Journal of Hazardous Materials 169 (2009) 217–220

Contents lists available at ScienceDirect

Journal of Hazardous Materials

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

A QSPR model for estimation of lower flammability limit temperature of pure compounds based on molecular structure

Farhad Gharagheizi [a,b,*]

[a] Department of Chemical Engineering, Faculty of Engineering, University of Tehran, P.O. Box 11365-4563, Tehran, Iran
[b] Department of Chemical Engineering, Medicinal Plants and Drugs Research Institute, Shahid Beheshti University, Evin, Tehran, Iran

ARTICLE INFO

Article history:
Received 9 December 2008
Received in revised form 15 March 2009
Accepted 18 March 2009
Available online 27 March 2009

Keywords:
Lower flammability limit temperature (LFLT)
Lower flash point
Lower explosion point
Temperature limit of flammability
Safety
Fire
Flammability

ABSTRACT

In this study, a quantitative structure–property relationship was presented to estimate lower flammability limit temperature (LFLT) of pure compounds. This relationship is a multi-linear equation and has six parameters. These chemical structure-based parameters were selected from 1664 molecular-based parameters by genetic algorithm multivariate linear regression (GA-MLR). Since 1171 compounds were used to develop this equation, the model can be used to estimate the LFLT of a wide range of pure compounds.

1. Introduction

To safe handling, transportation, and storage of flammable compounds, information about flammability of these compounds is needed [1]. One of the most important parameters used to this purpose is lower flammability limit temperature (LFLT) [2].

The lower flammability limit temperature (LFLT) or temperature limit of flammability or lower explosion point or lower flash point is the temperature related to 1.01325 bar of pressure, at which the concentration of a saturated vapor/air mixture equals the lower flammability limit. In other words, the LFLT is the minimum temperature at which liquid or solid compounds evolve sufficient vapor to form a flammable mixture with air under equilibrium conditions. The LFLT is measured under ASTM Test E 1232. In this test method, the LFLT is defined as the lowest temperature (corrected to standard atmospheric pressure of 101 kPa) at which application of an ignition source causes a homogenous mixture of a gaseous oxidizer and vapors in equilibrium with liquid (or solid) sample to ignite and propagate a flame away from the ignition source [3].

The LFLT of a pure compound can be calculated from its vapor pressure curve and the lower flammability limit. Practically, LFLT is the lowest temperature at which the mixtures of vapor or gas with air, if ignited will just propagate flame.

The definition of the LFLT is like the definition of flash point and in principle, the FP (The FP is defined as the temperature at which it can form an ignitable mixture with air.) and LFLT should be the same, but due to the differences in determination methods, the LFLT is lower than the FP.

There is an important difference between these two properties (LFLT and FP). The FP is reached when a flame propagates from an ignition source such as external flame through the vapor–air mixture but, LFLT is essentially independent of the ignition source strength [1,2]. As a result, it can be concluded that the LFLT always has lower value in comparison with the FP. This result has been experimentally confirmed [2]. Therefore, attention to this result is very important and it can be found that LFLT is more important than FP in evaluation of safely operating an industrial processes. In other words, operating at temperatures below the FP may not be sufficient safety, but operating at temperatures below the LFLT gives sufficient safety [2].

The LFLT is one of the important safety parameters used in design safe operational conditions in those equipments such as vessels and storage tanks which, the equilibrium conditions occur [3].

The performed literature survey showed that there is no computational method to estimate or predict LFLT. Then the

* Correspondence address. Department of Chemical Engineering, Faculty of Engineering, University of Tehran, P.O. Box 11365-4563, Tehran, Iran.
Tel.: +98 21 66957784; fax: +98 21 66957784.
E-mail addresses: fghara@ut.ac.ir, fghara@gmail.com.

0304-3894/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhazmat.2009.03.083

estimation of LFLT of pure compounds is the subject of this study.

One of the most widely used methods applied to relate physical and chemical properties to the chemical structure of compounds is quantitative structure–property relationship (QSPR). In this methodology, the desired property is correlated using molecular-based parameters called "molecular descriptors". Molecular descriptors are computed only from chemical structure of a molecule using the known mathematical algorithms. Application of this methodology to correlate various physical and chemical properties has been showed promising results [4–11]. Therefore in this study, this methodology is used to develop a molecular-based model to predict LFLT of pure compounds.

