Available online at www.sciencedirect.com
ScienceDirect
Thermochimica Acta 469 (2008) 8–11
thermochimica acta
www.elsevier.com/locate/tca

A new molecular-based model for prediction of enthalpy of sublimation of pure components

Farhad Gharagheizi *

Department of Chemical Engineering, Faculty of Engineering, University of Tehran, P.O. Box 11365-4563, Tehran, Iran

Received 8 October 2007; received in revised form 4 December 2007; accepted 12 December 2007
Available online 23 December 2007

Abstract

A quantitative structure property relationship (QSPR) study was performed to develop a model for prediction of enthalpy of sublimation of pure components. For developing this model, 1348 pure components were used, and for each of them, 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 five-parameter multi-linear equation that has a squared correlation coefficient of 0.9746 ($R^2 = 0.9746$).

Keywords: QSPR; GA-MLR; Enthalpy of sublimation

1. Introduction

The enthalpy of sublimation of a component may be defined as the molar change in enthalpy when the solid is isothermally converted into a gas at its triple point. This property is of certain practical interest for the chemistry of the crystalline state and, in particular, for resolving the problems associated with dispersion of materials, and such ecological problems as transport of organic contaminant in the atmosphere, etc. [1–5].

There are several models to estimate the enthalpy of sublimation of pure components. Rice et al. [6] presented a molecular-based model to estimate enthalpy of sublimation using the properties associated with quantum mechanically determined electrostatic potentials of isolated molecules. The root mean squares of error and the maximum deviation of their model over 35 pure components are respectively 15 and 52 kJ/mol (the unit is converted from kcal/mol in Ref. [6] to kJ/mol). Politzer et al. [7], Mathie and Somonetti [8] and Kim et al. [9] used different modifications of van der Waals electrostatic surface potentials and their derived properties to correlate the enthalpy of sublimation. Their models showed good results over 34 pure components used to their studies. Ouvrad and Mitchell [10] presented a simple model using the number of occurrences of different atom types as descriptors for prediction of the enthalpy of sublimation. The squared correlation coefficient of their model over 226 pure components used as training set, and 35 pure components as test set, are 0.925 and 0.937. Politzer et al. [11] presented a model to estimate the enthalpy of sublimation of pure components on the basis of the calculated electrostatic potential on the molecular surface. The average absolute deviation of their correlation over 66 pure components is 11.7 kJ/mol (the unit is converted from kcal/mol in Ref. [11] to kJ/mol). Recently, Byrd and Rice [12] presented a model to estimate the enthalpy of sublimation using quantum chemical data. The root mean squares of error and the maximum deviation of their model over 35 pure components are respectively 12.5 and 217.7 kJ/mol (the unit is converted from kcal/mol in Ref. [12] to kJ/mol).

Another type of correlations used to predict enthalpy of sublimation is quantitative structure property relationships (QSPR). QSPR relates a property of interest, defined quantitatively by a numerical measure, to characteristic molecular descriptors derived theoretically from the chemical structures of the components.

From this type of correlations, Charlton et al. [1] used neural networks, theoretical crystal packing calculations, and multi-linear regression for prediction of enthalpy of sublimation for a set of 62 organic components. Puri et al. [4] applied three-dimensional quantitative structure property relationship (3D-QSPR) using comparative molecular field analysis

* Fax: +98 21 66957784.
E-mail address: fghara@engmail.ut.ac.ir.

0040-6031/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.tca.2007.12.005

(CoMFA) for prediction of enthalpy of sublimation of polychlorinated biphenyls. Zhokhova et al. [5] used fragment approach based on QSPR for prediction of the enthalpy of sublimation for a set of 72 organic components.

All of the previously presented models are useful but most of them have been developed over a small data set. On the other hand, in each work, small types of molecular-based parameters have been considered to develop these models.

As a result, in this work, using a large data set of pure components, and also, using a large pool of molecular-based parameters, enthalpy of sublimation is correlated.

2. Materials and methods

2.1. Data set

Evaluated databases such as DIPPR 801 database [13] are useful tools for developing new property prediction models. DIPPR 801 is recommended by American Institute of Chemical Engineers (AIChE). In this study, 1348 pure components were selected and their values of enthalpy of sublimation were extracted. These components and their values of enthalpy of sublimation are presented as Supplementary materials.

