ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/231390835

Prediction of Triple-Point Temperature of Pure Components Using their Chemical Structures

ARTICLE *in* INDUSTRIAL & ENGINEERING CHEMISTRY RESEARCH · DECEMBER 2009

Impact Factor: 2.59 · DOI: 10.1021/ie901029m

CITATIONS
24

READS
12

2 AUTHORS:

Farhad Gharagheizi
Texas Tech University

168 PUBLICATIONS 2,911 CITATIONS

SEE PROFILE

Mehdi Sattari
University of KwaZulu-Natal

40 PUBLICATIONS 325 CITATIONS

SEE PROFILE

Available from: Mehdi Sattari
Retrieved on: 31 January 2016

CORRELATIONS

Prediction of Triple-Point Temperature of Pure Components Using their Chemical Structures

Farhad Gharagheizi*,[†] and Mehdi Sattari[‡]

Department of Chemical Engineering, Faculty of Engineering, University of Tehran, P.O. Box 11365-4563, Tehran, Iran, and Division of Polymer Science and Technology, Research Institute of Petroleum Industry (RIPI), P.O. Box 14665-1998, Tehran, Iran

A quantitative structure property relationship study was performed to develop a model for the prediction of triple-point temperature of pure components. For developing this model, 638 pure components were used, and, for whichever, 1664 molecular descriptors were determined. As a standard tool for subset variable selection, genetic algorithm-based multivariate linear regression (GA-MLR) technique was used. The obtained model is a seven parameters multilinear equation that has a squared correlation coefficient of 0.9410 ($R^2 = 0.9410$).

Introduction

Physical and thermodynamic properties of substances are needed in the design and development of chemical and petrochemical plants. Measurement of these needed properties only for a large number of industrially important substances needs considerable time and cost expenditures. During the last hundred years, many of these needed properties have been measured and most of them have been published in the literature. But the majority of needed properties have not been measured or at least have not been reported in the literature for public use. As a result, use of computational techniques and predictive tools are needed for scientists and engineers who need these properties.

One of the most widely used computational techniques for this purpose is quantitative structure−property relationships methodology (QSPR). In this method, the interested property is correlated by a collection of chemical structure-based parameters. These chemical structure-based parameters are calculated from chemical structure of substances using known mathematical algorithms. These parameters are called "molecular descriptors". Finally, the obtained correlation can be used to estimate or even predict the property. Of course, there are certain, rather obvious limitations to its use: (i) the family of compounds used to derive the QSPR should be chemically similar, and (ii) realistic predictions can only be made for compounds that are chemically related to some of those from which the QSPR model was derived; that is, predictions should be of interpolations or short extrapolations.

One of the most important thermodynamic properties of substances is triple-point temperature (TPT). The TPT is defined as the temperature at which solid, liquid, and vapor of the substance are all in equilibrium. This temperature is widely used in phase equilibrium thermodynamics. For example, below the TPT, solid vaporize without melting (sublime). This fact is used to determine the enthalpy of sublimation. Also, the TPT is used to determine the enthalpy of fusion, vapor solid pressure, and solubility of solids in liquids.[1,2]

In the present study, a QSPR study is performed to predict TPT of pure components. As a standard tool, genetic algorithm based multivariate linear regression is used to develop a multilinear correlation.

Materials and Methods

Materials. Evaluated databases such as DIPPR 801 database are useful tools for developing new property prediction models.[3] DIPPR 801 is recommended by AIChE (American Institute of Chemical Engineers). In this study, 638 pure components were found and their values of TPT were extracted. These components and their TPT are presented as Supporting Information.

Determination of Molecular Descriptors. Molecular descriptors are defined as numerical characteristics associated with chemical structures. The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number applied to correlate physical properties.

