Journal of Hazardous Materials 167 (2009) 507–510

Contents lists available at ScienceDirect

Journal of Hazardous Materials

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

Prediction of upper flammability limit percent of pure compounds from their molecular structures

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 20 September 2008
Received in revised form
26 December 2008
Accepted 7 January 2009
Available online 15 January 2009

Keywords:
Upper flammability limit percent
Flame
Quantitative structure–property relationship (QSPR)
Genetic algorithm-multivariate linear regression (GA-MLR)

ABSTRACT

In this study, a quantitative structure–property relationship (QSPR) is presented to predict the upper flammability limit percent (UFLP) of pure compounds. The obtained model is a five parameters multilinear equation. The parameters of the model are calculated only from chemical structure. The average absolute error and squared correlation coefficient of the obtained model over all 865 pure compounds used to develop the model are 9.7%, and 0.92, respectively.

1. Introduction

Knowledge of flammability limits is needed for safe and economical operation of some chemical and petrochemical processes. This case would be more important when the process is dealing with flammable or combustible materials [1–5].

A flammable gas burns in air only over a limited range of composition. Below a certain concentration of the flammable gas, the mixture lacks sufficient fuel (substance) to burn. This is sometimes called the lower flammability limit percent (LFLP). On the other hands, above the upper flammable limit percent (UFLP) the mixture of substance and air is too rich in fuel (deficient in oxygen) to burn. As a result, the concentrations between these limits constitute the flammable range [1–5].

UFLP flammability limit percent is one of the most important parameters used to evaluate the potential for fire and explosion of industrial materials in the chemical and petrochemical industries [1–5].

There are several methods for estimation of the UFLP of pure compounds. These methods can be classified into several, categories containing empirical correlations, critical flame temperature correlations, structural group contribution models, and neural network models. These methods are reviewed by Vidal et al. [2].

The most important disadvantage of these correlations is their limitations in use. These correlations are obtained based on an especial family of compounds or a small group of compounds. Therefore, the range of applicability of these correlations is very limited.

Quantitative structure–property relationship (QSPR) analysis is now a well-established and useful technique to correlate various simple and complex physicochemical properties of a compound with its molecular structure, through a variety of molecular descriptors (these molecular descriptors are calculated using known mathematical algorithms from molecular structure of every compound). The basic strategy of QSPR analysis is to find optimum quantitative relationships, which can then be used for the prediction of the property from molecular structures. Once a reliable relation has been obtained, it is possible to use it to predict that same property for other structures not yet measured or even not yet prepared. The use of this relation has certain, rather obvious limitation: (i) the family of compounds used to derive the QSPR (the "training set") should be chemically similar and (ii) realistic predictions can only be made for compounds that are chemically related to those from which the QSPR model was derived, i.e., predictions should be of interpolations or short extrapolations [6].

In this study, a quantitative structure–property relationship study is presented to develop a model for predicting the UFLP of a large number of pure compounds. Application of a large num-

* Correspondence address: Department of Chemical Engineering, Faculty of Engineering, University of Tehran, P.O. Box 11365-4563, Tehran, Iran.
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.2009.01.002

ber of compounds can help to extend applicability of the obtained model.

2. Materials and methods

2.1. Data set

Evaluated databases such as DIPPR 801 database [7] are useful tools for developing new property prediction models. DIPPR 801 is recommended by AIChE (American Institute of Chemical Engineers) for physical properties of pure compounds. In this study, 865 pure compounds were extracted and their UFLP were used as main dataset. These compounds and their UFLP values are presented as supplementary materials.

2.2. Determination of molecular descriptors

In this step, the molecular structures of all 865 pure compounds were drawn into Hyperchem software [8] and optimized using the MM+ molecular mechanics force field. Thereafter, using these optimized molecular structures; molecular descriptors were calculated by Dragon software [9]. Dragon software can calculate 1664 molecular descriptors for every molecule. Of course, these molecular descriptors have been calculated for about 2,34,000 pure compounds using Dragon software and are accessible from milano chemometrics and QSAR research group web site.[1] This web site can be searched for desired pure compounds. 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 [9].

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 of UFLP) from all available variables (all molecular descriptors) so that can predict UFLP, 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. This algorithm has been presented by Leardi et al. for the first time [10].

In this study, the GA-MLR technique presented by Leardi et al. [10] with RQK function presented by Todeschini et al. [11] was used to subset variable selection. This methodology has been extensively presented in the previous works of the author and the results are satisfactory [12–25].

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 is 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 865 compounds, 693 compounds are in the training set and 172 compounds are in the test set). The selection was randomly done.

