Available online at www.sciencedirect.com
ScienceDirect
Chemosphere 66 (2007) 2067–2076
CHEMOSPHERE
www.elsevier.com/locate/chemosphere

QSPR model of Henry's law constant for a diverse set of organic chemicals based on genetic algorithm-radial basis function network approach

Hassan Modarresi [a], Hamid Modarress [a], John C. Dearden [b,*]

[a] Department of Chemical Engineering, Amirkabir University of Technology (Tehran Polytechnic), 424 Hafez Avenue, Tehran, Iran
[b] School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK

Received 31 January 2006; received in revised form 22 September 2006; accepted 25 September 2006
Available online 20 November 2006

Abstract

Six quantitative structure–property relationship (QSPR) models for a diverse set of experimental data of Henry's law constant (H) of organic chemicals under environmental condition ($T = 25$ °C; water–air system) have been developed based on four different molecular descriptor sets. Three different models based on the descriptors of CODESSA (Comprehensive Descriptors for Structural and Statistical Analysis), Tsar, and Dragon software and a model based on a combined descriptor set from these packages, and in addition from HYBOT software, have been established using the stepwise regression method. The combined descriptors set model gave the best results. Furthermore, a genetic algorithm was used for descriptor selection from a combined set of descriptors, and a radial basis function network was utilized to establish a model with a low root mean square error (RMSE). The results of this study were compared with the well-known bond contribution and group contribution methods. The group contribution method failed to predict Henry's law constant of 170 from all 940 compounds in the data-set. RMSEs of 0.693, 0.798, and 0.564 were achieved for bond contribution, group contribution and the best QSPR model of this study, respectively, based on logarithm of H. Analysis of different QSPR models showed that hydrogen bonding between the organic solute and water as a solvent has the greatest influence on this partitioning phenomenon.

Keywords: Henry's law constant; QSPR; Molecular descriptor; Solvation free energy; Hydrogen bonding; Multi-linear regression

1. Introduction

As the air–water partition coefficient, Henry's law constant (H) represents a key physical property of a compound with respect to its distribution and fate in the environment as well as to the applicability of potential treatment methods such as air-stripping for treatment of contaminated ground water (Staudinger and Roberts, 1996; Baker et al., 2002). However, according to the latest literature, directly measured H data are known for a relatively small number of organic chemicals (fewer than 1200 compounds) out of the over 70,000 that are in current use (Brennan et al., 1998). Therefore, the need for additional experimental data on the one hand and for validated estimation methods for this property on the other hand is crucial. The estimation methods for H for environmental purposes can be categorized as (1) property–property relationships (PPR) methods; (2) bond and group contribution methods; (3) continuum-solvation methods; (4) UNIFAC (universal quasi-chemical functional group activity coefficient) and structural, quantum chemical or physicochemical descriptor-based quantitative structure–property relationships (QSPR) methods.

The most well known PPR is the VP/AS (vapor pressure/aqueous solubility) method (Mackay et al., 2000). Although for compounds with low solubility and low vapor pressure measured at a desired temperature, the VP/AS method, also called indirectly measured H (Mackay and Shiu, 1981), has

* Corresponding author. Tel.: +44 151 231 2066; fax: +44 151 231 2170.
E-mail address: J.C.Dearden@ljmu.ac.uk (J.C. Dearden).

0045-6535/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.chemosphere.2006.09.049

excellent results, the above-mentioned ratio from calculated values of vapor pressure and aqueous solubility can lead to large errors (Cramer, 1980). The first bond contribution method (Hine and Mookerjee, 1975) has been improved by expanding the number of bond definitions from 34 to 59 and with 15 correction factors (Meylan and Howard, 1991) and finally, in recent revisions, the bond method contains 64 bond definitions and 57 correction factors, whilst the group contribution method contains 93 group definitions (Meylan and Howard, 2000). Continuum-solvation models (SM*x*) are based on a thermodynamically linear relationship of the logarithm of H and the solvation free energy (ΔG_s). A comprehensive review of performances of SM*x* models (Dearden and Schüürmann, 2003) using a data-set including 700 experimental H points has revealed that despite the high computational cost of these methods, large errors even up to 30 orders of magnitude arise for H. UNIFAC is a semi-empirical, thermodynamics based QSPR–PPR model used to calculate activity coefficient. For environmental applications, UNIFAC has been used directly to calculate infinite dilution activity coefficient (γ^{∞}) for aqueous solution (Shimotori and Arnold, 2002) or indirectly by extrapolation of vapor-liquid equilibrium data obtained at higher solute concentrations (Örnektekin et al., 1996). The value of γ^{∞} is then used with vapor pressure and total pressure ratio value (P^{sat}/P_T) to calculate H. Although the UNIFAC approach is able to consider temperature effects, it requires interaction parameters that are obtained from model fit to experimental phase-equilibrium data, which are often lacking for chemicals of environmental interest.

Alternatively, QSPRs based solely on calculated molecular descriptors, which represent the quantitative features of a molecule, provide a promising method for the accurate estimation of H. Such QSPR studies have provided satisfactory models for the prediction of Henry's law constants of rather small data-sets and specific chemical classes (Dearden and Schüürmann, 2003); however, developing a comprehensive QSPR for a wide range of chemicals still remains a challenge for researchers. Previous QSPR models (e.g. Abraham et al., 1994; Katritzky et al., 1996; Dearden et al., 1997, 2000; English and Carroll, 2001; Yao et al., 2002; Yaffe et al., 2003) have been developed based on different numbers of data-set points up to 495 compounds (Yaffe et al., 2003) and have different performances.

The performance of the models is usually evaluated by means of standard error (SE or s) or root mean square error (RMSE) for training and test data-sets and seldom by means of absolute average error (AAE); however this should not be overinterpreted by, for example, comparison of the RMSE or AAE of different models with different number of compounds in each data-set. Strictly speaking, the performances of models should be compared only when the same data-set is used.

The objectives of the present study are to analyse and compare the performances of different molecular descriptor sets, to evaluate the application of a genetic algorithm (GA) for descriptor selection and radial basis function network (RBFN) for QSPR model development, to establish a satisfactory QSPR model for H of organic compounds of environmental interest, and to achieve some insight into the main molecular features of such compounds that influence the partitioning phenomenon between air and water.

