
standard deviation error	3.6
root-mean-square error	0.9
N	143
Training + Validation + Test Set	
R^2	0.941
absolute average relative deviation, %	3.4
standard deviation error	3.4
root-mean-square error	0.9
N	1438

[a] R^2, squared correlation coefficient. [b] AAD = $[100/(N - n)]\{\sum_i^N [|\mathrm{rep}(i)/\mathrm{pred}(i) - \exp(i)|/\exp(i)]\}$, where n is the number of model parameters. [c] Number of data points.

by the GFA[79,84] model have been considered as the input parameters of the nonlinear LSSVM[96] strategy. The two main parameters of this algorithm are σ^2 and γ, which are supposed to be optimized using a proper optimization method. However, selecting the best optimization method for this purpose is still a challenge. To select the most efficient optimization method, the following characteristics of the corresponding algorithm should be taken into account:[1,101−111]

- ability to handle nondifferentiable, nonlinear, and multimodal cost functions
- no requirement of extensive problem formulation, while in traditional methods (such as integer programming, geometric programming, branch and bound methods, etc.) special mathematical formulation is necessary for solving a problem
- ease of use, i.e., few control variables to steer the minimization, and these variables should also be robust and easy to choose
- no sensitivity to starting point
- good convergence properties, i.e., consistent convergence to the global optimum in consecutive independent trials

Due to the preceding characteristics and the high nonlinearity of the SVM algorithm, application of nonpopulation based optimization methods such as the simplex simulated annealing algorithm (M-SIMPSA)[112] and Levenberg−Marquardt (LM)[83] may not be appropriate.

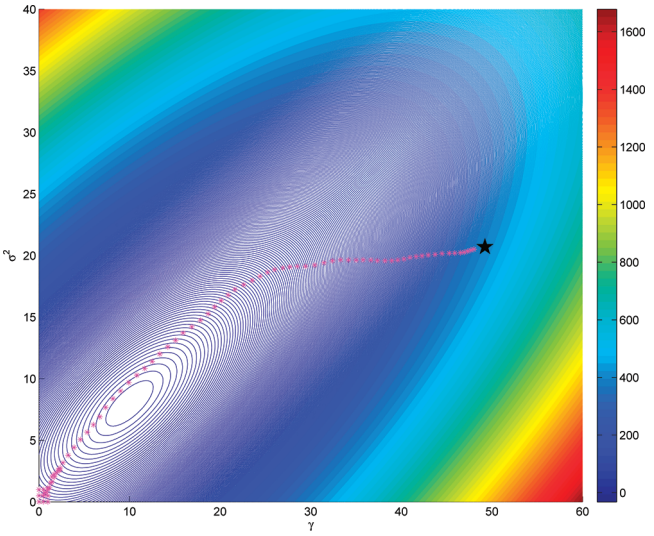

Figure 6. Contour route of the GA[85] optimization algorithm in LSSVM[96] mathematical approach: ∗, computational route to converge to the global optimum of the problem; ★, global optimum of the problem.

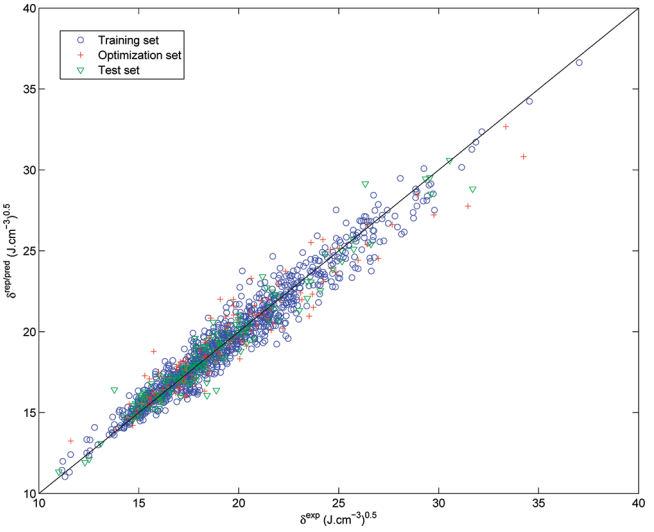

Figure 7. Comparison between the represented/predicted results of the developed QSPR-LSSVM model and experimental values[78] of solubility parameters of investigated chemical compounds.

In this work, we have modified the optimization part of the LSSVM[96] algorithm developed by Pelckmans et al.[97] and Suykens and co-workers[98] to use the robust genetic algorithm[85] method. This modification not only results in more quick computational steps but also results in no sensitivity to starting points as we are interested in. For this purpose, the GA optimization toolbox of MATLAB software was implemented, which is able to perform parallel computations. This feature may effectively decrease the required time of the optimization to converge to the global optimum. The number of populations of the optimization algorithm applied in this work was set to 1000. In order to ensure that the global optimum has been obtained, the optimization procedure was repeated several times.

The consequent results of a particular computational route to achieve the most probable optimum parameters of the LSSVM[96] model are shown in Figure 6. As can be inferred from this contour figure, the problem contains many local optimums (minimum/maximum). This fact shows the capability of the LSSVM[96] method to converge to the most probable global optimum even with many local optimums. The optimized values of the

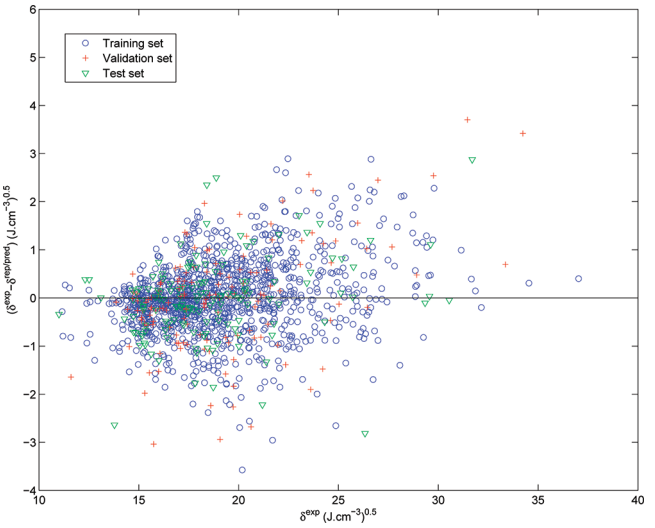

Figure 8. Absolute deviations of the determined solubility parameter values of investigated chemical compounds by the QSPR-LSSVM model from experimental values.[78]