2. Materials and methods

2.1. Data set

Evaluated databases such as DIPPR 801 database [12] are useful tools for developing new property prediction models. DIPPR 801 is recommended by American Institute of Chemical Engineers (AIChE) for physical properties of pure compounds. In this study, 1171 pure compounds were found in this database and their LFLT were extracted and used as main dataset. These compounds and their LFLT values are presented as supplementary materials.

2.2. Determination of molecular descriptors

In this step, the molecular structures of all 1171 pure compounds were drawn into Hyperchem software [13] and optimized using the MM+ molecular mechanics force field. Since the values of some types of molecular descriptors are dependent to bonds lengths and bonds angles, the real values for these parameters are needed, therefore, the optimized chemical structures are necessary to obtain true values for molecular descriptors. Thereafter, using these optimized molecular structures; molecular descriptors were calculated by Dragon software [14]. Dragon software can calculate 1664 molecular descriptors for every molecule. Of course, these molecular descriptors have been calculated for approximately 234,000 pure compounds using Dragon software and are accessible from Milano chemometrics and QSAR research group web site (http://michem.disat.unimib.it/mole_db). 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 [14].

2.3. GA-MLR calculations

Usually, in QSPR methodology, after computing 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 of LFLT) from all available variables (all molecular descriptors) so that can predict LFLT, with minimum error in comparison with the available 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. Application of the genetic algorithm for subset variable selection was presented by Leardi et al. for the first time [15].

In this study, the GA-MLR technique presented by Leardi et al. [15] with RQK objective function presented by Todeschini et al. [16] was used to subset variable selection. This methodology has been extensively presented in the previous works of the author and the results are satisfactory [4–11].

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 1171 compounds, 937 compounds are in the training set and 234 compounds are in the test set). The selection was randomly done.

The inputs of our program are the pool of molecular descriptors, the LFLT of pure compounds, 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 six parameters. This equation is:

$$\begin{aligned}
&\mathrm{LFLT} = 17.1766(\pm 4.8632) + 213.3319(\pm 8.6995)\mathrm{Mv} + 5.0667(\pm 00.0676)\mathrm{CID} + 23.4601(\pm 0.7897)\mathrm{EEig02d} \\
&\quad -7.9145(\pm 0.3629)\mathrm{GGI1} + 38.5377(\pm 1.42097)\mathrm{nROH} + 20.692(\pm 1.0474)\mathrm{nHDon} \\
&n_{\mathrm{training}} = 937;\ n_{\mathrm{test}} = 234;\ R^2_{\mathrm{training}} = 0.9459;\ Q^2_{\mathrm{LOO}} = 0.9448;\ Q^2_{\mathrm{BOOT}} = 0.9443;\ Q^2_{\mathrm{EXT}} = 0.9495; \\
&\quad s = 15.613;\ a = -0.017;\ F = 8229.781; \\
&\mathrm{RQK\ function\ parameters:}\ \ \Delta K = 0.075;\ \Delta Q = 0.000;\ R^P = 0.001;\ R^N = 0.000
\end{aligned} \tag{1}$$

where LFLT is in Kelvin unit.

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

"Mv" is mean atomic van der Waals volume. This parameter is a measure of the size of a molecule. When the size of a molecule increases the LFLT of that molecule increases. In other words, the flammability of a molecule decreases when its size increases. "CID" is a molecular ID number. These types of molecular descriptors are proposed to unequivocally identify a molecule by a single real number. When this parameter increases, the LFLT increases, too. "EEig02d" belongs to edge adjacency indices. It is a measure of polarity of a molecule. When the polarity of molecule is increases the "EEig02d" increases and therefore, the LFLT is increases. "GGI1" is a topological charge index. These molecular descriptors are proposed to evaluate the charge transfer between pairs of atoms, and therefore the global charge transfer in a molecule. Increase in this parameter in a molecule causes decrease in the LFLT of that molecule. "nROH" and "nHDon" are related to the functional groups. These descriptors are measures of special types of interactions such as hydrogen bonds and other related interactions. Existence of these

