Fluid Phase Equilibria 427 (2016) 194–201

Contents lists available at ScienceDirect

Fluid Phase Equilibria

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

Prediction of the normal boiling point of oxygen containing organic compounds using quantitative structure–property relationship strategy

Liangjie Jin, Peng Bai*

School of Chemical Engineering and Technology, Tianjin University, Tianjin, 300072, PR China

ARTICLE INFO

Article history:
Received 3 March 2016
Received in revised form
17 July 2016
Accepted 19 July 2016
Available online 20 July 2016

Keywords:
Normal boiling point
Multiple linear regression
Radial basis network
QSPR

ABSTRACT

Quantitative structure–property relationship (QSPR) models were applied to predict the normal boiling point (NBP) of oxygen containing organic compounds, including alcohols, phenols, ethers, aldehydes, ketones, carboxylic acids and esters. The total 432 compounds were divided into 3 subsets according to their structure features. For each subset, 8 significant descriptors were selected from the pool of descriptors. Sequentially, the multiple linear regression (MLR) method as well as the non-linear radial basis network (RBN) was used to correlate and predict the NBP of the compounds. RBN model showed higher accuracy with respect to MLR model and Constantinou-Gani (C-G) group contribution method. Comparison with previous QSPR models indicated that the present models could be more general for NBP prediction of organic compounds with certain oxygen containing functional group. In addition, QSPR models for all the 432 compounds were also deduced, and the results confirmed that RBN model performed better in the field of QSPR modeling.

1. Introduction

The normal boiling point (NBP) can be defined as the temperature at which the vapor pressure of a pure liquid is 760 mmHg. The NBP is usually used to estimate many key physical and physicochemical properties such as critical temperature, enthalpy of vaporization and vapor pressure [1,2], etc. Accurate knowledge of the NBP is essential for the process and equipment design based on fluid phase equilibria in chemical industry. The experimental NBP values of many compounds are often missing in literature due to the costly, laborious, or dangerous measurement procedure for the researcher or the environment. For the reason that the NBP is directly correlated to the chemical structure of the molecule [3], the methods for NBP estimation based on molecular structure are of great significance. There are two main approaches to tackle the problem: group contribution methods (GC methods) and quantitative structure-property relationship models (QSPR models).

GC methods such as those proposed by Lydersen [4], Joback and Reid [5], Klincewicz and Reid [6], Lyman et al. [7] and Constantinou and Gani [8] have been considered as classical approaches to NBP estimation. In these approaches, molecules are considered as made of some predefined fundamental groups, each of which gives a constant contribution to the value of NBP. GC methods have the advantage of quick estimates without requiring experimental data and they provide promising results for small and non-electrolyte molecules [9]. However, the application of GC methods is still limited because not all group-contributions data are available, or stereo-isomers are not distinguished, or the interactions between different groups are not considered.

In addition to GC methods, QSPR models are considered as important complementary tools for the estimation of NBP. Developing a QSPR model first requires a database of compounds in which all NBP values are available and the molecular structures information can be numerically characterized with suitable software. Sequentially, via a series of mathematical or statistical methods, a correlation is quantified between the NBP values and some selected variables (molecular descriptors) which represent the molecular structures information. The first success in QSPR studies was achieved by Wiener [10], who quantified a correlation between the NBP of alkanes and two structure-related parameters.

* Corresponding author.
E-mail addresses: jinliangjie868@sina.com (L. Jin), chemeng114_tju@163.com (P. Bai).

http://dx.doi.org/10.1016/j.fluid.2016.07.015
0378-3812/

Since then various studies on QSPR have been reported. In these papers, the majority of the QSPR models were achieved by traditional multiple linear regression (MLR) approaches. For example, Katritzky et al. [11] developed an eight-parameter MLR equation for 612 organic compounds with $R^2 = 0.965$ and a standard prediction error of 15.5 °C using CODESSA PRO. The MLR equation provided confident prediction of the NBP of organic compounds on the basis of their chemical structure alone. Sola et al. [12] focused on predicting the boiling points for 155 compounds by regression analysis techniques. The final eight-parameter model showed better performance than the most sophisticated Marrero-Gani's GC method. In recent years, computational neural networks have become an important QSPR modeling technique. Compared to MLR approaches, neural networks algorithm could incorporate nonlinear and cross-product terms into QSPR model. In order to predict the NBP of a very large database, Gharageizi et al. [13] optimized a three-layer feed forward artificial neural network with 44 molecular descriptors as the inputs and 40 neurons in the hidden layer. The general performance of the final model was satisfactory and the results indicated the artificial neural network would be a promising strategy to predict the NBP of pure chemicals. However, to our knowledge, the QSPR models that utilize neural networks algorithm to predict the NBP are lacking compared to MLR models, especially those concentrating on molecules with similar functional groups. Moreover, there are few literature referring to the comparison between MLR and neural networks focused on the same compound database.