Table 2. Statistical Parameters of the QSPR-LSSVM Nonlinear Model

statistical parameter	value
Training Set	
R^2	0.951
absolute average relative deviation, %	2.9
standard deviation error	0.8
root-mean-square error	0.8
N	1152
Validation Set	
R^2	0.923
absolute average relative deviation, %	4.0
standard deviation error	1.1
root-mean-square error	1.1
N	143
Test Set	
R^2	0.947
absolute average relative deviation, %	3.2
standard deviation error	0.8
root-mean-square error	0.8
N	143
Training + Validation + Test Set	
R^2	0.947
absolute average relative deviation, %	3.1
standard deviation error	0.0
root-mean-square error	0.8
N	1438

Table 3. Ranges and Average Absolute Deviations of the Obtained Results Using the Three Developed Models from Solubility Parameter Experimental Values[78] for Each Investigated Chemical Family

		$\delta^{rep/pred}$ $(J\cdot cm^{-3})^{0.5}$ using QSPR-GFA[79,84]			$\delta^{rep/pred}$ $(J\cdot cm^{-3})^{0.5}$ using QSPR-ANN			$\delta^{rep/pred}$ $(J\cdot cm^{-3})^{0.5}$ using QSPR-LSSVM[96]			
no.	family	AARD,[a] %	min value[b]	max value[c]	AARD, %	min value	max value	AARD, %	min value	max value	N
1	1-alkenes	3.2	16.20	17.28	2.0	14.99	17.32	1.8	15.20	17.28	17
2	2,3,4-alkenes	4.1	16.89	17.19	2.5	16.03	17.26	2.6	16.02	17.25	24
3	acetates	3.8	17.17	21.95	2.6	18.03	22.27	2.3	17.83	22.46	25
4	aldehydes	4.1	17.27	25.50	3.3	17.37	25.69	3.7	17.43	25.11	36
5	aliphatic ethers	6.6	15.83	18.83	3.1	15.27	18.06	3.4	15.71	18.14	34
6	alkylcyclohexanes	2.6	16.69	18.04	2.1	17.06	17.97	2.3	17.14	18.12	11
7	alkylcyclopentanes	1.5	16.79	17.92	1.2	16.86	17.65	1.4	16.79	17.69	7
8	alkynes	3.5	15.92	18.93	2.4	15.69	18.74	2.4	15.91	18.37	17
9	anhydrides	4.6	19.49	26.10	2.7	19.45	26.68	3.4	19.50	26.82	8
10	aromatic alcohols	5.8	17.50	24.21	4.9	17.57	25.29	4.2	17.52	24.37	27
11	aromatic amines	3.9	18.53	30.04	3.4	18.82	29.79	3.4	18.86	27.73	36
12	aromatic carboxylic acids	4.1	21.74	29.56	4.9	22.12	28.43	4.0	21.37	28.23	7
13	aromatic chlorides	6.0	20.26	23.09	2.4	19.91	22.05	2.8	20.25	21.97	13
14	aromatic esters	5.0	16.49	22.08	3.2	17.27	22.11	3.2	17.65	22.22	31
15	C, H, Br compounds	5.4	16.67	22.71	3.8	16.82	21.34	3.6	16.68	21.73	16
16	C, H, F compounds	6.4	12.12	18.04	3.8	12.94	19.58	3.2	14.01	19.85	18
17	C, H, I compounds	6.0	16.74	21.80	5.1	16.85	21.21	4.5	16.54	21.61	6
18	C, H, multihalogen compounds	5.3	14.98	23.07	3.6	15.79	21.95	3.1	15.91	22.23	17
19	C, H, NO_2 compounds	5.3	21.49	25.50	3.3	20.80	24.86	2.4	20.62	24.51	20
20	C1/C2 aliphatic chlorides	3.8	17.19	21.87	3.2	17.49	22.35	3.9	18.11	22.19	15
21	C3 and higher aliphatic chlorides	2.7	16.86	21.59	2.7	16.34	21.50	2.4	16.60	21.41	26
22	cycloaliphatic alcohols	8.3	20.57	26.24	4.6	19.24	24.90	2.4	18.59	24.85	4
23	cycloalkanes	6.9	18.68	19.08	3.1	17.94	18.47	3.5	17.92	18.83	5
24	cycloalkenes	5.5	18.20	19.93	3.3	17.75	19.28	3.8	18.06	19.42	9
25	dialkenes	5.6	16.54	20.80	3.8	15.69	20.59	3.6	15.68	20.46	27
26	dicarboxylic acids	6.7	27.43	31.08	4.5	25.41	33.00	4.2	25.00	31.81	11
27	dimethylalkanes	2.0	13.61	16.46	2.4	12.77	16.63	1.3	14.27	16.33	21
28	diphenyl/polyaromatics	4.4	18.22	20.75	2.9	18.15	20.66	2.6	17.72	20.38	19
29	elements	3.7	17.72	17.72	3.8	19.03	19.03	1.8	19.49	19.49	1
30	epoxides	6.4	20.11	24.76	4.9	18.89	25.16	3.7	19.74	24.17	11
31	ethyl and higher alkenes	1.5	15.21	16.37	1.5	14.71	16.42	1.3	14.98	16.37	12
32	formates	7.1	18.02	22.62	6.5	18.34	23.01	7.2	18.41	22.81	13
33	inorganic acids	6.2	16.71	33.56	3.8	16.46	37.76	5.2	16.46	34.15	8
34	inorganic gases	5.8	20.13	25.98	1.1	19.68	24.00	0.4	19.48	23.15	3
35	inorganic halides	5.3	18.04	21.48	5.5	17.65	21.33	4.2	18.81	20.81	4
36	isocyanates/diisocyanates	7.2	19.87	23.75	6.2	19.46	23.81	5.4	19.40	23.86	10
37	ketones	5.9	15.96	25.58	6.7	16.02	26.20	6.4	16.25	25.85	41
38	mercaptans	2.3	16.68	23.60	2.4	16.17	23.23	2.1	16.32	23.46	23
39	methylalkanes	1.7	14.70	16.45	1.3	13.65	16.70	0.7	14.49	16.57	17
40	methylalkenes	3.5	16.03	16.85	2.3	14.93	16.94	2.1	15.17	16.91	23
41	multiring cycloalkanes	1.7	18.28	18.28	1.0	18.15	18.15	5.4	18.77	18.77	1
42	*n*-alcohols	4.0	19.61	22.14	1.8	18.40	23.64	1.0	18.80	23.23	17
43	*n*-aliphatic acids	3.3	19.74	23.40	2.5	19.69	24.58	2.8	19.39	24.56	17
44	*n*-aliphatic primary amines	4.1	17.91	22.31	2.8	17.80	23.26	1.3	18.02	22.48	13
45	*n*-alkanes	3.7	13.40	18.32	1.8	12.87	16.90	1.7	14.20	17.07	30
46	*n*-alkylbenzenes	2.3	18.02	20.88	2.2	17.60	20.67	0.9	17.32	20.12	19
47	naphthalenes	3.6	18.20	20.49	3.1	18.35	20.20	2.5	17.91	20.24	17
48	nitriles	4.3	20.23	27.85	3.3	19.55	29.21	3.1	19.46	29.21	26
49	nitroamines	5.3	22.27	25.44	4.1	20.23	25.91	3.2	20.53	25.79	6
50	organic salts	5.5	18.35	27.06	4.3	19.01	26.37	3.9	19.02	27.27	14