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

ID	Molecular descriptor	Type	Definition
1	Mv	Constitutional descriptors	Mean atomic van der Waals volume (scaled on carbon atom)
2	CID	Walk and path counts	Randic I.D. number
3	EEig02d	Edge adjacency indices	Eigenvalue 02 from edge adjacency matrix weighted by dipole moment
4	GGI1	Topological charge indices	Topological charge index of order 1
5	nROH	Functional group counts.	Number of hydroxyl groups
6	nHDon	Eigenvalue-based index	Number of donor atoms for H-bonds (N and O)

types of interactions causes to increase stability of a molecule and therefore increase in the LFLT of that molecule [16].

$n_{\text{trainiing}}$ and n_{test} are the number of compounds 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 [16]. 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_{training} show that the obtained model is a good model and has good predictive power [16]. Also the intercept value of the y-scrambling technique has low value ($a = -0.017$) that reveals the validity of the model (The y-scrambling, bootstrapping, and external validation techniques have been extensively presented by Todeschini et al. [16].).

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

To evaluate the applicability domain (AD) of a QSPR model, application of a plot of standardized cross-validated residuals versus leverage (Hat diagonal) values was suggested by Gramatica [17]. This plot is called William plot. This method has been explained in details in Ref. [17]. This simple plot helps to identify both the response outliers and structurally influential chemicals in the model. As stated by Gramatica, those compounds with cross-validated standardized residuals grater than three standard deviation units are response outliers. Also, those compounds with Hat values grater than a critical Hat value are influential compounds in the model. Gramatica used 2.5 times of average of Hat values for this critical Hat value.

The William plot for Eq. (1) is presented in Fig. 1. Based on the explanations presented by Gramatica, in Fig. 1, compound 58 (1,4-benzendiamine) is truly predicted by the model but because high leverage value, as defined by the Hat vertical line it is outside of the AD. As can be found there is no outlier in the test set used in this study. Prediction of all other 10 compounds belong to test set and lie between 2 vertical lines and has Hat value greater than critical Hat value are reliable because in this area there are three compounds belong to the training set. Therefore, these three compounds are influential in model development. Also, compound 827 (furan) is wrongly predicted but it belongs to the AD of the model. This erroneous prediction could probably be attributed to wrong experimental data rather than to molecular structure.

The predicted values of LFLT using Eq. (1) in comparison with the DIPPR 801 data are presented in Fig. 2. The values of the predicted LFLT in comparison to the DIPPR 801 data are presented as supplementary materials. Also the values of the descriptors and status of all of the pure compounds (training set or test set) are presented as supplementary materials.

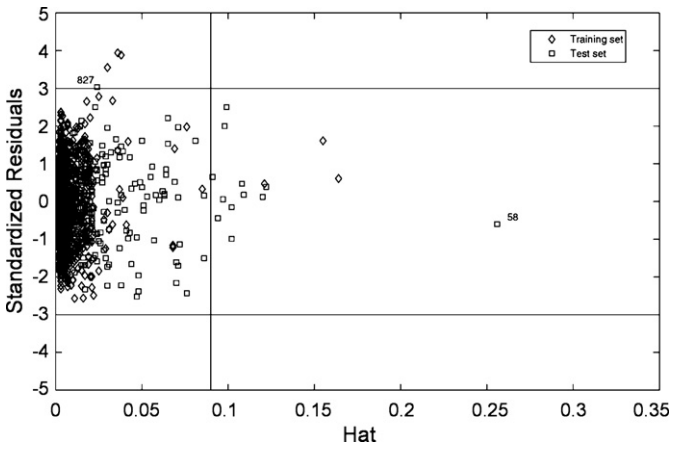

Fig. 1. The William plot for the Eq. (1) as stated by Gramatica [17].

