Thermochimica Acta 655 (2017) 112–116

Contents lists available at ScienceDirect

Thermochimica Acta

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

Quantitative structure-property relationship (QSPR) study for predicting gas-liquid critical temperatures of organic compounds

Lulu Zhou[a,b], Beibei Wang[b], Juncheng Jiang[a,**], Yong Pan[c,**], Qingsheng Wang[b,d,*]

[a] College of Safety Science and Engineering, Nanjing Tech University, Nanjing, Jiangsu, 210009, China
[b] Department of Fire Protection & Safety, Oklahoma State University, Stillwater, OK 74078, USA
[c] Jiangsu Key Laboratory of Hazardous Chemicals Safety and Control, Nanjing Tech University, Nanjing, Jiangsu, 210009, China
[d] Department of Chemical Engineering, Oklahoma State University, Stillwater, OK 74078, USA

ARTICLE INFO

Keywords:
Critical temperature
Quantitative structure-property relationship
Multiple linear regression
Support vector machine

ABSTRACT

Gas-liquid critical temperature is an important parameter of critical state. Organic compounds are under rapid phase changes leading to explosions when conditions are changed at their critical states. Therefore, for safety purposes it is important to study the gas-liquid critical properties for different organic compounds, especially their critical temperatures. In this work, critical temperatures of 692 organic compounds were collected and applied to build quantitative structure-property relationship (QSPR) models. Dragon software was used to obtain their molecular structure information. Methods of multiple linear regression (MLR) and support vector machine (SVM) were applied to build the models, combined with genetic algorithm method. Between these two models, the MLR model has better internal robustness and the SVM model has better goodness-of-fit predictive ability. The results show the developed models have great performance in predicting the gas-liquid critical temperatures. With these models, critical temperatures of organic compounds can be predicted solely based on their molecular structures.

1. Introduction

Gas-liquid critical properties are basic thermodynamic parameters, which include critical temperature, critical pressure, critical volume, critical density, and so on. These thermodynamic data at critical states are important for the design of chemical processes near critical regions, such as petrol fractionation, supercritical extraction and reactions. These data are also used to calculate other properties and applied to equations of state. When chemicals are at the gas-liquid critical state while the condition or environment changes, their phases could be changed rapidly leading to explosions [1]. Therefore, it is very important to study the gas-liquid critical properties. Critical temperature (T_c) is the temperature above which a gas cannot be liquefied. Experimental determination of their values is a challenge, especially for large molecules that can decompose at high critical temperatures [2].

Quantitative structure-property relationship (QSPR) studies have been widely used to predict the desired properties in chemical engineering and safety science [3–5]. In the 1990s, some research work was done by Leanne et al. [6], Lowell et al. [7], Brian et al. [8] and Alan et al. [9] to build QSPR models using the same 165 compounds from the DIPPR (Design Institute for Physical Properties) database to predict critical temperatures. The major differences among their work were the descriptors and software used during the model building processes. In 2001, Espinosa et al. [10] used fuzzy ARTMAP-based QSPR to predict Tc for 530 compounds. The group contribution methods also have been applied in QSPR models by Nannoolal et al. [11] in 2006 and Gharagheizi et al. [12] in 2011. Some researchers focused on building QSPR models for a specific group of compounds, such as hydrocarbons [13], alkyl benzenes [14], 1-alcohol series [15], sulfur compounds [16] and refrigerants [17]. However, all the above mentioned QSPR models did not have the application domain (AD) process. In their studies, only squared correlation coefficient (R^2), root mean square error (RMSE) and average absolute error (AAE) were used to evaluate the fitness of the obtained models. Organization for Economic Cooperation and Development (OECD) has published five principles for QSPR studies and it shows that the model validation and AD processes are important and necessary [18]. There is one study in 2002 that followed all the necessary processes to build a reliable QSPR model to predict the Tc values, but there was no application domain process [19]. In 2016, Saaidpour performed a complete QSPR study to predict critical temperatures for refrigerant compounds, but not for all types of organic compounds [20].

* Corresponding author at: Department of Fire Protection & Safety, Oklahoma State University, Stillwater, OK 74078, USA.
** Corresponding authors.
E-mail addresses: jcjiang@njtech.edu.cn (J. Jiang), yongpan@njtech.edu.cn (Y. Pan), qingsheng.wang@okstate.edu (Q. Wang).

http://dx.doi.org/10.1016/j.tca.2017.06.021
Received 9 March 2017; Received in revised form 15 June 2017; Accepted 20 June 2017
Available online 21 June 2017

In QSPR studies, the selection of appropriate modeling techniques that can be applied for model development is one of the key problems. At present, many different technologies, such as multiple linear regression (MLR), partial least squares (PLS), and support vector machine (SVM) have been widely used in the QSPR modeling. These techniques can be used for inspection of linear and nonlinear relations between interested property and molecular descriptors. The SVM method was developed from the machine learning community by Vapnik and coworkers [21,22]. The method was originally developed for classification problems. With introduction of ε-insensitive loss function, SVM has been extended to solve regression problems and has shown great performance in QSPR studies because of its remarkable ability to interpret the nonlinear relationships between molecular structure and properties. There have been a lot of applications based on the SVM method in different research fields, such as least-squares SVM [23–27] and LibSVM [28].