Table 3. Continued

		$\delta^{rep/pred}$ $(J\cdot cm^{-3})^{0.5}$ using QSPR-GFA[79,84]			$\delta^{rep/pred}$ $(J\cdot cm^{-3})^{0.5}$ using QSPR-ANN			$\delta^{rep/pred}$ $(J\cdot cm^{-3})^{0.5}$ using QSPR-LSSVM[96]			
no.	family	AARD,[a] %	min value[b]	max value[c]	AARD, %	min value	max value	AARD, %	min value	max value	*N*
51	organic/inorganic compounds	8.0	15.60	22.51	7.4	16.80	23.25	6.3	16.85	23.02	6
52	other aliphatic acids	4.7	20.09	25.74	3.9	20.27	27.15	3.2	19.45	27.30	18
53	other aliphatic alcohols	5.4	19.19	22.92	3.0	19.32	26.16	2.9	19.38	24.14	37
54	other aliphatic amines	5.4	15.23	19.81	2.4	16.07	18.78	2.3	16.23	19.00	21
55	other alkanes	3.5	14.73	16.22	2.7	14.75	16.41	1.9	15.23	16.34	24
56	other alkylbenzenes	3.9	16.37	19.10	2.6	16.91	19.40	2.3	16.52	19.46	48
57	other amines, imines	6.2	17.92	27.18	4.6	16.79	26.34	4.1	17.24	25.76	29
58	other condensed rings	6.0	20.68	21.52	4.9	20.42	21.22	4.5	20.20	21.11	10
59	other ethers/diethers	5.2	17.02	22.11	5.0	18.05	22.20	4.4	17.57	21.87	22
60	other hydrocarbon rings	5.7	17.75	20.66	5.8	18.13	20.55	5.4	18.30	20.44	13
61	other monoaromatics	4.4	17.58	20.14	3.6	17.91	20.34	2.8	17.80	20.11	19
62	other polyfunctional C, H, O	6.2	16.46	26.31	3.8	17.22	28.47	3.8	17.33	27.26	47
63	other polyfunctional organics	5.6	21.00	22.79	3.8	19.91	22.98	2.7	19.95	22.31	3
64	other saturated aliphatic esters	5.1	15.94	24.85	2.9	17.28	25.29	2.7	17.16	25.41	22
65	peroxides	6.7	16.61	23.83	5.4	15.90	24.01	5.2	16.38	24.15	12
66	polyfunctional acids	6.0	24.20	37.00	3.7	23.91	34.73	3.2	23.99	32.61	16
67	polyfunctional amides/amines	5.7	20.81	31.64	4.1	20.09	32.59	4.0	20.42	31.03	22
68	polyfunctional C, H, N, halide, (O)	4.0	20.50	25.63	3.4	19.93	24.66	2.6	20.38	24.76	11
69	polyfunctional C, H, O, halide	4.3	17.24	29.37	2.5	16.35	29.58	2.4	17.06	31.81	36
70	polyfunctional C, H, O, N	6.4	21.72	28.14	4.4	22.46	29.77	3.7	22.45	29.14	19
71	polyfunctional C, H, O, S	3.4	21.25	29.04	3.8	21.62	29.38	3.1	21.20	30.17	8
72	polyfunctional esters	3.8	16.25	27.05	3.4	16.24	25.61	3.6	17.22	26.32	22
73	polyfunctional nitriles	4.0	22.68	28.35	3.3	23.34	30.62	5.0	22.72	29.53	4
74	polyols	6.3	23.63	33.03	2.7	21.87	34.73	2.3	23.03	33.79	25
75	propionates and butyrates	1.8	17.21	20.16	1.7	17.17	19.91	1.6	17.10	19.81	13
76	silanes/siloxanes	6.7	11.06	20.72	4.2	11.43	23.14	3.1	12.32	21.84	20
77	sulfides/thiophenes	2.7	16.49	23.65	2.6	15.63	23.23	2.5	16.12	23.18	46
78	terpenes	2.7	17.37	17.84	2.3	17.62	18.20	2.1	17.72	18.35	8
79	unsaturated aliphatic esters	4.2	16.35	21.94	3.4	17.31	22.07	3.4	17.37	21.79	24

[a] Refer to the DIPPR[78] database for observing the experimental solubility parameter values. [b] Minimum determined solubility parameter value for the compounds present in each chemical family. [c] Maximum determined solubility parameter value for the compounds present in each chemical family.