2. Materials and methods

2.1. Data-set

The data-set used in this study comprises a diverse set of 940 organic compounds including a large set of nitro compounds, which have been used only rarely in previous studies. The data were collected from different sources (Meylan and Howard, 2000; Lin and Sandler, 2002), and were compiled in the units of atm m^3 mol^{-1}, and presented as the logarithm of H at 25 °C, the values of which range from −11.475 to 1.307. The data-set is a mix of directly and indirectly measured H, in which indirectly measured data were selected meticulously according to reliability of solubility and vapor pressure data.

The data-set was randomly divided into a training set of 770 compounds and a test set of 170 compounds for linear regression analysis, and three subsets of 770, 110, and 60 compounds as the training, cross validation, and test data-sets for RBFN analysis, in such a way that test sets included compounds representative of all chemical groups.

2.2. Molecular optimization and descriptor generation

The SMILES (Simplified Molecular Input Line Entry System) strings of all compounds were entered into the Tsar 3.3 (© 2000 Oxford Molecular Limited, and now available from Accelrys Inc.) software to generate the 3D structures of molecules. Then the *mol* files of compounds were exported from Tsar to AMPAC (Austin Method PACkage) software (Semichem, Inc.), where the molecular structures were optimized using the PM3 (Parameterized Model number 3) Hamiltonian in vacuo using the Polak-Ribiere algorithm until the root mean square gradient was 0.01. All calculations were carried out at restricted Hartree-Fock level with no configuration interaction. The optimized geometries were transferred into CODESSA (© 2002 Hypercube, Inc.), Dragon (Web version, © 2003 Talete srl), HYBOT (© 2000 Dr. Sergei V. Trepalin), and Tsar packages to calculate 313, 1352, 14 and 113 molecular descriptors respectively, for each compound.

2.3. Genetic algorithm for descriptor subset selection

Genetic algorithm is inspired by Darwin's theory of evolution. The algorithm begins with a set of chromosomes called a population. Here, bit mask vectors, in which the dimension of vectors is equal to the number of all descriptors, serve as chromosomes. Vector element is one if the corresponding descriptor is included in the model and is

zero otherwise. Vectors from one population are taken and used to form a new population. This is motivated by a hope that the new population will yield a better model than the old one. Subsets, which are then selected to form new vectors (offspring), are selected according to their performance in cross-validation; the more suitable they are, the more chances they have to reproduce and survive. There are two basic parameters of GA: cross-over probability and mutation probability. Cross-over operates on selected genes (binary components of chromosomes) from parent chromosomes and creates new offspring. The simplest way to do this is to choose randomly some cross-over point, copy everything before this point from the first parent, and then copy everything after the cross-over point from the other parent. Following the cross-over phase, mutation is applied with a very low probability (e.g. 1%) to all genes in the population. In the mutation phase, each bit may be inverted from 0 to 1 and vice versa. The above steps of the algorithm are called an evolution, and the algorithm usually is stopped after a predefined number of evolutions. In this study, a MATLAB code (Orr, 1996a) was used for the descriptor selection process with GA based on partial least squares (PLS) regression.

2.4. Radial basis function neural network

A RBFN has a similar form to the multi-layer perceptrons neural network, but with just one hidden layer. Thus it can be described as a three-layer feed-forward structure. As presented schematically in Fig. 1, the input layer does not process the information; it only distributes the p elements (e_k) of n input vectors (x_i) from the matrix data-set of X to the hidden layer. The hidden layer contains u radial basis function units, usually a statistical transformation based on a Gaussian distribution. Therefore, each of the units should have center (c_j) and width (r_j). As a simplification in the calculation process, the width of the units can be of identical value (r), which is the case in this study. A RBF is a nonlinear transfer function and operates by measuring the Euclidean distance between input vector and radial basis function center. The output of each unit (z_{ij}) is a scalar element of $n \times (u + 1)$ matrix, known as a design matrix in the training process, Z, as follows:

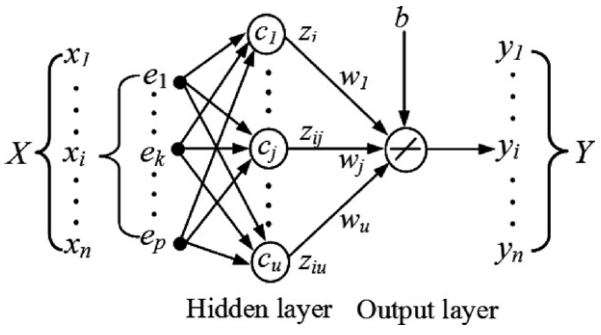

Fig. 1. Schematic representation of RBFN.

$$z_{ij} = \exp\left(\frac{-\|x_i - c_j\|^2}{r^2}\right) \quad (1)$$

The elements of the last column of design matrix are unity for inclusion of bias (b). When the design matrix is found based on training input vectors, selected centers and width, the optimal parameters of the network (weights and b) can be calculated by a least squares training method, which leads to the following equation:

$$W = (Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}Y \quad (2)$$

where Z^{T} is the transposition of matrix Z, W is the network parameter $(u + 1) \times 1$ vector, including u weights (w_{ij}) and a b as a final element. The optimum vector of the parameters can be used for generating the output of any appropriate input, whether training or external test data-set. This is accomplished by the output layer. The operation of the output layer is linear and is achieved with the following matrix internal multiplication:

$$Y = ZW \quad (3)$$

where the elements of vector Y are the corresponding outputs (y_i) of x_i. Therefore, designing a RBFN involves selecting centers, width, weights, and number of hidden units. A forward subset selection routine (Orr, 1996a,b) was used to select the centers from the training data-set. The width of units can be selected by the analysing the performance of RBFN (RMSE) regarding the number of hidden units.

3. Results

In the search for the best descriptor subset by stepwise regression from a large set of the descriptors, a major problem is connected with the mutual collinearity of descriptors, which leads to instability of the regression coefficients, overestimated standard errors, and a critical loss of predictive information. In addition, descriptors with low correlation coefficient (CO) with the modeled property ($\log H$) in a QSPR model do not exactly describe the behavior of the property and almost cover the prediction offset (or errors) of the models, which for external prediction may lead to large errors. Therefore, MATLAB codes were developed for excluding those descriptors that had less than 5% correlation with the $\log H$ data and those descriptors that had more than 70% pair-wise collinearity. Of the inter-correlated descriptors, that one was excluded that had the lower correlation with $\log H$. A combined set of descriptors from the four packages was set up and the above refining procedure was applied. After the refining phase, the total number of descriptors was 250, 315, 57 and 150 for CODESSA, Dragon, Tsar, and combined (including HYBOT) descriptor sets, respectively. Then forward–backward stepwise regression was used to select the 10 best descriptors of each descriptor set according to statistical t and p tests and to develop four QSPR models (CSW/MLR, DSW/MLR, TSW/MLR, HSW/MLR stand-

ing for models based on stepwise selection method of descriptors from CODESSA, Dragon, Tsar, and combined set of all above descriptors including HYBOT descriptors and established by multi-linear regression (MLR)).