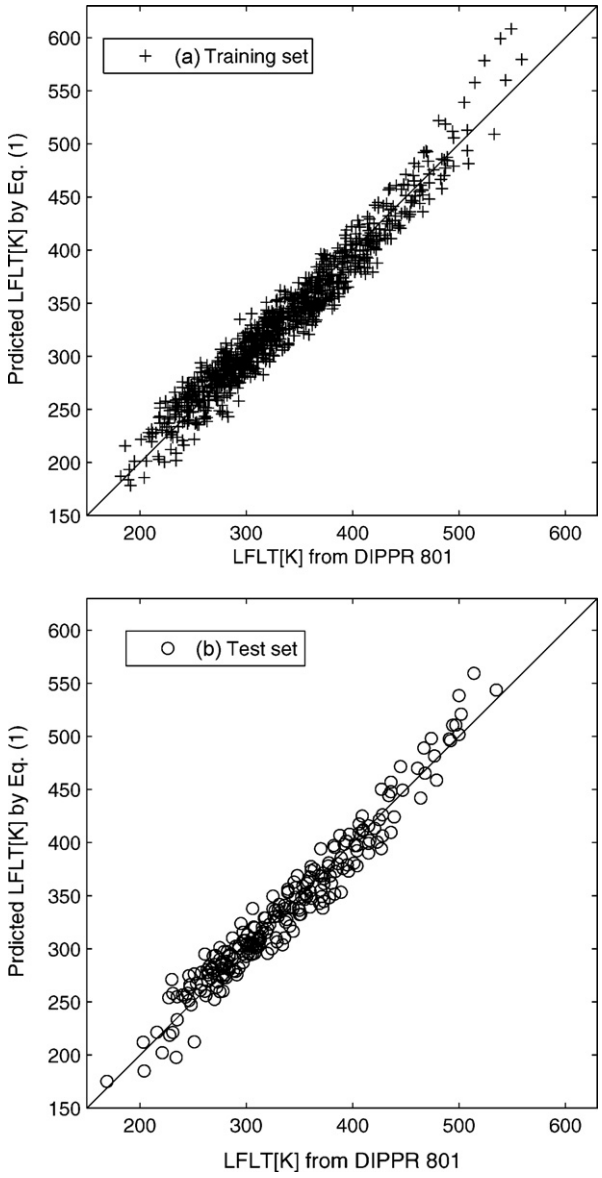

Fig. 2. Comparison between the predicted LFLT by Eq. (1) and DIPPR 801 data.

Table 2
Statistical parameters of the obtained model.

Statistical parameter	Value
Training set	
R^2	0.9459
Average absolute deviation	3.98%
Standard deviation error	15.554
Root mean square error	15.613
n	937
Test set	
R^2	0.9527
Average absolute deviation	3.91%
Standard deviation error	15.174
Root mean square error	15.406
n	234
Training set + test set	
R^2	0.9468
Average absolute deviation	3.96%
Standard deviation error	15.565
Root mean square error	15.611
n	1171

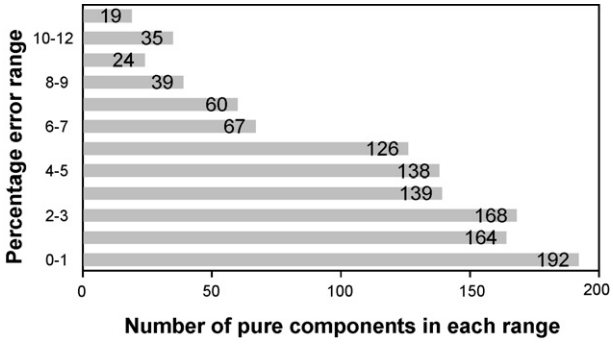

Fig. 3. Percent error of predicted LFLT by Eq. (1) over all of 1171 pure compounds used in this study.

The results obtained by model are presented in Table 2. These results show that the squared correlation coefficient, average absolute deviation, standard deviation error, and root mean square error of the model over the training set and the test set and the main data set are, respectively, 0.9459, 0.9527, 0.9468, 3.98%, 3.91%, 3.96%, 15.554, 15.174, 15.565, 15.613, 15.406, and 15.611.