LSSVM[96] algorithm have been calculated as follows: $\gamma = 49.2305$ and $\sigma^2 = 20.6970$. The numbers of the reported digits of the two aforementioned parameters are normally obtained by sensitivity analysis of the overall errors of the optimization procedure.

The determined solubility parameters and their absolute deviations using the developed QSPR-LSSVM model in comparison with the experimental values[78] are reported in Figures 7 and 8, respectively. Moreover, the statistical parameters of this model are reported in Table 2. It has been found out that the squared correlation coefficients, absolute average relative deviations, standard deviation errors, and root-mean-square errors of the model over the training set, the validation (optimization) set, the test (prediction) set, and the main data set for the QSPR-LSSVM model are 0.951, 0.923, 0.947, 0.947, 2.9%, 4.0%, 3.2%, 3.1%, 0.8, 1.1, 0.8, 0, 0.8, 1.1, 0.8, and 0.8, respectively. All of the determined parameter values by the developed QSPR-LSSVM nonlinear model, the corresponding deviations from experimental values,[78] and the detailed allocation of the molecular descriptors in each organic compound have been extensively presented as Supporting Information. The *mat* files (MATLAB file format) of the three obtained models and the instructions for running the developed computer programs are freely available upon request to the authors (presented as computer programs).

Table 3 reports the absolute average deviations of the represented/predicted solubility parameter values of each investigated chemical family from corresponding experimental values.[78] It is inferred from the results of the two nonlinear models that, for the problem of interest in this work, the LSSVM[96] algorithm leads to a more accurate model in comparison with the model developed based on the FFAAN method. However, this may not be a general conclusion for similar problems. It merits further scrutiny for other scientific/engineering problems.

Careful investigation of Figures 3, 5, and 8 shows that the absolute deviations of the calculated/estimated solubility parameter values do not pursue similar trends by increasing the experimental values regarding the three developed models. This is mainly due to the different mathematical bases of these models. For instance, the GFA[85] linear model leads the absolute deviations to increase linearly in comparison with the experimental values because this model has been developed based on solutions of a system of linear equations as explained before. However, the aforementioned trend is different for the

QSPR-ANN and QSPR-LSSVM nonlinear models, which depend on the transfer function of the neural network method and the radial basis of the support vector machine method,[96] respectively.

Another element to consider is that we have tried to use the largest available data set[78] for the solubility parameter values in the literature. However, the existing difficulties and possible errors in experimental measurements, which lead to uncertainties in obtained experimental values (refer to the Supporting Information for observing the uncertainties of the data for each data source), may contribute to decreasing the accuracy and prediction capability of the developed models. This would literally demonstrate the need to design more accurate experimental apparatus and produce reliable experimental data to provide the industry with more accurate and predictive models.

4. CONCLUSION

Three molecular models were presented for representation/prediction of the solubility parameters of nonelectrolyte organic compounds at 298.15 K and atmospheric pressure. A large data set regarding the solubility parameter values of nonelectrolyte organic compounds (DIPPR 801)[78] was applied to develop and validate the proposed models. The genetic function approximation mathematical method[79,84] was used to select the most appropriate model parameters from a domain of the parameters (1438). The required parameters of the linear model are the numbers of 11 molecular descriptors in each investigated molecule. Two mathematical tools were applied to present more accurate nonlinear models including three-layer feed forward artificial neural networks (3FFANN) and least-squares support vector machine (LSSVM).[96] Furthermore, the Levenberg–Marquardt (LM)[83] and the genetic algorithm (GA)[85] optimization methods were respectively implemented to optimize the 3FFANN and LSSVM[96] models. The statistical parameters of the obtained models show that they are reliable, comprehensive, and predictive models to represent/predict the one-component solubility parameters of nonelectrolyte organic compounds, which are especially applied in the chemical, petroleum, and polymer industries. The results also prove that the LSSVM[96] algorithm leads to presenting a more accurate QSPR model than the FFANN method. Another issue to point out is the effect of uncertainties of the experimental data, applied for developing the model, on the obtained predicted results that cannot be ignored. The more accuracy there is in the experimental measurements of the data, the more accuracy of the developed model is expected.

■ ASSOCIATED CONTENT

Supporting Information. Detailed results of the three developed models in XLS format. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: amir-hossein.mohammadi@mines-paristech.fr. Tel.: + (33) 1 64 69 49 70. Fax: + (33) 1 64 69 49 68.

■ ACKNOWLEDGMENT

A.E. wishes to thank MINES ParisTech for providing him a Ph.D. scholarship.

■ REFERENCES

(1) Eslamimanesh, A.; Esmaeilzadeh, F. Estimation of solubility parameter by the modified ER equation of state. *Fluid Phase Equilib.* **2010**, *291*, 141–150.

(2) Scatchard, G. Equilibria in Non-electrolyte solutions in relation to the vapor pressures and densities of the components. *Chem. Rev.* **1931**, *8*, 321–333.

(3) Hildebrand, J. H.; Scott, R. L. *The Solubility of Nonelectrolytes*; Rheinhold Publishing Corp.: New York, 1950.

(4) Hansen, C. M. The three dimensional solubility parameter—key to paint component affinities I.—Solvents, plasticizers, polymers, and resins. *J. Paint Technol.* **1967**, *39*, 104–117.