The list of descriptors along with their coefficients, statistical values of the coefficients, and correlation coefficients with $\log H$ are presented for four models in Tables 1–4. The statistical results of the four models show that the HSW/MLR model has the best quality. Therefore, the selection routine of GA (Leardi et al., 1992; Leardi, 1994; Leardi and Gonzalez, 1998) was applied on the combined descriptor-set to select the 10 best descriptors from MLR. The population, cross-over probability, and mutation probability of GA were set to 30 chromosomes, 50%, and 1%, respectively. There was not much difference in results by increasing the number of chromosomes; however, it slowed down the evolutionary process. The performance of PLS over a leave-group-out cross-validation procedure (with three elements out in each epoch of cross-validation phase) was used as a fitness measure of GA, followed by the breeding (cross-over and mutation) process. After 5000 evolutions the percentage of survival chance of descriptors in all generations was used to select the 10 best descriptors. Table 5 presents the selected descriptors, their sources and survival chances in the GA, their coefficients, statistical values of the coefficients in the MLR, and correlation coefficients with $\log H$.

The results and performance of five linear QSPR models are presented in Table 6, which shows that the MLR model based on selected descriptors from combined descriptor set by means of GA (GA/MLR model) has good statistical results and prediction ability. Therefore, the radial basis function networks were used to develop a nonlinear model based on the same subset of descriptors which was used for the GA/MLR model, to achieve a more accurate and more generalized QSPR model. As mentioned before, the designing of a RBFN involves selection of centers, number of nodes in the hidden layers, optimum width, and weights. Thirty centers were found by forward subset selection routine from the training data-set. The advantages of this method are that it can determine the number of hidden

Table 1
List of descriptors, coefficients, *t*-values, standard errors (SE), and correlation coefficients (CO) with $\log H$ of CSW/MLR model

No.	Descriptor symbol	Coefficient	*t*-value	SE	CO
	Intercept	−1.132	−11.3	0.100	
1	n_F	0.717	14.4	0.050	0.146
2	n_O	−0.426	−9.1	0.047	−0.492
3	n_N	−0.776	−12.0	0.065	−0.410
4	n_R	−0.513	−11.5	0.045	−0.248
5	*DPSA-1*	0.003	9.6	0.000	0.128
6	*PPSA-3*	0.016	4.1	0.004	−0.266
7	*HA_HDSA-2*	−0.277	−19.4	0.014	−0.645
8	Q_{min}	13.096	16.0	0.819	0.564
9	*HDCA*	−0.198	−9.4	0.021	−0.475
10	*WNSA-1*	−0.015	−13.4	0.001	−0.191

Name of descriptors (category): 1 – Number of fluorine atoms (constitutional); 2 – number of oxygen atoms (constitutional); 3 – number of nitrogen atoms (constitutional); 4 – number of rings (constitutional); 5 – difference in charged partial surface areas [partial positively charged surface area–partial negatively charged surface area], based on quantum chemical partial charge (electrostatic); 6 – atomic charge-weighted partial positively surface area, based on Zefirov's partial charge (electrostatic); 7 – hydrogen-acceptors dependent area-weighted surface charge of hydrogen-bonding donor atoms, based on quantum chemical partial charge (electrostatic); 8 – minimum most negative partial charge (electrostatic); 9 – hydrogen-donors charged surface area, based on quantum chemical partial charge (electrostatic); 10 – weighted partial negatively charged surface area [partial negatively charged surface area × total molecular surface area/1000], based on Zefirov's partial charge (electrostatic).

Table 2
List of descriptors, coefficients, *t*-values, standard errors (SE), and correlation coefficients (CO) with $\log H$ of DSW/MLR model

No.	Descriptor symbol	Coefficient	*t*-value	SE	CO
	Intercept	−0.896	−8.7	0.104	
1	H-050	−1.081	−16.1	0.067	−0.563
2	O-058	−0.927	−10.0	0.093	−0.331
3	*Mor15p*	−1.759	−14.5	0.121	−0.350
4	$R1e^+$	−3.826	−10.1	0.379	−0.463
5	n_N	−0.817	−13.4	0.061	−0.416
6	*RDF020e*	−0.298	−13.2	0.023	−0.288
7	*MAXDP*	−0.578	−14.9	0.039	−0.522
8	n_{COOH}	−1.694	−9.3	0.182	−0.279
9	H-046	0.130	17.5	0.007	0.330
10	*SEigp*	−0.272	−14.0	0.019	0.254

Name of descriptors (category): 1 – H attached to heteroatom (atom-centered fragments); 2 – O = (atom-centered fragments); 3 – 3D MoRSE-signal 15/weighted by atomic polarizabilities (MoRSE); 4 – *R* maximal autocorrelation of lag 1/weighted by atomic Sanderson electronegativities (GETAWAY); 5 – number of N atoms (constitutional); 6 – radial distribution function − 2.0/weighted by atomic Sanderson electronegativities (RDF); 7 – maximal electrotopological positive variation (topological); 8 – number of aliphatic carboxylic acids (constitutional); 9 – H attached to CO(sp3) no X attached to next C (atom-centered fragments); 10 – eigenvalue sum from polarizability weighted distance matrix (eigenvalue-based indices).