2.2. Determination of molecular descriptors

In this step, the molecular structures of all 1348 pure components were drawn into Hyperchem software [14] and optimized using the MM+ [15] molecular mechanics force field. Thereafter, using these optimized molecular structures; molecular descriptors were calculated by Dragon software [16]. Dragon software can calculate 1664 molecular descriptors for every molecule. For more information about the types of the molecular descriptors which Dragon can calculate, and the procedure of calculation of the descriptors, refer to Dragon software user's guide [16].

2.3. GA-MLR calculations

Generally, in QSPR studies, after calculating molecular descriptors, the problem is to find a linear equation that can predict the desired property with the least number of variables as well as with the highest accuracy.

In other words, the problem is to find a subset of variables (most statistically effective molecular descriptors on enthalpy of sublimation) from all available variables (all molecular descriptors) so that can predict enthalpy of sublimation, with minimum error in comparison to 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 used to select best subset variables with respect to an objective function. This algorithm was presented by Leardi et al. for the first time [17].

In this study, the GA-MLR technique presented by Leardi et al. [17] with RQK function presented by Todeschini et al. [18] was used to subset variable selection. This methodology has been extensively presented in the previous works of the author and the results are satisfactory [19–27].

Before performing GA-MLR technique, the data set must be divided into two new collections. First one is allocated for training and second one is allocated for testing. By means of the training set, the best model is found and then the predictive power of the obtained model was checked by the test set as external dataset. In this work, 80% of the database was used for training set and 20% for test set (from 1348 components, 1079 components are in the training set and 269 components are in the test set). The selection was randomly done.

The inputs of our program are the pool of molecular descriptors, the enthalpy of sublimation of pure components, and the number of molecular descriptors which we want to enter into our final model.

To obtain the best multivariate linear equation, all molecular descriptors must be introduced to the program and the minimum number of possible variables must be tested at the starting point. So running the program is started with one variable. After running the program, we must obtain the best multivariate linear model. In the next steps, we increase the number of desired variables to two, three, four, and so on, and we must repeat all calculations for them.

When we saw that increasing in the number of variables has no considerable effect on the accuracy of the best-obtained model, the calculations must be stopped, because the best multivariate linear model has been obtained.

3. Results and discussion

By presented procedure, the best multivariate linear equation was obtained. This multivariate linear model has five parameters. This equation is

$$\begin{aligned}\Delta H_{\text{sub}} = {} & 15.3238(\pm 0.4246) - 2.046(\pm 0.05)\text{ZM1} \\ & + 5.1782(\pm 0.4199)\text{X1sol} + 12.3669(\pm 0.3263)n\text{ROH} \\ & + 0.401(\pm 0.0085)\text{TPSA(NO)} \\ & + 12.3991(\pm 0.3781)\text{VRv1} \end{aligned} \quad (1)$$

$n_{\text{training}} = 1079$; $n_{\text{test}} = 269$; $R^2 = 0.9746$; $Q^2_{\text{LOO}} = 0.9740$; $Q^2_{\text{BOOT}} = 0.9737$; $Q^2_{\text{EXT}} = 0.9758$; $s = 5.46$; $a = -0.029$; $F = 8229.781$; where ΔH_{sub} is the enthalpy of sublimation in kJ/mol unit.

Table 1
The five molecular descriptors entered into the best obtained multi-linear equation (Eq. (1))

ID	Molecular descriptor	Type	Definition
1	ZM1	Topological descriptor	First Zagreb index M1
2	X1sol	Connectivity index	Solvation connectivity index χ1
3	*n*ROH	Functional group count	Number of hydroxyl groups
4	TPSA(NO)	Molecular property	Topological polar surface area using N, O polar contributions
5	VRv1	Eigenvalue-based index	Randic-type eigenvector-based index from van der Waals weighted distance matrix