There are many software packages which calculate the molecular descriptors of any desired chemical structure. A review of these software packages has been presented in ref 4. One of the most widely used is Dragon software.[5] Dragon calculates 1664 molecular descriptors for many common chemical structures. Since the values of many descriptors are related to the bonds length and bonds angles, etc., the chemical structure of every molecule must be optimized before calculating its molecular descriptors. For this reason, chemical structures of all 638 pure components were drawn in Hyperchem software[6] and optimized using the MM+ molecular mechanics force field.

After optimizing the chemical structures of all 638 pure components, the molecular descriptors were calculated using Dragon. The inputs to this software are the optimized chemical structures of molecules obtained by MM+ optimization.

Developing Model. Usually, in QSPR studies, after calculating the molecular descriptors from optimized chemical structures of all the components available in the data set, the problem is to find a linear equation that can predict the desired property with the least number of variables as well as highest accuracy.

In other words, the problem is to find a subset of variables (most statistically effective molecular descriptors on TPT) from

* To whom correspondence should be addressed. Fax: +98 21 66957784. E-mail: fghara@ut.ac.ir; fghara@gmail.com.

[†] University of Tehran.

[‡] Research Institute of Petroleum Industry (RIPI).

10.1021/ie901029m © 2010 American Chemical Society
Published on Web 12/22/2009

all available variables (all molecular descriptors) that can predict TPT, with minimum error in comparison with the experimental data.

A generally accepted method for this problem is genetic algorithm-based multivariate linear regression (GA-MLR). In this method, genetic algorithm is applied for selection of best subset variables with respect to an objective function. This algorithm was presented by Leardi et al. for the first time.[7]

There are many standard fitness functions such as R^2, adjusted R^2, Q^2, Akaike information content, LOF function, and so on, which are used as objective function in GA-MLR technique.[8] RQK fitness function is a suitable fitness function for model searching proposed to avoid unwanted model properties, such as chance correlation, presence in the models of noisy variables, and other model pathologies that cause lack of model prediction power.[8] This fitness function is a constrained fitness function based on Q^2_{Loo} (leave-one-out cross validated variance) statistics and four tests that must be fulfilled contemporarily. The Q^2_{Loo} is defined as

$$Q^2_{\text{Loo}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{ic})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (1)$$

where y_i, $\bar{y}$, and $\hat{y}_{ic}$ are TPT for ith pure component, mean value of TPT of all pure component, and response of the ith object estimated using a model obtained without using the ith object, respectively.

These four constrains were presented by Todeschini et al.[8] as below:

$$\Delta K = K_{XY} - K_X > 0 \qquad \text{(quick rule)} \qquad (2)$$

$$\Delta Q = Q_{\text{Loo}}{}^2 - Q_{\text{ASYM}}{}^2 > 0 \qquad \text{(asymptotic } Q^2 \text{ rule)} \qquad (3)$$

$$R^{\text{P}} > 0 \qquad \text{(redundancy RP rule)} \qquad (4)$$

$$R^{\text{N}} > 0 \qquad \text{(overfitting RN rule)} \qquad (5)$$

These four constrains have been extensively explained by Todeschini et al.[8]

Since many conditions during this algorithm are checked, we can ensure that the final model is valid, has the predictive power, and is not a chance correlation. In this study, GA-MLR with RQK fitness function was used based on satisfactory results in the author's previous works with this technique.[9−26] To perform GA-MLR a program was written based on the MATLAB software (Mathworks Inc. software).

Before performing GA-MLR, the data set must be divided into two collections. The first one is applied for training and the second one is applied for testing. By means of the training set, the best model is found, and then the predictive power of this obtained model was checked by the test set. In this work, 80% of the database was used for the training set (511 pure components) and 20% (127 pure components) of the database was used for the test set. These components were randomly selected.

To obtain a valid model, several validation techniques should be used. The most widely used techniques have been presented by Todeschini et al.[4,8] Of those techniques, the bootstrapping, y-scrambling, and external validation techniques are used in this study.