Table 3
List of descriptors, coefficients, *t*-values, standard errors (SE), and correlation coefficients (CO) with $\log H$ of TSW/MLR model

No.	Descriptor symbol	Coefficient	*t*-value	SE	CO
	Intercept	−1.679	−16.3	0.103	
1	n_{HD}	−1.910	−25.4	0.075	−0.600
2	μ	−0.312	−8.9	0.035	−0.416
3	n_{NO_2}	2.804	19.1	0.147	−0.140
4	$\log P$	0.374	11.4	0.033	0.153
5	n_F	1.772	29.1	0.061	0.146
6	n_N	−0.582	−9.0	0.065	−0.410
7	n_h	−0.144	−5.6	0.026	0.060
8	n_{CH_3}	0.340	10.6	0.032	0.200
9	$\sum E_{state}$	−0.167	−29.0	0.006	−0.418
10	$^2\kappa$	0.159	8.1	0.020	−0.116

Name of descriptors (category): 1 – Number of hydrogen-bond donor (constitutional); 2 – total dipole moment (electrostatic); 3 – group count for nitro (constitutional); 4 – logarithm of octanol–water partition coefficient (physicochemical); 5 – number of fluorine atoms (constitutional); 6 – number of nitrogen atoms (constitutional); 7 – number of halogen atoms (constitutional); 8 – group count for methyl (constitutional); 9 – sum of electrotopological state indices for whole molecule (topological); 10 – Kier shape index 2 (topological).

Table 4
List of descriptors, coefficients, *t*-values, standard errors (SE), and correlation coefficients (CO) with $\log H$ of HSW/MLR model

No.	Descriptor symbol	Coefficient	*t*-value	SE	CO	Source
	Intercept	−1.183	−9.8	0.121		
1	$\sum C_{ad(o)}$	−0.763	−29.2	0.026	−0.816	HYBOT
2	n_{NO_2}	2.487	20.1	0.124	−0.140	DT[a]
3	*GATS1e*	0.946	12.8	0.074	0.315	Dragon
4	*HA_HDSA-2*	−1.365	−6.5	0.209	−0.638	CODESSA
5	n_F	0.425	12.7	0.033	0.146	CDT[b]
6	PNSA-1	−0.005	−16.2	0.000	−0.151	CODESSA
7	Max($C_{a(o)}$)	−0.657	−11.4	0.057	−0.578	HYBOT
8	*RPCG*	−1.511	−8.5	0.178	−0.111	CODESSA
9	n_{R6}	−0.445	−12.5	0.035	−0.238	DT
10	n_{OH}	−0.510	−7.3	0.070	−0.494	DT

Name of descriptors (category): 1 – Sum of absolute C_a and C_d values [hydrogen-bond free energy acceptor and donor factors, respectively] for all H-bond donor and acceptor atoms in molecule based on octanol–water partition coefficient (thermodynamic); 2 – nitro group count (constitutional); 3 – Geary autocorrelation–lag 1/weighted by atomic Sanderson electronegativities (2D autocorrelations); 4 – hydrogen-acceptors dependent area-weighted surface charge of hydrogen bonding donor atoms based on quantum chemical partial charge; 5 – number of fluorine atoms (constitutional); 6 – partial negatively charged surface area based on Zefirov's partial charge (electrostatic); 7 – largest C_a factor value in molecule (thermodynamic); 8 – relative positive charge based on quantum chemical partial charge (electrostatic); 9 – number of six-membered rings (constitutional); 10 – group count for hydroxyl (constitutional).
[a] Dragon or Tsar.
[b] CODESSA, Dragon, or Tsar.

Table 5
List of descriptors, their percentage of survival chance (%) in the GA (SC), coefficients, *t*-values, standard errors (SE), sources, and correlation coefficients (CO) with $\log H$ of GA/MLR model

No.	Descriptor symbol	SC	Coefficient	*t*-value	SE	CO	Source
	Intercept		−2.057	17.2	0.119		
1	$\sum C_{ad(o)}$	54.4	−0.673	−25.7	0.026	−0.816	HYBOT
2	n_{NO_2}	24.8	1.919	18.0	0.106	−0.140	DT[a]
3	Max($C_{a(o)}$)	23.8	−0.549	−9.6	0.057	−0.578	HYBOT
4	*MLOGP*	19.4	0.301	10.6	0.028	0.249	Dragon
5	*HA_HDCA-1*	17.4	−0.181	−11.9	0.015	−0.609	CODESSA
6	*GATS1e*	16.0	0.861	11.7	0.073	0.315	Dragon
7	*PNSA-1*	14.8	−0.006	−15.6	0.000	−0.151	CODESSA
8	n_F	13.4	0.404	12.7	0.032	0.146	CDT[b]
9	n_{R6}	10.8	−0.414	−10.4	0.040	−0.238	DT
10	$^3\chi_c^v$	10.2	−0.132	−5.2	0.025	−0.273	DT

Name of descriptors (category): 1 – Sum of absolute C_a and C_d values [hydrogen-bond free energy acceptor and donor factors, respectively] for all H-bond donor and acceptor atoms in molecule based on octanol–water partition coefficient (thermodynamic); 2 – nitro group count (constitutional); 3 – largest C_a factor value in molecule (thermodynamic); 4 – Moriguchi octanol–water partition coefficient (physicochemical); 5 – hydrogen-acceptors dependent hydrogen bonding donor ability of the molecule, based on quantum chemical partial charge (electrostatic); 6 – Geary autocorrelation–lag 1/weighted by atomic Sanderson electronegativities (2D autocorrelations); 7 – partial negatively charged surface area based on Zefirov's partial charge (electrostatic); 8 – number of fluorine atoms (constitutional); 9 – number of six-membered rings (constitutional); 10 – valence 3rd order cluster chi index (topological).
[a] Dragon or Tsar.
[b] CODESSA, Dragon, or Tsar.

layer units simultaneously and there is no need to fix the number of hidden layer units in advance. By adding the units of hidden layer one by one and evaluating the performance of the models with different widths, the optimum number of units in the hidden layer can be found.

The performance of a RBFN model was measured by RMSE of the model over 110 cross-validation data. Table 7 presents the minimum RMSE of the cross-validation and the corresponding number of units in the hidden layer with variation of the width. This table clearly shows that the optimum number of the units is 12 and the optimum width should be between 6 and 8. The exact value of width could be found by fixing the number of units and calculating the RMSE of the cross-validation with width variation. An optimum of 6.8 for the width was found by these calculations. After determination of the centers and width, the weights and bias of the RBFN can be calculated easily by setting up the design matrix from Eq. (1) and putting it in Eq. (2). The last element of vector W is the bias and is equal to 0.783. The 12 selected centers and corresponding weights are presented in Table 8.