(5) Hansen, C. M. The Universality of the Solubility Parameter. *Ind. Eng. Chem. Prod. Res. Dev.* **1969**, *8*, 2–11.

(6) Gharagheizi, F.; Sattari, M.; Angaji, M. T. Effect of Calculation Method on Values of Hansen Solubility Parameters of Polymers. *Polym. Bull.* **2006**, *57*, 377–384.

(7) Hildebrand, J. H. Solubility: XII. Regular Solutions. *J. Am. Chem. Soc.* **1929**, *51*, 66–80.

(8) Mousavi-Dehghani, S. A.; Mirzayi, B.; Mousavi, S. M. H.; Fasih, M. An applied and efficient model for asphaltene precipitation in production and miscible gas injection. *Pet. Sci. Technol.* **2010**, *28*, 113–124.

(9) Novosad, Z.; Costain, T. G. Experimental and modeling studies of asphaltene equilibria for a reservoir under CO_2 injection. *Proceedings of the 65th Annual Technical Conference and Exhibition of the SPE, New Orleans, LA, USA*; 1990; Paper SPE 20530, Richardson, TX.

(10) Nor-Azian, N.; Adewumi, M. A. Development of asphaltene phase equilibrium predictive model. *Proceedings of the Eastern Regional Conference and Exhibition of the SPE, Richardson, TX, USA*; 1993; Paper SPE 26905, Richardson, TX.

(11) MacMillan, D. J.; Tackett, J. E.; Jessee, M. A.; Monger-McClure, T. G. A unified approach to asphaltene precipitation: laboratory measurement and modeling. *Proceedings of the SPE International Symposium on Oilfield Chemistry, San Antonio, TX, USA*; 1995; Paper SPE 28990, Richardson, TX.

(12) Yang, Z.; Ma, C. F.; Lin, S. X.; Yang, J. T.; Guo, T. M. Experimental and modeling studies on the asphaltene precipitation in degassed and gas-injected reservoir oils. *Fluid Phase Equilib.* **1999**, *157*, 143–158.

(13) Alboudwarej, H.; Akbarzadeh, K.; Beck, J.; Svercek, W. Y.; Yarranton, H. W. Regular solution model of asphaltene precipitation from bitumen. *AIChE J.* **2003**, *11*, 2948–2956.

(14) Akbarzadeh, K.; Alboudwarej, H.; Svercek, H. Y.; Yarranton, H. W. A generalized regular solution model for the prediction of asphaltene precipitation from n-alkane diluted heavy oils and bitumens. *Fluid Phase Equilib.* **2005**, *232*, 159–170.

(15) Yarranton, H. W.; Masliyah, J. H. Molar mass distribution and solubility modeling of asphaltenes. *AIChE J.* **1996**, *42*, 3533–3543.

(16) Mohammadi, A. H.; Richon, D. The Scott-Magat polymer theory for determining onset of precipitation of dissolved asphaltene in the solvent + precipitant solution. *Open Thermodyn. J.* **2008**, *2*, 13–16.

(17) Prausnitz, J. M.; Lichtenthaler, R. N.; de Azevedo, E. G. *Molecular Thermodynamics of Fluid Phase Equilibria*, 3rd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 1999.

(18) Van Arkel, A. E. Mutual solubility of liquids. *Trans. Faraday Soc.* **1946**, *42B*, 81–84.

(19) Small, P. A. Some factors affecting the solubility of polymers. *J. Appl. Chem.* **1953**, *3*, 71–80.

(20) Anderson, R. Polar Organic Solvents and Aromatic Hydrocarbons. Ph.D. Thesis, Department of Chemical Engineering, University of California, Berkeley, 1961.

(21) Blanks, R. F.; Prausnitz, J. M. Thermodynamics of polymer solubility in polar and nonpolar systems. *Ind. Eng. Chem. Fundam.* **1964**, *3*, 1–8.

(22) Prausnitz, J. M.; Anderson, R. Thermodynamics of solvent selectivity in extractive distillation of hydrocarbons. *AIChE J.* **1961**, *7*, 96–101.