The performance of SVM for regression depends on the combination of several parameters. They are kernel function type and its corresponding parameters, capacity parameter C, and ε of ε-insensitive loss function. C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the training error. If C is too small, then insufficient stress will be placed on fitting the training data. If C is too large, then the algorithm will over fit the training data. The optimal value for ε depends on the type of noise present in the data and the number of resulting support vectors. The bigger ε, the fewer support vectors are selected [29].

In this work, for the first time a complete and reliable process for predicting the critical temperature for 692 organic compounds based on the QSPR methodology was presented. The effective variable selection method of genetic algorithm (GA) was applied to select an optimal subset of descriptors that had significant contributions to the overall critical temperatures; then methods of MLR and SVM were employed to fit the possible quantitative relationship that existed between selected descriptors and critical temperatures. In addition, the AD was discussed in this study.

2. Methods

2.1. Dataset

The dataset consists of 692 organic compounds that include hydrocarbons, aldehydes, alcohols, acids, amines, benzenes, sulfur compounds and so on [30]. The values of critical temperature of these compounds range from 190.6 to 908 K. The dataset is randomly divided into a training set with 630 compounds and an external prediction set with 62 compounds. The training set is used for variable selection and model development, while the external prediction set is used for model validation.

2.2. Molecular descriptors calculation and selection

There are many software and programs that can be used to calculate the molecular descriptors, such as Gaussian, CODESSA, and Dragon [31–33]. Dragon is the most widely used molecular descriptors calculation program. It can calculate 18 types (1481 kinds) of molecular descriptors which are shown in Table 1. For a detailed description of these descriptors, refer to Dragon software user's guide [34].

Table 1
The types and numbers of descriptors calculated by Dragon.

Type	Number	Type	Number
Constitutional descriptors	47	Geometrical descriptors	58
Topological descriptors	262	RDF descriptors	150
Molecular walk counts	21	3D-MoRSE descriptors	160
BCUT descriptors	64	WHIM descriptors	99
Galvez topology charge indices	21	GETAWAY descriptors	197
2D autocorrelations	96	Functional groups	121
Charge descriptors	14	Atom-centered fragments	120
Aromaticity indices	4	Empirical descriptors	3
Randic molecular profiles	41	Properties	3

In this work, the Dragon program was used to calculate the molecular descriptors and used to search for the best model of the critical temperature prediction. The calculation was on the basis of the minimum energy molecular geometries optimized by the HyperChem software using MM+ molecular mechanics force field and AM1 semi empirical method [35]. After the calculation of the molecular descriptors, constant and near constant descriptors for all molecules were eliminated, and pairs with a correlation coefficient greater than 0.95 were considered to be inter-correlated so one of them in each correlated pair was deleted. Finally, a total set of 641 remaining descriptors was achieved and used to select an optimal subset of descriptors that have significant contributions to the critical temperatures.

Genetic algorithm (GA) is a well-accepted method to select the optimal subset of descriptors. GA is a powerful optimization method to search for the global optima of solutions. It is developed to mimic some of the processes observed in natural evolution [29].

2.3. Model building

Two kinds of model building methods, which are MLR and SVM, were used in this work. MLR process was performed by SPSS, which is classic statistical analysis software [36]. The SVM model was implemented based on the program Libsvm.

2.4. Model validation

The model validation included three parts: First, the goodness of fit was presented by R^2, RMSE and AAE. Second, the internal robustness was presented by Leave-Many-Out (LMO) method in this study. The result of LMO was the cross-validated correlation coefficient which was Q^2_{LMO}. Third, the external predictive ability was judged by an external $Q^2{}_{ext}$. The detailed calculation processes of $Q^2{}_{LMO}$ and $Q^2{}_{ext}$ were introduced in previous work [30,34].

2.5. Application domain

There are many methods of definition of AD, such as range based, distance based, geometrical based, probability based and so on. One of the most common definitions is determining the leverage values for each compound. In this study, the definition of AD was based on the leverage values which were assumed to follow Gaussian distribution. The advantage of this method was that the application of the model could be quantified and presented with the visual graph named the Williams plot. The leverage value of a compound in the original variable space is defined as

$$h_i = X_i(X^TX)^{-1}X_i^T\,(i = 1,2,3\ldots n)$$

Where X_i is the descriptor row-vector of the considered compound, X is the descriptor matrix derived from the training set descriptor values and n is the number of the compounds. The standard value h* is defined as

$$h^* = \frac{3(k+1)}{m}$$

Where k is the number of model variables and m is the number of the training set compounds. If the leverage value (h) of a compound is higher than the standard value (h*), the predicted value of the compound could be regarded as out of the application range of the model and the predicted result may be not reliable [37,38].