Therefore to calculate $\log H$ for a compound or set of compounds, one should set up a Z vector or matrix according to Eq. (1) and enter the unit elements as bias coefficients and apply them to Eq. (3). The statistical results and performance of the GA/RBFN model are also presented in Table 6. In addition, the scatter diagrams of training and test data-set for all six models are presented in Fig. 2.

Table 6
Statistical results and performances of all models

Model	r^2%			RMSE				Fisher value
	Training	CV[a]	test[b]	Training	CV	Test	Total[c]	
CSW/MLR	85.4	84.2	79.8	0.811	0.842	0.829	0.814	371
DSW/MLR	83.0	82.4	78.1	0.873	0.873	0.856	0.870	372
TSW/MLR	83.1	82.1	80.0	0.872	0.894	0.800	0.859	443
HSW/MLR	92.5	92.1	87.5	0.582	0.598	0.645	0.594	931
GA/MLR	92.8	92.5	90.0	0.570	0.582	0.574	0.571	972
GA/RBFN	92.9	88.7	98.4	0.564	0.592	0.520	0.564	
Group contribution							0.798[d]	
Bond contribution							0.697	

[a] Cross-validation (external test) over 110 data for GA/RBFN model and leave-one-out (internal test) for linear models over 770 data.
[b] External test over 60 and 170 data for RBFN and MLR models, respectively.
[c] Total RMSE of 940 training and external test data.
[d] For 770 data (failed for 170 data).

Table 7
Effect of width variation on the minimum RMSE of 110 cross-validation (CV) data

Width	Minimum RMSE of CV	Number of hidden units
1	0.8003	29
2	0.6067	12
3	0.5974	12
4	0.5945	12
5	0.5930	12
6	0.5925	12
7	0.5924	12
8	0.5925	12
9	0.5927	12
10	0.5929	12
15	0.5935	12
20	0.5938	12

Table 8
List of centers selected for RBFN and corresponding weights

No.	Compound	Weight
747	Tetrafluoromethane	1.286
763	Trifluoroacetic acid	−15.692
552	Hydrocyanic acid	13.999
653	*N*-Nitrosomorpholine	−7.915
290	2-Pentanone	−26.422
646	Nitromethane	7.033
435	Chlorfluzuron	−2.468
339	3-Pentanone	4.752
3	1,1,1,3,3,3-Hexafluoropropan-2-ol	9.697
768	Tripropylamine	15.353
218	2,4-Dinitrophenol	−3.103
765	Trifluralin	2.935

The performance of GA/RBFN model as the best QSPR model of this study was compared with improved group and bond contribution methods (Table 6), which are the popular methods in calculation of H of compounds and have satisfactory precision (Dearden and Schüürmann, 2003). HENRYWIN software (Meylan and Howard, 2000) was used to calculate the H of all compounds in the training and test data-sets. The group contribution method failed for calculating of H for 170 of the 940 compounds in the data-set. This is one of the most important restrictions of the group contribution method since it does not involve all functional groups and fails if a compound contains a functional group that is not in the original training set of the method. More drawbacks about group contribution method can be found in the literature (e.g. Lin and Sandler, 2002). The RMSE of the group contribution method for 770 compounds was 0.798. The bond contribution method was applied to all 940 compounds and yielded an RMSE of 0.693 in comparison with 0.564 for the GA/RBFN model.

A complete list of all 940 compounds and results of calculated H for all QSPR models, i.e. CSW/MLR, DSW/MLR, TSW/MLR, HSW/MLR GA/MLR, and GA/RBFN models, group and bond contribution methods is deposited as Supporting material and is available upon request.

4. Discussion

As the behavior of air under environmental conditions is close to that of an ideal gas, H depends primarily on interactions in the aqueous phase. Therefore, H is related to ΔG_s (Dearden and Schüürmann, 2003):

$$\log \frac{H}{RT} = \frac{\Delta G_s}{2.3RT} \tag{4}$$

Also according to the universal solvation model (SMx), ΔG_s is partitioned as follows:

$$\Delta G_s = \Delta G_{ENP} + G_{CDS} \tag{5}$$

where ΔG_{ENP} is the change in the solute electronic and nuclear energy and solvent electronic polarization energy (electrostatic) and G_{CDS} is a cavitation-dispersion-solvent structure term (non-electrostatic). Fig. 3 shows the logarithm of the absolute ratio of electrostatic and non-electrostatic terms of ΔG_s for all 940 compounds in the data-set. In addition, the data regarding this figure are listed in the Supporting material. Fig. 3 shows that the electrostatic term of Eq. (5) is at least 10 times greater than the non-electrostatic term for most (ca. 60%) of the compounds in the