(23) Weimer, R. F.; Prausnitz, J. M. Complex formation between carbon tetrachloride and aromatic hydrocarbons. *J. Chem. Phys.* **1965**, *42*, 3643–3644.
(24) Prausnitz, J. M.; Shair, F. H. A thermodynamic correlation of gas solubilities. *AIChE J.* **1961**, *7*, 682–687.
(25) Lyckman, E. W.; Eckert, C. A.; Prausnitz, J. M. Generalized liquid volumes and solubility parameters for regular solution application. *Chem. Eng. Sci.* **1965**, *20*, 703–706.
(26) Fedors, R. F. A method for estimating both the solubility parameter and molar volumes of liquids. *Polym. Eng. Sci.* **1974**, *14*, 147–154.
(27) Carnahan, N. F.; Starling, K. E. Equation of state for nonattracting rigid spheres. *J. Chem. Phys.* **1969**, *51*, 635–636.
(28) Lozada, C. M.; Monfort, J. P.; Del Río, F. Calculation of solubility parameters from an equation of state. *Chem. Phys. Lett.* **1977**, *45*, 130–133.
(29) Hansen, C. M.; Beerbower, A. Solubility parameters. In *Kirk-Othmer Encyclopedia of Chemical Technology, Supplementary Volume*,2nd ed.; Standen, A., Ed.; Interscience: New York, 1971.
(30) Beerbower, A. Environmental Capability of Liquids. In *Interdisciplinary Approach to Liquid Lubricant Technology*; NASA Publication SP-318; 1973.
(31) Van Krevelen, D. W.; Hoftyzer, P. J. *Properties of Polymers: Their Estimation and Correlation with Chemical Structure*, 2nd ed.; Elsevier: Amsterdam, 1976.
(32) Allada, S. R. Solubility parameters of supercritical fluids. *Ind. Eng. Chem. Prod. Res. Dev.* **1984**, *23*, 344–348.
(33) Panayiotou, C. Solubility parameter revisited: An equation-of-state approach for its estimation. *Fluid Phase Equilib.* **1997**, *131*, 21–35.
(34) Williams, L. L.; Rubin, J. B.; Edwards, H. W. Calculation of Hansen Solubility Parameter Values for a Range of Pressure and Temperature Conditions, Including the Supercritical Fluid Region. *Ind. Eng. Chem. Res.* **2004**, *43*, 4967–4972.
(35) Bozdogan, A. E. A method for determination of thermodynamic and solubility parameters of polymers from temperature and molecular weight dependence of intrinsic viscosity. *Polymer* **2004**, *45*, 6415–6424.
(36) Utracki, L. A.; Simha, R. Statistical thermodynamics predictions of the solubility parameter. *Polym. Int.* **2004**, *53*, 279–286.
(37) Stefanis, E.; Tsivintzelis, I.; Panayiotou, C. The partial solubility parameters: An equation-of-state approach. *Fluid Phase Equilib.* **2006**, *240*, 144–154.
(38) Zeng, Z.-Y.; Xu, Y.-Y.; Li, Y.-W. Calculation of Solubility Parameter Using Perturbed-Chain SAFT and Cubic-Plus-Association Equations of State. *Ind. Eng. Chem. Res.* **2008**, *47*, 9663–9669.
(39) Code, J. E.; Holder, A. J.; Eick, J. D. Direct and Indirect Quantum Mechanical QSPR Hildebrand Solubility Parameter Models. *QSAR Comb. Sci.* **2008**, *27*, 841–849.
(40) Vargas, F. M.; Chapman, W. G. Application of the One-Third rule in hydrocarbon and crude oil systems. *Fluid Phase Equilib.* **2010**, *290*, 103–108.
(41) Esmaeilzadeh, F.; Roshanfekr, M. A new cubic equation of state for reservoir fluids. *Fluid Phase Equilib.* **2006**, *239*, 83–90.
(42) Bonyadi, M.; Esmaeilzadeh, F. A modification of the alpha function (α), and the critical compressibility factor (ζ_c) in ER (Esmaeilzadeh−Roshanfekr) equation of state. *Fluid Phase Equilib.* **2008**, *273*, 31–37.
(43) Gharagheizi, F.; Eslamimanesh, A.; Mohammadi, A. H.; Richon, D. Use of artificial neural network-group contribution method to determine surface tension of pure compounds. *J. Chem. Eng. Data* **2011**, *56*, 2587–2601.
(44) Gharagheizi, F.; Sattari, M. Prediction of triple-point temperature of pure components using their chemical structures. *Ind. Eng. Chem. Res.* **2010**, *49*, 929–932.
(45) Gharagheizi, F.; Abbasi, R.; Tirandazi, B. Prediction of Henry's law constant of organic compounds in water from a new group-contribution-based model. *Ind. Eng. Chem. Res.* **2010**, *49*, 10149–10152.
(46) Eslamimanesh, A.; Gharagheizi, F.; Mohammadi, A. H.; Richon, D. Artificial neural network modeling of solubility of supercritical carbon dioxide in 24 commonly used ionic liquids. *Chem. Eng. Sci.* **2011**, *66*, 3039–3044.
(47) Gharagheizi, F.; Eslamimanesh, A.; Mohammadi, A. H.; Richon, D. Artificial neural network modeling of solubilities of 21 mostly-used industrial solid compounds in supercritical carbon dioxide. *Ind. Eng. Chem. Res.* **2011**, *50*, 221–226.
(48) Gharagheizi, F.; Eslamimanesh, A.; Mohammadi, A. H.; Richon, D. Representation/prediction of solubilities of pure compounds in water using artificial neural network−group contribution method. *J. Chem. Eng. Data* **2011**, *56*, 720–726.
(49) Gharagheizi, F.; Eslamimanesh, A.; Mohammadi, A. H.; Richon, D. QSPR approach for determination of parachor of non-electrolyte organic compounds. *Chem. Eng. Sci.* **2011**, *66*, 2959–2967.
(50) Gharagheizi, F.; Eslamimanesh, A.; Mohammadi, A. H.; Richon, D. Determination of parachor of various compounds using an artificial neural network-group contribution method. *Ind. Eng. Chem. Res.* **2011**, *50*, 5815–5823.
(51) Chouai, A.; Laugier, S.; Richon, D. Modeling of thermodynamic properties using neural networks: Application to refrigerants. *Fluid Phase Equilib.* **2002**, *199*, 53–62.
(52) Piazza, L.; Scalabrin, G.; Marchi, P.; Richon, D. Enhancement of the extended corresponding states techniques for thermodynamic modelling. I. Pure fluids. *Int. J. Refrig.* **2006**, *29*, 1182–1194.
(53) Scalabrin, G.; Marchi, P.; Bettio, L.; Richon, D. Enhancement of the extended corresponding states techniques for thermodynamic modelling. II. Mixtures. *Int. J. Refrig.* **2006**, *29*, 1195–1207.
(54) Chapoy, A.; Mohammadi, A. H.; Richon, D. Predicting the hydrate stability zones of natural gases using Artificial Neural Networks. *Oil Gas Sci. Technol.* **2007**, *62*, 701–706.
(55) Mohammadi, A. H.; Richon, D. Hydrate phase equilibria for hydrogen + water and hydrogen + tetrahydrofuran + water systems: Predictions of dissociation conditions using an artificial neural network algorithm. *Chem. Eng. Sci.* **2010**, *65*, 3352–3355.
(56) Mohammadi, A. H.; Richon, D. Estimating sulfur content of hydrogen sulfide at elevated temperatures and pressures using an artificial neural network algorithm. *Ind. Eng. Chem. Res.* **2008**, *47*, 8499–8504.
(57) Mohammadi, A. H.; Richon, D. A Mathematical model based on artificial neural network technique for estimating liquid water−hydrate equilibrium of water−hydrocarbon System. *Ind. Eng. Chem. Res.* **2008**, *47*, 4966–4970.
(58) Mohammadi, A. H.; Afzal, W.; Richon, D. Determination of critical properties and acentric factors of petroleum fractions using artificial neural networks. *Ind. Eng. Chem. Res.* **2008**, *47*, 3225–3232.
(59) Mohammadi, A. H.; Richon, D. Use of artificial neural networks for estimating water content of natural gases. *Ind. Eng. Chem. Res.* **2007**, *46*, 1431–1438.
(60) Mohammadi, A. H.; Martínez-López, J. F.; Richon, D. Determining phase diagrams of tetrahydrofuran+methane, carbon dioxide or nitrogen clathrate hydrates using an artificial neural network algorithm. *Chem. Eng. Sci.* **2010**, *65*, 6059–6063.
(61) Mehrpooya, M.; Mohammadi, A. H.; Richon, D. Extension of an Artificial Neural Network algorithm for estimating sulfur content of sour gases at elevated temperatures and pressures. *Ind. Eng. Chem. Res.* **2010**, *49*, 439–442.
(62) Mohammadi, A. H.; Belandria, V.; Richon, D. Use of an artificial neural network algorithm to predict hydrate dissociation conditions for hydrogen + water and hydrogen + tetra-n-butyl ammonium bromide + water systems. *Chem. Eng. Sci.* **2010**, *65*, 4302–4305.
(63) Gharagheizi, F. A new group contribution-based method for estimation of lower flammability limit of pure compounds. *J. Hazard. Mater.* **2009**, *170*, 595–604.
(64) Gharagheizi, F. New Neural Network Group Contribution model for estimation of lower flammability limit temperature of pure compounds. *Ind. Eng. Chem. Res.* **2009**, *48*, 7406–7416.
(65) Gharagheizi, F.; Sattari, M. Estimation of molecular diffusivity of pure chemicals in water: A quantitative structure-property relationship study. *SAR QSAR Environ. Res.* **2009**, *20*, 267–285.
(66) Gharagheizi, F. Prediction of standard enthalpy of formation of pure compounds using molecular structure. *Aust. J. Chem.* **2009**, *62*, 376–381.

