Fluid Phase Equilibria 379 (2014) 96–103

Contents lists available at ScienceDirect

Fluid Phase Equilibria

journal homepage: www.elsevier.com/locate/fluid

Accurate prediction of the solubility parameter of pure compounds from their molecular structures

Tareq A. Albahri*

Chemical Engineering Department, Kuwait University, P.O. Box 5969, Safat 13060, Kuwait

ARTICLE INFO

Article history:
Received 10 January 2014
Received in revised form 20 April 2014
Accepted 15 July 2014
Available online 23 July 2014

Keywords:
Group contribution
Neural networks
QSPR
Solubility parameter
Structure property correlation

ABSTRACT

A quantitative structure property relation (QSPR) method for predicting the solubility parameter (δ) of pure compounds is presented. Artificial neural network (ANN) model was developed and used to probe the structural groups that have significant contribution to the overall solubility of pure compounds and arrive at the set of groups that can best represent the solubility parameter for about 418 substances. The 36 atom-type structural groups listed can predict the solubility parameter of pure compounds from the knowledge of the molecular structure alone with a correlation coefficient of 0.998 and an absolute standard deviation and error of 0.109 and 0.67%, respectively. The results are further compared with those of the traditional structural group contribution (SGC) method based on multivariable regression as well as other methods in the literature. The method is very useful in predicting the solubility potential of various compounds and has advantages in terms of combined accuracy and simplicity.

1. Introduction

Food, medical and petroleum industries have recently placed more focus on the solubility of raw materials, undesirable gases etc. to improve medical drugs, reduce environmental emissions, and extract vegetable oil from plant seeds. Solubility plays key role in designing purification process like absorbers, strippers, distillation columns, extraction and leaching equipment [1].

There are many solubility scale systems in the literature including the solubility grade, aromatic character, aniline cloud point, wax number, heptane number, Kaouri-Butanol number, and the solubility parameter (δ) which is perhaps the most widely applicable of all.

Numerically, the solubility parameter is defined by the following equation [2]:

$$\delta = \left(\frac{\Delta U^{vap}}{V^L}\right)^{1/2} \quad (1)$$

where, ΔU^{vap} is internal energy change on vaporization to the ideal gas, in cal/mol, and V^L is the liquid molar volume at 25 °C, in cm^3/mol.

An approximation of the internal energy change yields [2]:

$$\delta = \left(\frac{\lambda - RT}{V^L}\right)^{1/2} \quad (2)$$

where, λ is the heat of vaporization at 25 °C, in cal/mol, R is the gas constant (1.9872 cal/mol K), and T is the absolute temperature, 298.15 K. This equation was used to calculate solubility parameter in units of $(cal/cm^3)^{1/2}$.

The above equation, also known as the Hildebrand [3] solubility parameter (δ) provides a numerical estimate of the degree of interaction between materials, and can be a good indication of solubility. Materials with similar values of δ are likely to be miscible. The Hildebrand solubility parameter is the square root of the cohesive energy density, which is the amount of energy needed to completely remove unit volume of molecules from their neighbors to infinite separation (an ideal gas), which is equal to the heat of vaporization divided by molar volume [3]. The cohesive energy density is a direct reflection of the degree of van der Waals forces holding the molecules of the liquid together. In order for a material to dissolve, these same interactions need to be overcome as the molecules are separated from each other and surrounded by the solvent.

Solubility parameter provides simple predictions of phase equilibrium based on a single parameter that is readily obtained for most materials. These predictions are often useful for non-polar and slightly polar systems without hydrogen bonding. For polar molecules, more complicated 3D solubility parameters, such as

* Tel.: +965 2481 7662; fax: +965 2483 9498.
E-mail address: toalbahri@gmail.com

http://dx.doi.org/10.1016/j.fluid.2014.07.016
0378-3812/

Hansen Solubility Parameters have been proposed [4]. The solubility parameter applies only to associated solutions (accounts only for positive deviations from Raoult's law); it cannot account for negative deviations.

2. Background

Many methods have been developed for predicting the solubility of pure compounds. The more recent methods use the molecular structure of the molecules and artificial intelligence (AI) [5–10]. Huuskonen's [5] developed a multivariable linear regression (MLR) and artificial neural network (ANN) models for estimating the aqueous solubility's of organic compounds using 6 intricate topological indices representing molecular connectivity and shape. He used 24 atom-type electro-topological (E-state) indices as structural parameters in order to improve the pharmaceutical application in preparing medical drugs. Using MLR method and Statistical Package for the Social Sciences (SPSS) software, the 30 structural parameters (6 topological and 24 atom type) were determined for the following equations

$$\log S = \sum(a_i S_i) - 1.350 \tag{3}$$

where S is the solubility, a_i are the regression coefficients, and S_i are the corresponding structural parameters. The MLR model was not very accurate with correlation coefficient of 0.88 and a standard deviation of 0.71. The same 30 structural parameters were then used as input to a 30-12-1 back-propagation ANN model that showed better results with a correlation coefficient of 0.92 and a standard deviation of 0.60.