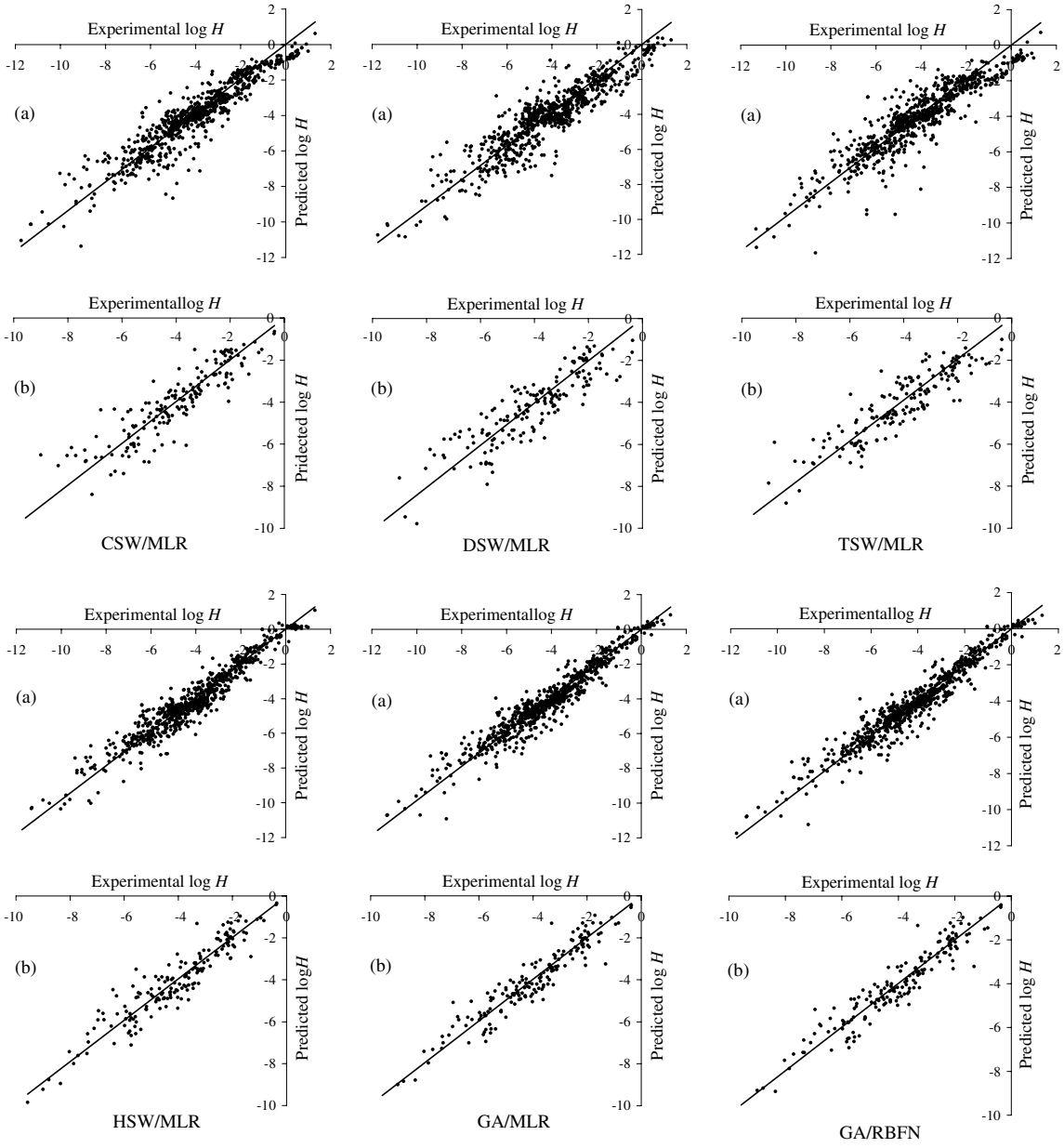

Fig. 2. Scatter diagrams of training (a) and external test (b) data-sets of all models.

data-set. Therefore, it is expected that those topological and geometrical descriptors that characterize the shape and size of a compound make only a weak contribution to the QSPR models of H. Tables 1–5 demonstrate this fact; thus, there is no such descriptor contribution in the CSW/MLR, DSW/MLR and HSW/MLR models (although some electrostatic descriptors such as charged partial surface area are combined shape and electronic information), and a very weak contribution in the TSW/MLR model, represented by $^{2}\kappa$ with lowest CO value of -0.116 and low t-value of 8.1, and in the GA/MLR model by $^{3}\chi_{c}^{v}$ with low CO value of -0.273 and lowest t-value of -5.2 in the corresponding models. Both of these topological descriptors are related to molecular complexity and branching. In other words, for this diverse data-set of chemicals the shape and size of the molecules are not as important for H as electrostatic features of molecules.

Although the different descriptor sources lead to different statistical results and performances, the selected subsets of descriptors are almost the same in nature. The CSW/MLR model (Table 1) has four constitutional descriptors, i.e. number of fluorine atoms (n_F), number of oxygen atoms (n_O), number of nitrogen atoms (n_N), and number of rings (n_R), of which two of them (n_O and n_N) are most probably connected with hydrogen bonding ability of a solute molecule with water molecules. *HA_HDSA-2* and *HDCA*, as electrostatic descriptors, are also related directly to hydrogen bonding ability, and the other descriptors encode the electrostatic features of the solute molecules that control the long-range interaction forces of a solute

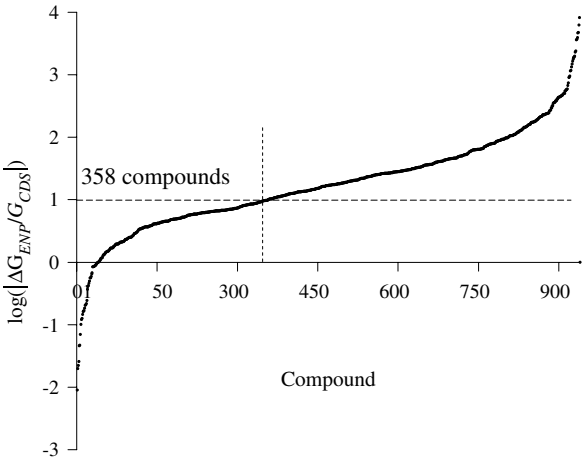

Fig. 3. Electrostatic and non-electrostatic free energy of solvation ratio for all data-set compounds.

molecule surrounded by solvent molecules. A weak contribution of molecular size is implicitly involved with charged partial surface area descriptors, i.e. *DPSA-1*, *PPSA-3*, *HDCA*, and *WNSA-1*. The most important descriptor of this model is *HA_HDSA-2* with the highest *t*-value of −19.4 and CO of −0.645.

The DSW/MLR model (Table 2) has two constitutional descriptors, i.e. number of carboxylic groups (n_{COOH}) and n_N. The group count descriptor, n_{COOH}, represents a molecule's capability to be a hydrogen-bond acceptor and n_N indirectly represents a molecule's capability to participate in hydrogen bonding. H-050, O-058, and H-046 are atom-centered fragment descriptors or atom-centered codes (ACCs). AAC describes each atom by its own atom type and the bond types and atom types of its first neighbors, and it has information regarding different functional groups. However, ACC cannot discriminate between different arrangements of functional groups within a molecule. Hydrogen attached to heteroatom (H-050) as an ACC is related to hydrogen-bond donation ability of a molecule, the most important descriptor of this model with a high *t*-value of −16.1 and highest CO of −0.563. *SEigp* has a combination of information about polarizability and branching of a molecule. The maximal electrotopological positive variation (*MAXDP*) is related to the electrophilicity of a molecule, and the remaining descriptors can be also related to electronegativity, polarizability and charge distribution in the molecule.