(67) Gharagheizi, F.; Tirandazi, B.; Barzin, R. Estimation of aniline point temperature of pure hydrocarbons: A Quantitative Structure-Property Relationship approach. *Ind. Eng. Chem. Res.* **2009**, *48*, 1678–1682.

(68) Gharagheizi, F.; Mehrpooya, M. Prediction of some important physical properties of sulfur compounds using QSPR models. *Mol. Diversity* **2008**, *12*, 143–155.

(69) Sattari, M.; Gharagheizi, F. Prediction of Molecular Diffusivity of Pure Components into Air: A QSPR Approach. *Chemosphere* **2008**, *72*, 1298–1302.

(70) Gharagheizi, F.; Alamdari, R. F.; Angaji, M. T. A new neural network-group contribution method for estimation of flash point. *Energy Fuels* **2008**, *22*, 1628–1635.

(71) Gharagheizi, F.; Fazeli, A. Prediction of Watson Characterization Factor of Hydrocarbon Compounds from Their Molecular Properties. *QSAR Comb. Sci.* **2008**, *27*, 758–767.

(72) Gharagheizi, F.; Alamdari, R. F. A molecular-based model for prediction of solubility of C60 fullerene in various solvents. *Fullerenes, Nanotubes, Carbon Nanostruct.* **2008**, *16*, 40–57.

(73) Gharagheizi, F. A New Neural Network Quantitative Structure-Property Relationship for prediction of θ (Lower Critical Solution Temperature) of polymer solutions. *e-Polym.* **2007**, *114*, 1–5.

(74) Gharagheizi, F. QSPR studies for solubility parameter by means of genetic algorithm-based multivariate linear regression and generalized regression neural network. *QSAR Comb. Sci.* **2008**, *27*, 165–170.

(75) Gharagheizi, F. A chemical structure-based model for estimation of upper flammability limit of pure compounds. *Energy Fuels* **2010**, *27*, 3867–3871.

(76) Vatani, A.; Mehrpooya, M.; Gharagheizi, F. Prediction of standard enthalpy of formation by a QSPR Model. *Int. J. Mol. Sci.* **2007**, *8*, 407–432.

(77) Mehrpooya, M.; Gharagheizi, F. A molecular approach for prediction of sulfur compounds solubility parameters. *Phosphorus, Sulfur Silicon Relat. Elem.* **2009**, *185*, 204–210.

(78) Project 801, Evaluated Process Design Data, Public Release Documentation, Design Institute for Physical Properties (DIPPR), American Institute of Chemical Engineers (AIChE), 2006.

(79) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics: Vol. I: Alphabetical Listing/Vol. II: Appendices, References*, 2nd revised and enlarged ed.; Wiley-VCH: Weinheim, Germany, 2009.

(80) Talete srl, Dragon for Windows (Software for Molecular Descriptor Calculation), version 5.5, 2006.

(81) Milano chemometrics and QSAR research group.

(82) Hyperchem Release 7.5 for Windows, Molecular Modeling System, Hypercube, Inc., 2002.

(83) Draper, N.; Smith, H. *Applied Regression Analysis*; John Wiley & Sons: New York, 1998.

(84) Rogers, D.; Hopfinger, A. J. Application of genetic function approximation to quantitative structure−activity relationships and quantitative structure−property relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.

(85) Holland, J. H. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: Ann Arbor, MI, USA, 1975.

(86) Shi, L. M.; Fan, Y.; Myers, T. G.; Paull, K. D.; Weinstein, J. N. J. Mining the NCI anticancer drug discovery databases: Genetic function approximation for the QSAR study of anticancer ellipticine analogues. *Chem. Inf. Comput. Sci.* **1998**, *38*, 189–199.