The inputs to our program are the pool of molecular descriptors, the UFLP 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, five, 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:

$$\begin{aligned}\mathrm{UFLP} &= 10.35415(\pm 0.31456) - 1.35486(\pm 0.08144)\mathrm{Jhetv} \\ &\quad - 42.28779(\pm 0.1.44928)\mathrm{PW5} + 18.59571(\pm 0.62369)\mathrm{SIC0} \\ &\quad - + 0.98203(\pm 0.0703)\mathrm{MATS4}m \\ &\quad - - 0.68363(\pm 0.03235)\mathrm{MLOGP} \qquad (1)\end{aligned}$$

$$n_{\mathrm{training}} = 693; \quad n_{\mathrm{test}} = 172; \quad R^2 = 0.9202;$$
$$Q^2_{\mathrm{LOO}} = 0.9184; \quad Q^2_{\mathrm{BOOT}} = 0.9172; \quad Q^2_{\mathrm{EXT}} = 0.9269;$$
$$s = 1.043; \quad a = 0.918; \quad F = 1586.34;$$

RQK function parameters $\Delta K = 0.095$; $\Delta Q = 0.000$; $R^P = 0.009$; $R^N = 0.000$, where UFLP is upper flammability limit percent in vol%.

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

"Jhetv" and "PW5" are of topological descriptors. Topological 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 these two descriptors increase the UFLP decreases.

"SIC0" is of information indices. These molecular descriptors are calculated as information content of molecules, based on the calculation of equivalence classes from the molecular graph. Among them, the indices of neighborhood symmetry take into account also neighbor degree and edge multiplicity. Increase in this descriptor increases the UFLP.

"MATS4m" is of 2D autocorrelations. 2D autocorrelations are spatial autocorrelations calculated on a H-depleted molecular graph weighted by atom physico-chemical properties.

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

ID	Molecular descriptor	Type	Definition
1	Jhetv	Topological descriptors	Balaban-type index from van der Waals weighted distance matrix
2	PW5	Topological descriptors	Path/walk 5 Randic shape index
3	SIC0	Information indices	Structural information content (neighborhood symmetry of 0-order)
4	MATS4m	2D autocorrelations	Moran autocorrelation-lag 4 weighted by atomic masses
5	MLOGP	Molecular properties	Moriguchi octanol–water partition coefficient (log P)

[1] http://michem.disat.unimib.it/mole_db.

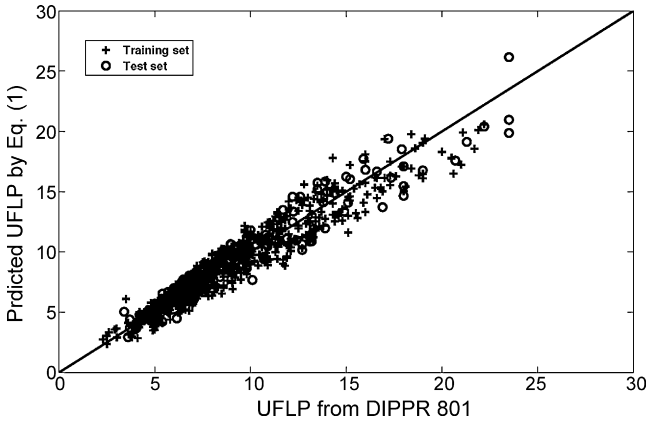

Fig. 1. Comparison between the predicted UFLP by Eq. (1) and DIPPR 801 data.

2D autocorrelations are molecular descriptors which describe how a considered property is distributed along a topological molecular structure. When this descriptor increases the UFLP increases.

"MLOGP" is of molecular properties. These descriptors include a set of heterogeneous molecular descriptors describing physicochemical and biological properties as well as some molecular characteristics obtained by literature models. Increase in this descriptor decreases the UFLP.

$n_{training}$ and n_{test} are the number of compounds of the training set and $n_{training}$ the test set, respectively. For checking validity of the model, more, bootstrap technique, y-scrambling, and external validation techniques were used [11]. 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 [11]. Also the intercept value of the y-scrambling technique has low value ($a = 0.918$) that reveals the validity of the model (The y-scrambling, bootstrapping, and external validation techniques have been extensively presented by Todeschini et al. [11]).

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

The predicted values of UFLP using Eq. (1) in comparison with the DIPPR 801 data are presented in Fig. 1. The values of the predicted UFLP 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.

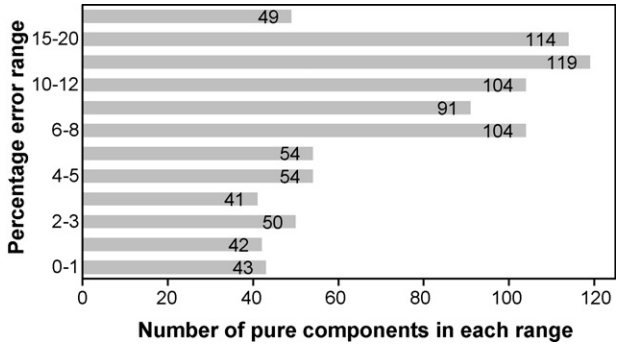

Fig. 2. Percent error of predicted UFLP by Eq. (1) over all of 865 pure compounds used in this study.