The TSW/MLR model (Table 3) has six constitutional descriptors i.e. number of hydrogen-bond donors (n_{HD}), number of nitro groups (n_{NO_2}), n_F, n_N, number of halogen atoms (n_h), and number of methyl groups (n_{CH_3}). n_{HD} directly, and (n_{NO_2}) and n_N indirectly, encode the ability of a molecule to form hydrogen bonds with water molecules. The log *P* descriptor is a measure of hydrophobicity of a compound and ${}^2\kappa$ is a measure of shape and molecular complexity and encodes information about the spatial density of atoms in a molecule. In this model, only μ and $\sum E_{state}$ reflect the effect of electrostatic features of solute molecules on *H*. The *E*-state index (E_{state}) gives information related to the electronic and topological state of an atom in the molecule. In other words, it is a measure of the electronic accessibility of the atom and can be interpreted as a probability of interaction with solvent (water) molecules. However, the index cannot be considered a pure electronic descriptor; it is, in fact, a descriptor of atom polarity and steric accessibility. The dipole moment (μ) is a vector quantity and encodes displacement with respect to the center of gravity of positive and negative charges in a molecule and is important in modeling solvation properties of compounds that depend on solute–solvent interaction. The most important descriptor of this model is n_{HD} with *t*-value of −25.4 and CO of −0.600.

The HSW/MLR model (Table 4) has four constitutional descriptors, i.e. n_{NO_2}, n_F, number of six-membered rings (n_{R6}), and number of hydroxyl groups (n_{OH}). n_{NO_2} most probably represents hydrogen-bond acceptor ability and n_{OH} reflects molecular capability for hydrogen-bond acceptance or donation. Two thermodynamically derived descriptors from the HYBOT software, namely $\sum C_{ad(o)}$ and Max($C_{a(o)}$), demonstrate marked influences on the performance of the QSPR model (Table 6). $C_{a(o)}$ and $C_{d(o)}$ are the hydrogen-bond acceptor and donor factor values respectively stemming from octanol–water partition coefficient, and $\sum C_{ad(o)}$ is the sum of absolute $C_{a(o)}$ and $C_{d(o)}$ values for all hydrogen-bond donor and acceptor atoms in a molecule. Furthermore, *HA_HDSA-2* is an electrostatic descriptor that is directly related to hydrogen-bond acceptor capability of a molecule. Electrostatic intermolecular forces between solute and solvent molecules in this model are characterized by the remaining descriptors of Table 4, i.e. *GATS1e*, *PNSA-1*, and *RPCG*. The most important descriptor of this model is $\sum C_{ad(o)}$ with *t*-value of −29.2 and CO of −0.816.

The GA/MLR (Table 5) model is developed based on the GA descriptor selection method and from the same set of descriptors which is used in developing the HSW/MLR model. Just three descriptors are different from those which have been achieved in the HSW/MLR model, i.e. Moriguchi octanol–water partition coefficient (*MLOGP*), hydrogen-acceptors dependent hydrogen bonding donor ability of the molecule (*HA_HDCA-1*), and 3rd order valence cluster molecular connectivity (${}^3\chi_c^v$). Again, $\sum C_{ad(o)}$ is the most important descriptor of the model, where its chance of survival in all generations of the GA is more than other descriptors and its *t*-value and CO are −25.7 and −0.816, respectively. More information about all the above-mentioned descriptors can be found in the literature (Todeschini and Consonni, 2000).

Due to the diversity of chemicals in the data-set, the presence of constitutional descriptors in all QSPR models was expected; however, some of them are strongly related to the hydrogen bonding ability of the compounds. This means that a QSPR model needs to model some groups of compounds in the data-set by means of specific atomic

or group indicators, and to shift the model prediction values accordingly, as their partitioning between water and air might be accomplished in entirely different ways from the others. Inclusion of hydrogen bonding descriptors in all QSPR models with high COs and *t*-values reveals that hydrogen bonding is the most important molecular feature of solvent–solute interaction in governing *H* of organic compounds in the air–water system. These results are completely consistent with the nature of water molecules as solvent, since they are very good hydrogen-bond acceptors and donors. Although some electrostatic descriptors appeared in all models, their importance and influence were not so great as those of hydrogen bonding.

According to Table 6, the three models of CSW/MLR, DCSW/MLR, and TSW/MLR, which are based on individual descriptor generator packages, i.e. CODESSA, Dragon, and Tsar, have no significant advantage over each other. In contrast, the HSW/MLR model based on the combined descriptors of all three packages, together with a very limited set of descriptors from the HYBOT package, has satisfactory statistical results and prediction ability. Statistical results from the GA/MLR are not very different from those of the HSW/MLR model, but the former is more generalized. Furthermore, applying the RBFN approach for modeling of *H* from the selected descriptor set from GA (GA/RBFN model) has improved the prediction ability of the model. Twelve (from 30) hidden units were found to be optimal for the GA/RBFN model. Although a greater number of hidden units decreased the RMSE of the model, it led to an over-fitted model. It should be mentioned that the cross-validation of the GA/RBFN model was carried out on 110 data points and the cross-validation of the other models was performed using the leave-one-out method. Therefore, the RMSEs of cross-validations should not lead to the deduction that the RMSE of GA/RBFN is larger than GA/MLR model. In fact, the RMSE of the GA/RBFN model for all the external data-set (170 data) is 0.567, lower than the RMSE of the GA/MLR model.

The results in Table 6 emphasize the importance of using the same data-set for comparison of models. For example, Lin and Sandler (2002), using a data-set of 395 compounds, obtained a RMSE of 0.34 for their $\log H$ model. However, they reported that the Meylan and Howard (2000) methods yielded, for the same compounds, RMSEs of 0.52 (group method) and 0.43 (bond method). The differences found are not as large as those found between our GA/RBFN model and those of Meylan and Howard (2000) for our 940-compound data-set.

The behavior of *H* in any linear QSPR model can be explained with the sign of the descriptor coefficients. However, the signs of descriptor coefficient, *t*-value and CO should be the same to guarantee the integrity of judgment about *H* variation against the specific descriptor. For instance, the same minus signs of coefficients, *t*-values, and COs for hydrogen-bond related descriptors in all linear models confirms that increasing the hydrogen-bond ability of an organic compound in water causes a decrease in *H* or in other words, an increase in its solubility in water.