In the present work, the MLR and neural networks were used to establish the quantitative relationship between molecular structure and the NBP of 432 oxygen containing organic molecules. First, the original set of 432 compounds was divided into 3 different subsets based on structure features. Then for each subset, a traditional MLR model and a novel radial basis network (RBN) model were developed for the NBP prediction. Both models obtained were validated and tested independently. The two obtained QSPR models were compared with each other, and they were also compared with Constantinou-Gani (C-G) group contribution method and other QSPR models in literature. In addition, the QSPR models for all the 432 compounds were also developed, and their performances of NBP prediction were also discussed. The ultimate objective was to establish reliable QSPR models for the NBP prediction of oxygen containing organic compounds.

2. Database and mathematical methods

2.1. Database

Experimental data set of the normal boiling points of 432 organic compounds containing C, H, O are taken from literature [14]. The data set, which includes alcohols, phenols, ethers, aldehydes, ketones, carboxylic acids and esters, was divided into 3 subsets described below according to the bonding types among atoms C and atoms O. In supplementary material tables all the chemicals in the experimental data sets, experimental and calculated boiling points are listed.

2.2. Determination of molecular descriptors

Molecular descriptors, which are numerical characteristics associated to the chemical structures of compounds [15], are necessary components for the development of a QSPR model. In order to calculate molecular descriptors, the 432 molecular structures were sketched using Materials Studio. Then these chemical structures were initially energy-minimized with compass molecular mechanics method and subsequently subjected to AM1 semi-empirical quantum chemical method for final geometry optimization. The three-dimensional structures with lowest energy conformation were ported to the E-dragon software, which can be used free of charge for molecular descriptor calculation online (www.vcclab.org/lab/edragon) [16]. For each compound 1666 descriptors can be calculated, grouped into 20 diverse blocks: Constitutional descriptors, Topological descriptors, Walk and path counts, Connectivity indices, Information indices, 2D-autocorrelation indices, Edge adjacency indices, Burden eigenvalue descriptors, Topological charge indices, Eigenvalue-based indices, Randic molecular profiles, Geometrical descriptors, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, Functional groups, Atom-centered fragments, Charge descriptors, and Molecular properties.

Among the huge number of calculated molecular descriptors, a pre-selection was performed to remove some information-poor descriptors. The first was to remove descriptors not available for all structures and those with constant values. Then, descriptors with the squared correlation coefficient (R^2) value of the one-parameter correlations lower than 0.1 were eliminated. Descriptors were considered collinear if their pair-correlation coefficient value was greater than 0.98. Among the collinear descriptors, the one having the highest R^2 value with the boiling points was retained while the rest were discarded.

2.3. Development of MLR model

The aim of the stage was to select a subset of molecular descriptors from all available descriptors, and found a strictly correlating mathematical equation between minimum number of variables and NBP. Assuming the contribution of each descriptor was linear, a multi-parameter correlation was developed with the following form:

$$Y = a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n$$

In this equation, Y is the NBP of a compound, X_1 through X_n are the calculated molecular descriptors for the compound, a_0 is the y-intercept of the regression model, and a_1 through a_n are the various coefficients for the descriptors determined by the regression model.

The MLR model was formulated as follows.

(1) All compounds were randomly partitioned to training and test sets with the size of 80% and 20% of studied data, respectively. The training set was used to establish the MLR model, while the test set was used to evaluate the prediction capability of the model.
(2) A multi-stepwise regression algorithm was applied to find an optimal subset of descriptors. All descriptors were listed in decreasing order according to the one-parameter R^2. Starting from the top descriptor, other descriptors were introduced one at a step to the regression equation. At each step, F-test was carried out to determine the entry or removal of descriptors. If the probability of the F-value is below 0.05, the descriptor was entered, and if the probability of the F-value is above 0.1, the descriptor was removed. The process was repeated until the addition of more descriptors decreased the average absolute relative deviation (*AARD*) by a threshold value less than 0.01.
(3) The molecular descriptors selected above were used in the MLR model development. Several statistical criteria such as R^2, *AARD*, and root mean square error (*RMSE*) were used as the results of model validation. The optimal MLR model was defined as that with low values of *AARD* and *RMSE*, and a high value of R^2.