(87) Nargotra, A.; Koul, S.; Sharma, S.; Khan, I. A.; Kumar, A.; Thota, N.; Koul, J. L.; Taneja, S. C.; Qazi, G. N. Quantitative structure−activity relationship (QSAR) of aryl alkenyl amides/imines for bacterial efflux pump inhibitors. *Eur. J. Med. Chem.* **2009**, *44*, 229–238.

(88) Bhattacharya, P.; Leonard, J. T.; Roy, K. Exploring QSAR of thiazole and thiadiazole derivatives as potent and selective human adenosine A3 receptor antagonists using FA and GFA techniques. *Bioorg. Med. Chem.* **2005**, *13*, 1159–1165.

(89) Roy, K.; Leonard, J. T. QSAR by LFER model of cytotoxicity data of anti-HIV 5-phenyl-1-phenylamino-1H-imidazole derivatives using principal component factor analysis and genetic function approximation. *Bioorg. Med. Chem.* **2005**, *13*, 2967–2973.

(90) Deswal, S.; Roy, N. Quantitative structure activity relationship of benzoxazinone derivatives as neuropeptide Y Y5 receptor antagonists. *Eur. J. Med. Chem.* **2006**, *41*, 552–557.

(91) Gharagheizi, F. QSPR analysis for intrinsic viscosity of polymer solutions by means of GA-MLR and RBFNN. *Comput. Mater. Sci.* **2007**, *40*, 159.

(92) Liu, H.; Yao, X.; Zhang, R.; Liu, M.; Hu, Z.; Fan, B. Accurate quantitative structure-property relationship model to predict the solubility of C_{60} in various solvents based on a novel approach using a least-squares support vector machine. *J. Phys. Chem. B* **2005**, *109*, 20565–20571.

(93) Yao, X.; Liu, H.; Zhang, R.; Liu, M.; Hu, Z.; Panaye, A.; Doucet, J. P.; Fan, B. QSAR and classification study of 1,4-dihydropyridine calcium channel antagonists based on least squares support vector machines. *Mol. Pharmaceutics* **2005**, *5*, 348–356.

(94) Manallack, D. T.; Livingstone, D. J. Neural networks in drug discovery; Have they Lived up to their promise? *Eur J. Med. Chem.* **1999**, *34*, 195–208.

(95) Gunn, S. R.; Brown, M.; Bossley, K. M. Network performance assessment for neurofuzzy data modeling. *Lect. Notes Comput. Sci.* **1997**, *1280*, 313–323.

(96) Suykens, J. A. K.; Vandewalle, J. Least squares support vector machine classifiers. *Neural Process. Lett.* **1999**, *9*, 293–300.

(97) Pelckmans, K.; Suykens, J. A. K.; Van Gestel, T.; De Brabanter, D.; Lukas, L.; Hamers, B.; De Moor, B.; Vandewalle, J. *LS-SVMlab: a Matlab/C Toolbox for Least Squares Support Vector Machines*; Internal Report 02-44, ESATSISTA; K. U. Leuven: Leuven, Belgium, 2002.

(98) Suykens, J. A. K.; Van Gestel, T.; De Brabanter, J.; De Moor, B.; Vandewalle, J. *Least Squares Support Vector Machines*; World Scientific: Singapore, 2002.

(99) Krzanowski, W. J. *Principles of Multivariate Analysis: A User's Perspective*; Oxford University Press: New York, 1988.

(100) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. Detecting "bad" Regression Models: Multicriteria Fitness Functions in Regression Analysis. *Anal. Chim. Acta* **2004**, *515*, 199–208.

(101) Price, K.; Storn, R. Differential Evolution. *Dr. Dobb's J.* **1997**, *22*, 18–24.

(102) Chiou, J. P.; Wang, F. S. Hybrid method of evolutionary algorithms for static and dynamic optimization problems with applications to a fed-batch fermentation process. *Comput. Chem. Eng.* **1999**, *23*, 1277–1291.

(103) Schwefel, H. P. *Numerical Optimization of Computer Models*; John Wiley & Sons: New York, 1981.

(104) Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley: Reading, MA, USA, 1989.

(105) Davis, L. *Handbook of Genetic Algorithms*; Van Nostrand Reinhold: New York, 1991.

(106) Storn, R. Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces. *J. Global Optim.* **1997**, *11*, 341–359.

(107) Eslamimanesh, A. A Semicontinuous Thermodynamic Model for Prediction of Asphaltene Precipitation in Oil Reservoirs. M.Sc. Thesis, Shiraz University, Shiraz, Iran, 2009 (in Persian).

(108) Eslamimanesh, A.; Shariati, A. A Semicontinuous thermodynamic model for prediction of asphaltene precipitation. Presented at the VIII Iberoamerican Conference on Phase Equilibria and Fluid Properties for Process Design (Equifase), Praia da Rocha, Portugal, October 2009.

(109) Belandria, V.; Eslamimanesh, A.; Mohammadi, A. H.; Richon, D. Gas hydrate formation in carbon dioxide + nitrogen + water system: compositional analysis of equilibrium phases. *Ind. Eng. Chem. Res.* **2011**, *50*, 4722–4730.

(110) Yazdizadeh, M.; Eslamimanesh, A.; Esmaeilzadeh, F. Thermodynamic modeling of solubilities of various solid compounds in supercritical carbon dioxide: Effects of equations of state and mixing rules. *J. Supercrit. Fluids* **2011**, *55*, 861–875.

(111) Belandria, V.; Eslamimanesh, A.; Mohammadi, A. H.; Théveneau, P.; Legendre, H.; Richon, D. Compositional analysis and hydrate dissociation conditions measurements for carbon dioxide + methane + water system. *Ind. Eng. Chem. Res.* **2011**, *50*, 5783–5794.

(112) Cardoso, M. F.; Salcedo, R. L.; Feyo de Azevedo, S.; Barbosa, D. A simulated annealing approach to the solution of minlp problems. *Comput. Chem. Eng.* **1997**, *21*, 1349–1364.