In conclusion, the five linear QSPR models of *H* show that hydrogen bonding, as a short-range intermolecular force, plays the main role in governing partitioning of an organic compound between air and water. Electrostatic intermolecular forces, as long-range forces, also have an important effect on *H*, whereas there is a much weaker correlation between *H* of a compound and the shape, size, and complexity of the molecular structure of compounds. The QSPR models of GA/MLR and GA/RBFN are based solely on calculated information from the molecular structures of the compounds and have better performances than the bond and group contribution methods. RBFN as a simple nonlinear modeling approach has improved the accuracy of the linear QSPR model, whereas establishing a MLPs neural network up to two hidden layers and up to 50 neurons in each layer was unsuccessful and led to over-fitted models.

Finally, for another comparison base, AAEs% are calculated for the developed models which are 30%, 33%, 37%, 16%, 14% and 14% for CSW/MLR, DSW/MLR, TSW/MLR, HSW/MLR, GA/MLR and GA/RBFN models, respectively.

Acknowledgment

The authors thank the British Council for supporting this work under a Partial Scholarship to H. Modarresi.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.chemosphere.2006.09.049.

References

Abraham, M.H., Andonian-Haftvan, J., Whiting, G.S., Leo, A., Taft, R.S., 1994. Hydrogen bonding. Part 34. The factors that influence the solubility of gases and vapours in water at 298 K, and a new method for its determination. J. Chem. Soc. Perkin Trans. 2, 1777–1779.

Baker, J.E., Bamford, H.A., Poster, D.L., Huie, R.E., 2002. Using extrathermodynamic relationships to model the temperature dependence of Henry's law constants of 209 PCB congeners. Environ. Sci. Technol. 36, 4395–4402.

Brennan, R.A., Nirmalakhandan, N., Speece, R.E., 1998. Comparison of predictive methods for Henry's law coefficients of organic chemicals. Water Res. 32, 1901–1911.

Cramer, R.D., 1980. BC (DEF) parameters. 2. An empirical structure-based scheme for the prediction of some physical properties. J. Am. Chem. Soc. 102, 1849–1859.

Dearden, J.C., Schüürmann, G., 2003. Quantitative structure–property relationships for predicting Henry's law constant from molecular structure. Environ. Toxicol. Chem. 22, 1755–1770.

Dearden, J.C., Cronin, M.T.D., Sharra, J.A., Higgins, C., Boxall, A.B.A., Watts, C.D., 1997. The prediction of Henry's law constant: a QSPR from fundamental considerations. In: Chen, F., Schüürmann, G. (Eds.), Quantitative Structure–Activity Relationships in Environmental Sciences-7, SETAC, Pensacola, FL, USA, pp. 135–142.

Dearden, J.C., Ahmad, S.A., Cronin, M.T.D., Sharra, J.A., 2000. QSPR prediction of Henry's law constant: improved correlation with new parameters. In: Gundertofte, K., Jørgensen, F.S. (Eds.), Molecular Modeling and Prediction of Bioactivity. Plenum, New York, NY, USA, pp. 273–274.

English, N.J., Carroll, D.G., 2001. Prediction of Henry's law constants by a quantitative structure property relationship and neural networks. J. Chem. Inf. Comput. Sci. 41, 1150–1161.

Hine, J., Mookerjee, P.K., 1975. The intrinsic hydrophilic character of organic compounds. Correlations in terms of structural contributions. J. Org. Chem. 40, 292–298.

Katritzky, A.R., Mu, L., Karelson, M., 1996. A QSPR study of the solubility of gases and vapors in water. J. Chem. Inf. Comput. Sci. 36, 1162–1168.

Leardi, R., 1994. Application of a genetic algorithm to feature selection under full validation conditions and to outlier detection. J. Chemometr. 8, 65–79.

Leardi, R., Gonzalez, A.L., 1998. Generic algorithms applied to feature selection in PLS regression: how and when to use them. Chemometr. Intell. Lab. Syst. 41, 195–207.

Leardi, R., Boggia, R., Terrile, M., 1992. Genetic algorithms as a strategy for feature selection. J. Chemometr. 6, 267–281.

Lin, S.T., Sandler, S.I., 2002. Henry's law constant of organic compounds in water from a group contribution model with multipole corrections. Chem. Eng. Sci. 57, 2727–2733.

Mackay, D., Shiu, W.S., 1981. A critical review of Henry's law constants for chemicals of environmental interest. J. Phys. Chem. Ref. Data 10, 1175–1199.

Mackay, D., Shiu, W.S., Ma, K.C., 2000. Henry's law constant. In: Boethling, R.S., Mackay, D. (Eds.), Handbook of Property Estimation Methods for Chemicals: Environmental and Health Sciences. Lewis, Boca Raton, FL, USA, pp. 69–87.

Meylan, W.M., Howard, P.H., 1991. Bond contribution method for estimating Henry's law constants. Environ. Sci. Technol. 10, 1283–1293.

Meylan, W.M., Howard, P.H., 2000. HENRYWIN 3.10, Syracuse Research, Syracuse, NY.

Örnektekin, S., Paksoy, H., Demirel, Y., 1996. The performance of UNIFAC and related group contribution models. Part II. Prediction of Henry's law constants. Thermochim. Acta 287, 251–259.

Orr, M.J.L., 1996a. MATLAB Routines for Subset Selection and Ridge Regression in Linear Neural Networks. Centre for Cognitive Science, Edinburgh University.

Orr, M.J.L., 1996b. Introduction to Radial Basis Function Networks. Centre for Cognitive Science, Edinburgh University.

Shimotori, T., Arnold, W.A., 2002. Henry's law constants of chlorinated ethylenes in aqueous alcohol solutions: measurement, estimation, and thermodynamic analysis. J. Chem. Eng. Data 47, 183–190.

Staudinger, J., Roberts, P.V., 1996. A critical review of Henry's law constants for environmental applications. Crit. Rev. Environ. Sci. Technol. 26, 205–297.

Todeschini, R., Consonni, V., 2000. Methods and Principles in Medicinal Chemistry, vol. 11. Handbook of Molecular Descriptors. Wiley-VCH Weinheim, Germany.

Yaffe, D., Cohen, Y., Espinosa, G., Arenas, A., Giralt, F., 2003. A fuzzy ARTMAP-based quantitative structure–property relationship (QSPR) for the Henry's law constant of organic compounds. J. Chem. Inf. Comput. Sci. 43, 85–112.

Yao, X., Liu, M., Zhang, X., Hu, Z., Fan, B., 2002. Radial basis function network-based quantitative structure–property relationship for the prediction of Henry's law constant. Anal. Chim. Acta 462, 101–117.