4. Conclusion

In this study a simple molecular-based model was presented to predict lower flammability limit temperature (LFLT) of pure compounds. Also, validity and predictive power of the model was checked by several techniques. As a result, obtained model has predictive power and can be used to predict the LFLT of pure compounds. The squared correlation coefficient and root mean squares of error obtained by this equation over 1171 pure compounds are 0.9468 and 15.61 K. Also, the maximum absolute deviation obtained by the model is equal to 17.9% and, it is related to furan. Also the average absolute error of the model over all 1171 pure compounds is equal to 3.96%. Also, the percentage error obtained by Eq. (1) is schematically shown in the Fig. 3.

Since the model has been obtained using 1171 pure compounds which belong to diverse chemical groups, it can be used to predict the LFLT of every regular compound with some limitations. These 1171 pure compounds cover many families of compounds therefore the model has a wide range of applicability but, application of the model is restricted to those compounds similar to the compounds used to develop this model. Application of the model to those compounds which is completely different from compounds used to develop the model is not recommended.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jhazmat.2009.03.083.

References

[1] M. Vidal, W.J. Rogers, J.C. Holste, M.S. Mannan, A review of estimations method for flash points and flammability limits, Process Saf. Prog. 23 (2004) 47–55.
[2] E. Brandes, M. Mitu, D. Pawel, The lower explosion point—a good measure for explosion prevention: experiment and calculation for pure compounds and some mixtures, J. Loss Prevent. Proc. 20 (2007) 536–540.
[3] L.G. Britton, K.L. Cashdollar, W. Fenlon, D. Furip, J. Going, B.K. Harrison, J. Niemeier, E.A. Ural, Hazard assessment. Part II. Flammability and Ignitability, Process Saf. Prog. 24 (2005) 12–28.
[4] F. Gharagheizi, A new accurate neural network quantitative-structure–property relationship for prediction of θ (lower critical solution temperature) of polymer solutions, e-polymers, 2007, Article Number 114.
[5] F. Gharagheizi, A simple equation for prediction of net heat of combustion of pure chemicals, Chemometr. Intell. Lab. Syst. 91 (2008) 177–180.
[6] F. Gharagheizi, A new molecular-based model for prediction of enthalpy of sublimation of pure components, Thermochim. Acta 469 (2008) 8–11.
[7] F. Gharagheizi, R.F. Alamdari, Prediction of flash point temperature of pure components using a quantitative structure–property relationship model, QSAR Comb. Sci. 27 (2008) 679–683.
[8] M. Sattari, F. Gharagheizi, Prediction of molecular diffusivity of pure components into air: a QSPR approach, Chemosphere 72 (2008) 1298–1302.
[9] A. Vatani, M. Mehrpooya, F. Gharagheizi, Prediction of standard enthalpy of formation by a QSPR model, Int. J. Mol. Sci. 8 (2007) 407–432.
[10] F. Gharagheizi, M. Sattari, Prediction of some important physical properties of sulfur compounds using QSPR models, Mol. Divers. 12 (2008) 143–155.
[11] F. Gharagheizi, B. Tirandazi, R. Barzin, Estimation of aniline point temperature of pure hydrocarbons: a quantitative structure–property relationship approach, Ind. Eng. Chem. Res. 48 (2009) 1678–1682.
[12] Project 801, Evaluated Process Design Data, Public Release Documentation, Design Institute for Physical Properties (DIPPR), American Institute of Chemical Engineers (AIChE) 2006.
[13] HyperChem Release 7.5 for Windows, Molecular Modeling System, Hypercube Inc., 2002.
[14] Talete srl, Dragon for windows (Software for molecular Descriptor Calculations), Version 5.4, 2006. (http://www/talete.mi.it/).
[15] R. Leardi, R. Boggia, M. Terrile, Genetic algorithms as a strategy for feature selection, J. Chemometr. 6 (1992) 267–281.
[16] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, Detecting "bad" regression models: multicriteria fitness function in regression analysis, Anal. Chim. Acta 515 (2004) 199–208.
[17] P. Gramatica, Principles of QSAR models validation: internal and External, QSAR Comb. Sci. 26 (2007) 694–701.