Yan and Gasteiger [6] developed two models to predict the solubility of organic compounds using MLR and ANN. The molecules were described by a set of 32 values of Radial Distribution Function (RDF) code representing the molecules 3D structure and eight additional descriptors. The 3D coordinates were obtained using a 3D structure generator that requires the connection table and optionally available stereo-chemical information to produce the Cartesian coordinates of the atoms. The RDF function used is the following

$$g(r) = f\sum_{i}^{N-1}\sum_{j>i}^{N} A_i A_j \cdot \exp\left[-B(r - r_{ij})\right]^2 \tag{4}$$

with

$$f = \frac{1}{\left(\sum_r [\chi g(r)]^2\right)^{1/2}} \tag{5}$$

where f is the scaling factor, N is the number of atoms, r_{ij} is the distance between the atoms i and j, B is the smoothing parameter, r is the atomic radius, A_i and A_j are the characteristic atomic properties A of atom i and j. The additional eight descriptors were calculated using yet another software to calculate the mean molecular polarizability, aromatic indicator of a molecule, aliphatic indicator of a molecule, highest hydrogen bond acceptor potential, highest hydrogen bond donor potential, number of hydrogen bond donor groups, and number of atoms of element nitrogen and oxygen. The robust 40-8-1 back-propagation ANN model predicted the solubility of organic compounds with a correlation coefficient of 0.92 and standard deviation 0.59 for which is disappointing considering the tedious effort. Furthermore, the MLR method predictions of the solubility were less accurate with a correlation coefficient of 0.82 and standard deviation 0.79 using 40 descriptors as input variables.

Raevsky et al. [7] developed a structure–property relation for predicting the solubility for 42 drugs according to the structural and physicochemical similarity of molecules as follows:

$$\log S = 0.42 - 0.275\alpha + 0.96\sum C_a - 0.27\sum C_d \tag{6}$$

where α is the polarizability, $\sum C_a$ is the hydrogen bond acceptor factor, $\sum C_d$ is the hydrogen bond donor factor. The results showed a predictive correlation coefficient of 0.966 and standard deviation of 0.35.

Bruneau [8] developed a method for predicting the solubility of pure compounds using 100 descriptors (topological, geometrical, and electronic) emphasizing surface properties for every compound. Bayesian learning of neural nets was used to select the most parsimonious models and train them from proprietary and public data sets. The predictive ability of the models were accessed using two new parameters; $NDD_{x,ref}$ the normalized smallest descriptor distance of a compound x to a reference data set and $NDD_{x,mod}$ the combination of $NDD_{x,ref}$ with the dispersion of the Bayesian neural nets calculations. The results show that it is possible to obtain a generic predictive model for the overall data set with a correlation coefficient of 0.95 and average deviation of 0.45. However, the robust and elaborate method and the 100 intricate (2D, 3D, and charge dependent) descriptors render the method unfavorable.

Katritzky et al. [9] developed a two-parameter quantitative structure–property relation (QSPR) equation to predict the solubility for 95 pure hydrocarbons with a correlation coefficient of 0.977. Less accurate results were obtained for the solubility of a larger set of organic compounds using the five-parameter equation shown below with a correlation coefficient of 0.941 and average absolute deviation and percentage error of 0.52 and 0.42%, respectively.

$$\begin{aligned} Log\,S \;&= (2.6 \pm 0.22) + (42.37 \pm 1.11)\mathrm{HDCA}(2) \\ &+ (0.65 \pm 0.02)[\mathrm{N(O)} + 2 * \mathrm{N(N)}] \\ &+ (-0.16 \pm 0.02)(E_{\mathrm{HOMO}} - E_{\mathrm{LUMO}}) \\ &+ (0.12 \pm 0.01)\mathrm{PCWT^E} + (0.82 \pm 0.06)N_{\mathrm{rings}} \end{aligned} \tag{7}$$

where S is the solubility, HDCA(2) is the hydrogen bonding related descriptor, [N(O)+2*N(N)] is the number of oxygen and nitrogen atoms in the molecule, $(E_{\mathrm{HOMO}} - \mathrm{E_{LUMO}})$ is the energy gap which relates to the dispersion energy of polar solutes in solution, $\mathrm{PCWT^E}$ is the most negative partial charge weighted topological electronic index, and N_{rings} is the number of rings in the molecule.

Finally, Gharagheizi [10] developed a method for predicting the solubility parameter for various pure compounds using Genetic Algorithm-Based Multivariate Linear Regression (GA-MLR), and Generalized Function Approximation Neural Network (GRNN). GA-MLR was used to select the molecular descriptors as inputs for GRNN. The obtained multivariate linear seven molecular-descriptors model by GA-MLR had a low correlation coefficient of 0.821, whereas GRNN model had a correlation coefficient of 0.980, which is the most accurate so far.

All the above models are too complex requiring intricate parameters, unconventional physiochemical descriptors and sophisticated software, yet producing less accurate results but for one. There is therefore need for more accurate yet simpler methods for calculating the solubility of pure compounds. The purpose of this work was to develop a simple yet accurate method to calculate the solubility parameter for pure compounds using least possible information and without having to resort to additional parameters, like the heat of vaporization or the molar volume, which are likewise difficult to determine. We can also deduce from the above that MLR models are less accurate than ANN models.

3. Method

The solubility parameter (δ) is not an easy property to estimate or correlate because of its complex dependency on the molecular structure and the intermolecular and intramolecular forces of the