Fluid Phase Equilibria 383 (2014) 108–114

Contents lists available at ScienceDirect

Fluid Phase Equilibria

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

# Solubility modeling in three supercritical carbon dioxide, ethane and trifluoromethane fluids by one set of molecular descriptors

Reza Tabaraki *, Aref Toulabi

*Department of Chemistry, Faculty of Science, Ilam University, Ilam, Iran*

ARTICLE INFO

*Article history:*
Received 12 February 2014
Received in revised form 22 August 2014
Accepted 5 October 2014
Available online 8 October 2014

*Keywords:*
Solubility
Supercritical fluid
Carbon dioxide
Ethane
Trifluoromethane
ANN
MLR

ABSTRACT

Quantitative structure property relationships (QSPR) were developed for the first time predicting of solubility in supercritical carbon dioxide, ethane and trifluoromethane over a wide range of pressures (5.1–36.2 MPa) and temperatures (308–343 K). A large number of descriptors were calculated and a subset of calculated descriptors was selected by genetic algorithm–multiple linear regression (GA–MLR). Four molecular descriptors and three experimental descriptors such as pressure, temperature and melting point were selected as the most feasible descriptors for prediction of solubility in three supercritical fluids. The data set consisted of 14 molecules in various temperatures and pressures, which form 586 solubility data. Modeling of the relationship between selected descriptors and solubility data was achieved by linear (multiple linear regression; MLR) and nonlinear (artificial neural network; ANN) methods. The artificial neural network architectures and their parameters were optimized simultaneously. The root mean squares error (RMSE) for supercritical carbon dioxide, ethane and trifluoromethane were 0.56, 0.68 and 0.72, respectively. The performance of the ANN models was also compared with multiple linear regression models and the results showed the superiority of the ANN over MLR model.



## 1. Introduction

Supercritical fluid technology (SFT) finds applications in chemical, biochemical, pharmaceutical and food processing industries. Supercritical fluids (SCFs) have diffusivities between that of gases and liquids; compressibility's comparable to gases, densities comparable to liquids and negligible surface tension. These properties make them attractive solvents for many industrial applications [1]. Solubility data in SCFs are important for the successful implementation of SFT. The experimental determination of solubility of organic solids in SCFs at various temperatures and pressures is expensive. Regarding the difficulties of solubility measurement in SCF, development of mathematical model to predict the solubility of new or even non-synthesized compounds is essential for saving both time and money. Therefore, modeling and prediction of solubility is essential.

In the mathematical modeling of solubility data in supercritical fluids, the solubility systems can be categorized in three groups, a single solute in a supercritical fluid, mixed solutes in a supercritical fluids and a single solute in mixed supercritical fluids or supercritical fluid plus an organic solvent. Different models have been presented for solubility in supercritical fluid and can be categorized into two groups, theoretical (such as equations of state and semi-empirical equations) and empirical equations (such as density based equations). Numerous theoretical models have studied solubility in binary solid-SC fluid systems such as cubic equations of state, perturbed hard-sphere equations of state, lattice models, Kirkwood–Buff solution theory, Monte Carlo simulation and mean field theory [2]. Equations of state often require properties such as critical temperature, critical pressure and acentric factor that are not available for the solid solutes. Also, the models require one or more temperature-dependent parameters which must be obtained from binary solid solubility data [3,4]. The empirical models are based on simple error minimization using least squares method and, for most of them; there is no need to use physicochemical properties [5].

One of the most successful approaches to the prediction of chemical properties with molecular structural information is quantitative structure property/activity relationship. In QSPR, the molecular structure is translated into the so-called molecular descriptors using chemical graph theory, information theory, quantum mechanics, etc., and mathematical equations are related chemical structure to a wide variety of physical, chemical and biological properties [6,7]. QSPR models can be used to predict properties of new compound. Major steps in constructing the QSPR models are (i) the proper calculation of molecular descriptors, which satisfactorily describe the properties of a set of chemical substances (ii) selection of the best descriptors (iii) constructing a

* Corresponding author. Tel.: +98 841 2227022; fax: +98 841 2227022.
*E-mail addresses:* rezatabaraki@yahoo.com, r.tabaraki@ilam.ac.ir (R. Tabaraki).