Table 1. The Molecular Descriptors Which Were Entered to the Best Multivariate Linear Equation and Their Definitions

molecular descriptor	type	definition
nCIC	constitutional descriptors	number of rings
MAXDP	topological descriptors	maximal electrotopological positive variation
BEHm2	Burden eigenvalues	highest eigenvalue n.2 of Burden matrix/weighted by atomic mases
nHDon	functional group counts.	number of donor atoms for H-bonds (N and O)
O-057	atom-centered fragments	phenol/enol/carboxyl OH
TPSA(Tot)	molecular properties	topological polar surface area using N, O, S, P polar contributions
MLOGP2	molecular properties	squared Moriguchi octanol−water partition coefficient.

By the bootstrapping technique, the original size of the data set (n) is preserved for the training set, by the selection of n objects with repetition. In this procedure, the training set usually consists of repeated objects and the evaluation set of the objects left out. The model is calculated on the training set, and responses are predicted on the evaluation set. All the squared differences between the true response and the predicted response of the objects of the evaluation set are collected "PRESS". This procedure of building training sets and evaluation sets is repeated thousands of time. "PRESS" is summed and the average predictive power is calculated.[4]

The y-scrambling technique is adopted to check models with chance correlation. This test is performed by calculating the quality of the model (usually the Q^2), randomly modifying the sequence of the response vector by assigning to each object a response randomly selected from the true responses. If the original model has no chance correlation, there is a significant deference in the quality of the original model and that associated with a model obtained with random responses. The procedure is repeated several hundreds of times.[4]

External validation technique is a validation technique where a test is retained to perform a further check on the predictive capabilities of a model obtained from a training set and with predictive power optimized by an evaluation set.[4]

Results and Discussion

By the presented procedure, the best multivariate linear equation was obtained. For obtaining this equation, first, the best one-molecular descriptors model was obtained. Then the best two-molecular descriptors model was obtained. This procedure was repeated to obtain the best three, four, five, and so on molecular descriptors model. The best multivariate linear model has seven parameters because the increase in the number of molecular descriptors has no significant effect on the accuracy of the best model. This equation and its statistical parameters are presented as

$$\begin{gathered}\text{TPT} = 93.6142(\pm 5.7537) + 50.2233(\pm 0.9453)\text{nCIC} + \\ 6.1567(\pm 0.6071)\text{maxDP} + 19.7186(\pm 2.0117)\text{BEHm2} + \\ 24.9384(\pm 1.1571)\text{nHDon} + 41.0616(\pm 2.3504)\text{O-57} + \\ 1.0684(\pm 0.0398)\text{TPSA(Tot)} + 1.7126(\pm 0.0894)\text{MlogP2} \\ n_{\text{training}} = 511;\ n_{\text{test}} = 127;\ R^2 = 0.9410\ Q_{\text{Loo}}{}^2 = 0.9389; \\ Q_{\text{LTO}}{}^2 = 0.9177\ Q_{\text{BOOT}}{}^2 = 0.9381;\ Q_{\text{EXT}}{}^2 = 0.9468 \\ Q_{\text{EXT}}{}^2 = 0.9468,\ s = 13.811;\ a = 0.939;\ F = 1147.01 \\ \Delta K = 0.084;\ \Delta Q = 0.000;\ R^{\text{P}} = 0.046;\ R^{\text{N}} = 0.000\end{gathered} \qquad (6)$$

Table 2. Correlation Matrix of Seven Selected Descriptors

	nCIC	MAXDP	BEHm2	nHDon	O−057	TPSA(Tot)	MLOGP2
nCIC	1						
MAXDP	0.128	1					
BEHm2	0.305	0.128	1				
nHDon	0.86	0.205	0.073	1			
O−057	0.024	0.278	0.009	0.339	1		
TPSA(Tot)	0.093	0.056	0.082	0.42	0.216	1	
MLOGP2	0.238	0.237	0.476	0.177	0	0.323	1

where s is a residual mean square error, a is the y-scrambling parameter, and F is a Fisher function. Also the molecular descriptors and their physical meanings are presented in Table 1. For more information about procedure of calculation of these molecular descriptors from the chemical structure of a compound, please refer to the Dragon software user's guide.[5]