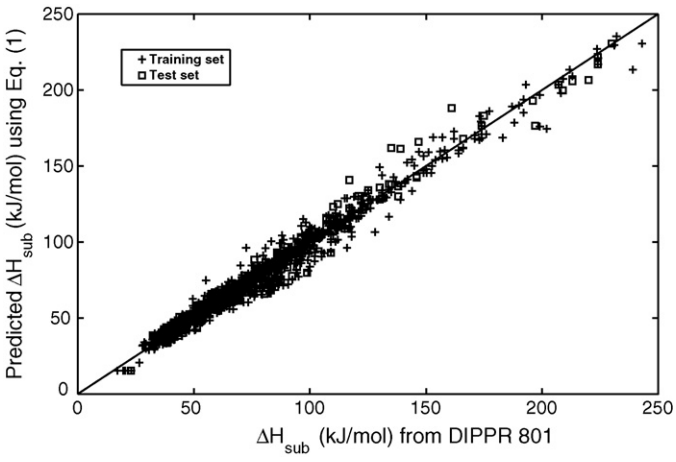

Fig. 1. Comparison between the best multi-linear results obtained by GA-MLR and the DIPPR 801 data.

The molecular descriptors and their physical meanings are presented in Table 1.

As can be found from Table 1 and Eq. (1), ΔH_{sub} increases with X1sol, *n*ROH, TPSA(NO) and VRv1, and it decreases with ZM1. X1sol is defined in order to model and describe dispersion interactions, also *n*ROH and TPSA(NO) describe the hydrogen bonding and some type of charge interactions [28]. VRv1 is useful molecular descriptor for presenting the size and shape of molecules [28]. For isomeric components, ZM1 presents the molecular branching [28].

$n_{training}$ and n_{test} are the number of components of the training set and the test set, respectively. For more checking validity of the model, bootstrap technique, y-scrambling, and external validation techniques were used [18,28]. The bootstrapping was repeated 5000 times. Also y-scrambling was repeated 300 times. As can be seen the difference between, Q^2_{LOO}, Q^2_{BOOT}, Q^2_{EXT} and R^2 show that the obtained model is a good model and has good predictive power [28]. Also the intercept value of the y-scrambling technique has low value ($a = -0.029$) that reveals the validity of the model (the y-scrambling, bootstrapping, and external validation techniques have been extensively presented by Todeschini et al. [18,28]).

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

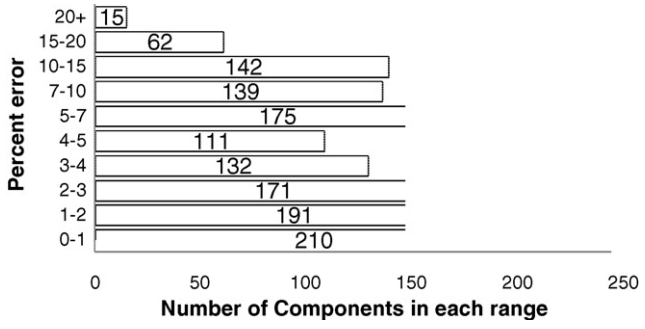

Fig. 2. Percent error obtained by Eq. (1) over all of 1348 pure components used in this study.

The predicted values of enthalpy of sublimation using Eq. (1) in comparison to the experimental data are presented in Fig. 1. The values of the predicted enthalpy of sublimation in comparison to the experimental data are presented as Supplementary materials. Also the values of the descriptors and status of all of the pure components (training set or test set) are presented as Supplementary materials.

4. Conclusion

In this study a simple QSPR model was presented based on molecular descriptors of Dragon software. GA-MLR technique with RQK fitness function was used to develop a multivariate linear model. Also, validity of the model was checked by several validation techniques. As a result, obtained model has predictive power and can be used to predict the enthalpy of sublimation of pure components. The squared correlation coefficient and root mean squares of error obtained by this equation over 1348 pure components are respectively, 0.9746 and 5.46 kJ/mol. Also, the maximum absolute deviation obtained by the model is equal to 27.56 kJ/mol and, it is related to dinonylphenol. Also the percent error obtained by Eq. (1) is schematically shown in Fig. 2.

The obtained results in this study are opposite to the conclusion presented by Byrd and Rice [12]. They concluded that QSPR models cannot be used more accurately than electrostatic potentials methods to estimate enthalpy of sublimation, but the presented model show better results than all previously presented models.