0378-3812/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.fluid.2014.10.008

**Table 1**
Molecular structure of organic solids.

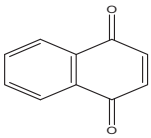
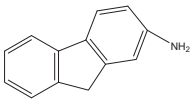
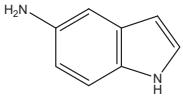
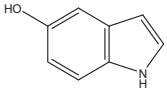
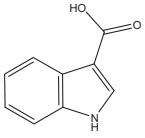
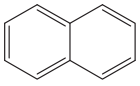
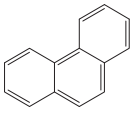
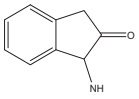
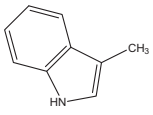
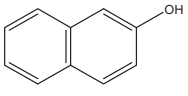
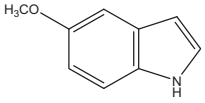
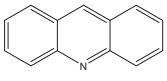
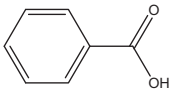
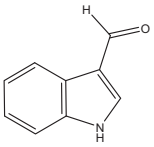
| | Molecules | Molecular structure |
|---|---|---|
| 1 | 1,4-Naphthoquinone |  |
| 2 | 2-Aminofluorene |  |
| 3 | 5-Aminoindole |  |
| 4 | 5-Hydroxyindole |  |
| 5 | Indole 3-carboxylic acid |  |
| 6 | Naphtalene |  |
| 7 | Phenanthrene |  |
| 8 | Oxindole |  |
| 9 | Skatole |  |
| 10 | 2-Naphthol |  |
| 11 | 5-Methoxyindole |  |
| 12 | Acridine |  |

**Table 1** (*Continued*)

| | Molecules | Molecular structure |
|---|---|---|
| 13 | Benzoic acid |  |
| 14 | Indole-3-aldehyde |  |

mathematical model having the best prediction of property data and (iv) validating the quality and predictivity of the model [8].

Genetic algorithms (GA) are optimization tools and randomized search techniques guided by the principles of evolution and natural genetics [9]. They have been proved to be a very efficient method in the feature selection problem. Variables are represented as genes on a chromosome and they are generally coded as binary strings. A population of strings is randomly created. In variable selection, each string is a row vector containing as many elements as there are variables. Each element was coded as 1 if the corresponding variable was selected and 0 if it was not selected. The fitness of the string is equal to the evaluation response that is based on the predictive ability with a given subset of selected variables. The method used in the GA variable selection is designed to select variables with lowest prediction error. Through natural selection and the genetic operators, mutation and recombination, chromosomes with better fitness are found. Natural selection guarantees that chromosomes with the best fitness will propagate in future populations. Mutation allows new areas of the response surface to be explored. GA offers a generational improvement in the fitness of the chromosomes and after many generations will create chromosomes containing the optimized variable settings. GA has several advantages such as the ability to move from local optima present on the response surface and require no knowledge or gradient information about the response surface and can be employed for a wide variety of optimization problems.

Artificial neural network consists of a large number of processing elements (neurons) and connections between them. Function f($x$) maps a set of given input values to some output values $y = f(x)$. A neural network tries to find the best possible approximation of the function f($x$). This approximation is coded in the neurons of the network using weights that are associated with each neuron. The weights of a neural network are learned using an iterative procedure during which examples of correct input–output associations are shown to the network and the weights are modified so that the network starts to mimic this desirable input–output behavior. Learning in a neural network then means finding an appropriate set of weights. This ability to learn from examples and based on this learning the ability to generalize to new situations is the most attractive feature of the neural network. One of the most popular learning algorithms is the back propagation algorithm. The architecture of a network used in connection with the back propagation algorithm is the feed forward layered network. In a feed forward layered network, the processing elements are divided into disjoint subsets, called layers**.** A feed forward network consists of layers (input, hidden and output layers). The input data flow through the network from the hidden layer towards the output layer. The number of hidden layers in a