"nCIC" is the number of rings (cyclomatic number). It is logical that increase in number of rings in a molecule raises the phase change temperatures such as TPT as can be found in eq 6. "MAXDP" is a topological descriptor. These descriptors are based on a graph representation of the molecule. They are numerical quantifiers of molecular topology obtained by the application of algebraic operators to matrices representing molecular graphs and whose values are independent of vertex numbering or labeling. They can be sensitive to one or more structural features of the molecule such as size, shape, symmetry, branching, and cyclicity and can also encode chemical information concerning atom type and bond multiplicity. When this descriptor increases, the TPT increases, too. "BEHm2" is a Burden eigenvalue descriptor. These molecular descriptors were originally proposed to address searching for chemical similarity/diversity on large databases. Increase in the value of this descriptor causes an increase in TPT. "nHDon" is the number of donor atoms for N and H bonds. It is a measure of the hydrogen-bonding ability of a molecule expressed in terms of number of possible hydrogen-bond donors. Specifically, it is calculated by adding up the hydrogens bonded to any nitrogen and oxygen in the molecule. It is clear that increase in value of this descriptor raises the TPT. "O-057" is the number of phenol/enol/carboxyl OH. When this descriptor increases, the TPT increases, too. "TPSA(Tot)" is a measure of polarity of a molecule. When this descriptor increases, There is an expected increase in phase change data such as TPT. "MLOGP2" is squared Moriguchi octanol−water partition coefficient and is a measure of hyrophobicity. When this descriptor increases, the TPT increases.[4]

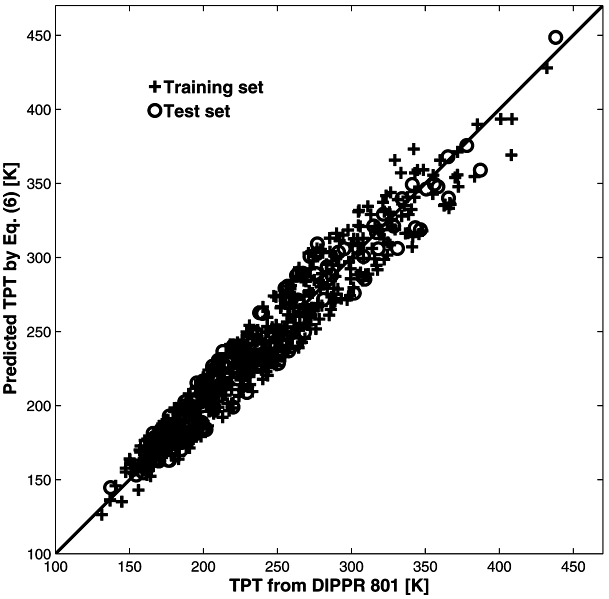

Figure 1. The comparison between the predicted values of TPT by the obtained model and DIPPR 801 data for training set and test set.

The $n_{\text{trainiing}}$ and n_{test} are the number of available pure components in the training set and the test set, respectively. For checking validity of the model, bootstrap technique, y-scrambling, and external validation techniques were used.[4,8] The bootstrapping was repeated 5000 times. Also y-scrambling was repeated 300 times. As can be seen, the difference between, Q_{Loo}^2, Q_{BOOT}^2, Q_{EXT}^2, and R^2 show that the obtained model is a good model and has good predictive power. When the number of the objects in the data set is quite large (such as in this work), the predictive ability obtained is too optimistic. This is due to a too small perturbation of the data when only one object is left out. Therefore, in these types of problems, the leave-more-out cross validation technique is used. The leave-10-out cross validation was used for this purpose. This technique was repeated 100 times over 100 random splits of training-test sets. The average of the cross validation coefficient was equal 0.9177 (Q_{LTO}^2). Also the intercept value of the y-scrambling technique has low value ($a = 0.939$) that reveals the model is valid. In addition the values of four constraints of the model are equal or greater than zero which shows that this model is valid and is not chance correlation.