Since the model has been obtained using 1348 pure components which belong to diverse chemical groups, it can be used to predict the enthalpy of sublimation of any regular components.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tca.2007.12.005.

References

[1] M.H. Charlton, R. Docherty, M. Hutchings, J. Chem. Soc. Perkin Trans. 2 (11) (1995) 2023–2030.
[2] A. Gavezzotti, J. Am. Chem. Soc. 111 (1989) 1835–1843.
[3] A. Gavezzotti, J. Phys. Chem. 95 (1991) 8948–8955.
[4] S. Puri, J.S. Chickos, W.J. Welsh, J. Chem. Inf. Comput. Sci. 42 (2002) 109–116.
[5] N.I. Zhokhova, I.I. Baskin, V.A. Payulin, A.N. Zeforov, N.S. Zefirov, Russian J. Appl. Chem. 76 (2003) 1966–1970.
[6] B.M. Rice, S.V. Pai, J. Hare, Combust. Flame 118 (1999) 445–458.
[7] P. Politzer, J.S. Murray, M.E. Grice, M. Deslvo, E. Miller, Mol. Phys. 91 (1997) 923–928.
[8] D. Mathie, P. Somonetti, Thermochim. Acta 384 (2002) 369–375.
[9] C.Y. Kim, K.A. Lee, K.H. Hyun, J. Park, I.Y. Kwack, C.K. Kim, H.W. Lee, B.S. Lee, J. Comput. Chem. 25 (2004) 2073–2079.
[10] C. Ouvrad, J.B.O. Mitchell, Acta Crystallogr. B 59 (2003) 676–685.
[11] P. Politzer, Y. Ma, P. Lane, M.C. Concha, Int. J. Quant. Chem. 105 (2005) 341–347.
[12] E.F.C. Byrd, B.M. Rice, J. Phys. Chem. A 110 (2006) 1005–1013.
[13] American Institute of Chemical Engineers, Project 801, Evaluated Process Design Data, Public Release Documentation, Design Institute for Physical Properties (DIPPR), American Institute of Chemical Engineers (AIChE), 2006.

[14] Hypercube Inc., HyperChem Release 7.5 for Windows, Molecular Modeling System, Hypercube Inc., 2002.
[15] J. Lii, S. Gallion, C. Bender, H. Wikström, N.L. Allinger, K.M. Flurchick, M.M. Teeter, J. Comput. Chem. 10 (1989) 503–513.
[16] Talete srl, Dragon for windows (Software for Molecular Descriptor Calculations), Version 5.4, 2006. http://www/talete.mi.it/.
[17] R. Leardi, R. Boggia, M. Terrile, J. Chemometr. 6 (1992) 267–281.
[18] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, Anal. Chim. Acta 515 (2004) 199–208.
[19] F. Gharagheizi, Comput. Mater. Sci. 40 (2007) 159–167.
[20] F. Gharagheizi, M. Mehrpooya, Energ. Convers. Manage. 48 (2007) 2453–2460.
[21] F. Gharagheizi, e-Polymers, 2007, no. 114.
[22] A. Vatani, M. Mehrpooya, F. Gharagheizi, Int. J. Mol. Sci. 8 (2007) 407–432.
[23] F. Gharagheizi, QSAR&Comb. Sci., doi:10.1002/qsar.200630159, in press.
[24] F. Gharagheizi, R.F. Alamdari, QSAR&Comb. Sci., doi:10.1002/qsar.200630110, in press.
[25] F. Gharagheizi, A. Fazeli, QSAR&Comb. Sci., doi:10.1002/qsar.200730020, in press.
[26] F. Gharagheizi, Chemometr. Intell. Lab. Syst., doi:10.1016/j.chemolab.2007.11.003, in press.
[27] F. Gharagheizi, R.F. Alamdari, Fuller. Nanotub. Car. N., doi:10.1080/15363830701779315, in press.
[28] R. Todeschini, V. Consonni, in: R. Manhold, H. Kubinyi, H. Temmerman (Eds.), Handbook of Molecular Descriptors, Wiley–VCH, Weinheim, 2000.