Table 2 presents the correlation matrix. It is clear that the seven selected descriptors are not highly correlated.

All of the validation techniques show the obtained model is a valid model and can be used to predict TPT of pure components.

The predicted values of TPT by eq 6 in comparison with the DIPPR 801 data are presented in Figure 1. As Supporting Information, the values of the predicted TPT in comparison with the DIPPR 801 data are presented. Also, the values of the

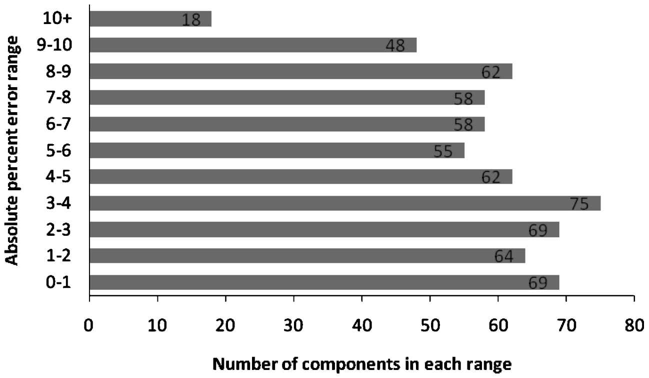

Figure 2. The absolute percent error of the obtained model over 638 pure components. The absolute percent error is defined as $100 \times |(y_{\text{exp}} - y_{\text{calcd}})/y_{\text{exp}}|$.

descriptors and status of all components (training set or test set) are presented as Supporting Information.

Conclusion

In this paper a new simple QSPR model was presented for prediction of the triple-point temperature (TPT) of pure components. This model is a multivariate linear model, which has seven variables (molecular descriptors). These seven molecular descriptors were selected using GA-MLR technique. These variables are calculated on the basisof the chemical structure molecules.

The absolute percent error of the obtained model is shown in Figure 2.

The presented model is simple and quite accurate. For developing this model, 638 pure components were used. As a result, the range of application of this model is wide and it can be used for prediction of TPT for any desired regular chemical structure.

Supporting Information Available: Values of the predicted TPT in comparison with the DIPPR 801 data; values of the descriptors and status of all components (training set or test set). This material is available free of charge via the Internet at http://pubs.acs.org.

Literature Cited

(1) Walas, S. M. *Phase Equilibria in Chemical Engineering*; Butterworth Publishers: Boston, MA, 1985.

(2) Poling, B. E.; Prausnitz, J. M.; O'Connell, J. P. *The Properties of Gases and Liquids*, 5th ed.; McGraw-Hill, New York, 2000.

(3) *Project 801*; Public Release Documentation; American Institute of Chemical Engineers (AIChE), Design Institute for Physical Properties (DIPPR): Brigham Young University, Utah, 2006.

(4) Todeschini, R.; Consonni, V. In *Handbook of Molecular Descriptors*; Manhold, R.; Kubinyi, H.; Temmerman H., Eds.; Wiley-VCH: Weinheim, Germany, 2000.

(5) Talete srl, Dragon for Windows (Software for Molecular Descriptor Calculation), version 5.4, http://www.talete.mi.it, 2006.

(6) *Hyperchem Release 7.5 for Windows*; Hypercube, Inc.: Gainesville, FL, 2002.

(7) Leardi, R.; Boggia, R.; Terrile, M. Genetic Algorithms as a Strategy for Feature Selection. *J. Chemom.* **1992**, *6*, 267.

(8) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. Detecting "Bad" Regression Models: Multicriteria Fitness Functions in Regression Analysis. *Anal. Chim. Acta* **2004**, *515*, 199.

(9) Gharagheizi, F. QSPR Analysis for Intrinsic Viscosity of Polymer Solutions by Means of GA-MLR and RBFNN. *Comput. Mater. Sci.* **2007**, *40*, 159.

(10) Gharagheizi, F. Quantitative Structure−Property Relationship for Prediction of the Lower Flammability Limit of Pure Compounds. *Energy Fuels* **2009**, *22*, 3037.

(11) Gharagheizi, F.; Mehrpooya, M. Prediction of Standard Chemical Exergy by a Three Descriptors QSPR Model. *Energy Convers. Manage.* **2007**, *48*, 2453.

(12) Gharagheizi, F.; Alamdari, R. F. A Molecular-Based Model for Prediction of Solubility of C_{60} Fullerene in Various Solvents. *Fullerenes, Nanotubes, Carbon Nanostruct.* **2008**, *16*, 40.

(13) Gharagheizi, F. QSPR Studied For Solubility Parameter By Means of Genetic Algorithm-Based Multivariate Linear Regression and Generalized Regression Neural Network. *QSAR Comb. Sci.* **2008**, *27*, 165.

(14) Gharagheizi, F. A Simple Equation For Prediction of Net Heat of Combustion of Pure Chemicals. *Chemom. Intell. Lab. Syst.* **2008**, *91*, 177.

(15) Gharagheizi, F. A New Molecular-Based Model for Prediction of Enthalpy of Sublimation of Pure Components. *Thermochim. Acta* **2008**, *469*, 8.

(16) Gharagheizi, F.; Alamdari, R. F. Prediction of Flash Point Temperature of Pure Components Using a Quantitative Structure−Property Relationship Model. *QSAR Comb. Sci.* **2008**, *27*, 679.

(17) Gharagheizi, F.; Fazeli, A. Prediction of the Watson Characterization Factor of Hydrocarbon Components from Molecular Properties. *QSAR Comb. Sci.* **2008**, *27*, 758.

(18) Sattari, M.; Gharagheizi, F. Prediction of Molecular Diffusivity of Pure Components into Air: A QSPR Approach. *Chemosphere* **2008**, *72*, 1298.

(19) Vatani, A.; Mehrpooya, M.; Gharagheizi, F. Prediction of Standard Enthalpy of Formation by a QSPR Model. *Int. J. Mol. Sci.* **2007**, *8*, 407.

(20) Gharagheizi, F.; Sattari, M. Estimation of Molecular Diffusivity of Pure Chemicals in Water: A Quantitative Structure−Property Relationship Study. *SAR QSAR Environ.* **2009**, *20*, 267.

(21) Gharagheizi, F.; Mehrpooya, M. Prediction of Some Important Physical Properties of Sulfur Compounds Using QSPR Models. *Mol. Diversity* **2008**, *12*, 143.

(22) Gharagheizi, F.; Tirandazi, B.; Barzin, R. Estimation of Aniline Point Temperature of Pure Hydrocarbons: A Quantitative Structure−Property Relationship Approach. *Ind. Eng. Chem. Res.* **2009**, *48*, 1678.

(23) Gharagheizi, F. A QSPR Model for Estimation of Lower Flammability Limit Temperature of Pure Compounds Based on Molecular Structure. *J. Hazard. Mater.* **2009**, *169*, 217.

(24) Gharagheizi, F. Prediction of Upper Flammability Limit Percent of Pure Compounds from Their Molecular Structures. *J. Hazard. Mater.* **2009**, *167*, 507.

(25) Gharagheizi, F.; Sattari, M. Prediction of θ(UCST) of Polymer Solutions: A Quantitative Structure−Property Relationship Study. *Ind. Eng. Chem. Res.* **2009**, *48*, 9054.

(26) Gharagheizi, F. Prediction of Standard Enthalpy of Formation of Pure Compounds Using Molecular Structure. *Aust. J. Chem.* **2009**, *62*, 374.

Received for review June 25, 2009
Revised manuscript received November 17, 2009
Accepted December 8, 2009

IE901029M