4. Conclusion

In this study a simple molecular-based model was presented to predict UFLP flammability limit percent (UFLP) 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 UFLP of pure compounds. The squared correlation coefficient and obtained by this equation over all 865 pure compounds are 0.92. Also the average absolute error of the model over all 865 pure compounds is equal to 9.7%. The percentage error obtained by Eq. (1) is schematically shown in Fig. 2.

Since the model has been obtained using 865 pure compounds which belong to diverse chemical groups, it can be used to predict the UFLP of every regular compound.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jhazmat.2009.01.002.

References

[1] F.P. Lees, Loss Prevention in the Process Industries, vol. 2, 2nd ed., Butterworth Heinemann, Oxford, 1996.
[2] M. Vidal, W.J. Rogers, J.C. Holste, M.S. Mannan, A review of estimations method for flash points and flammability limits, Process Saf. Progress 23 (2004) 47–55.
[3] L. Catoire, V. Naudet, Estimation of temperature-dependent UFLP flammability limit of pure organic compounds in air at atmospheric pressure, Process Saf. Progress 24 (2005) 130–137.
[4] M. Vidal, W. Wong, W.J. Rogers, M.S. Mannan, Evaluation of UFLP flammability limits of fuel–air-diluent mixtures using calculated adiabatic flame temperatures, J. Hazard. Mater. 130 (2006) 21–27.
[5] F. van den Schoor, R.T.E. Hermanns, J.A. van Oijen, F. Verplaetsen, L.P.H. de Goey, Comparison and evaluation of methods for the determination of flammability limits, applied to methane/hydrogen/air mixtures, J. Hazard. Mater. 150 (2008) 573–581.
[6] A.R. Katritzky, D.C. Fara, How chemical structure determines physical, chemical, and technological properties: an overview illustrating the potential of quantitative structure–property relationships for fuels science, Energy Fuels 19 (2005) 922–935.
[7] Project 801, Evaluated Process Design Data, Public Release Documentation, Design Institute for Physical Properties (DIPPR), American Institute of Chemical Engineers (AIChE) 2006.
[8] HyperChem Release 7.5 for Windows, Molecular Modeling System, Hypercube Inc. 2002.
[9] Talete S.R.L., Dragon for windows (Software for Molecular Descriptor Calculations), Version 5.4, 2006 (http://www/talete.mi.it/).
[10] R. Leardi, R. Boggia, M. Terrile, Genetic algorithms as a strategy for feature selection, J. Chemometr. 6 (1992) 267–281.
[11] 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.
[12] F. Gharagheizi, QSPR analysis for intrinsic viscosity of polymer solutions by means of GA-MLR and RBFNN, Comput. Mat. Sci. 40 (2007) 159–167.
[13] F. Gharagheizi, A new accurate neural network quantitative-structure–property relationship for prediction of θ(UFLP critical solution temperature) of polymer solutions, e-polymers, 2007, article number 114.
[14] F. Gharagheizi, M. Mehrpooya, Prediction of standard chemical exergy by a three descriptors QSPR model, Energ. Convers. Manage. 48 (2007) 2453–2460.
[15] F. Gharagheizi, R.F. Alamdari, A molecular-based model for prediction of solubility of C_{60} fullerene in various solvents, Fuller. Nanotub. Car. N. 16 (2008) 40–57.
[16] F. Gharagheizi, QSPR studied for solubility parameter by means of genetic algorithm-based multivariate linear regression and generalized regression neural network, QSAR Comb. Sci. 27 (2008) 165–170.
[17] F. Gharagheizi, A simple equation for prediction of net heat of combustion of pure chemicals, Chemometr. Intell. Lab. Sys. 91 (2008) 177–180.
[18] F. Gharagheizi, A new molecular-based model for prediction of enthalpy of sublimation of pure components, Thermochim. Acta 469 (2008) 8–11.
[19] 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.
[20] F. Gharagheizi, A. Fazeli, Prediction of the Watson characterization factor of hydrocarbon components from molecular properties, QSAR Comb. Sci. 27 (2008) 758–767.

[21] M. Sattari, F. Gharagheizi, Prediction of molecular diffusivity of pure components into air: a QSPR approach, Chemosphere 72 (2008) 1298–1302.
[22] A. Vatani, M. Mehrpooya, F. Gharagheizi, Prediction of standard enthalpy of formation by a QSPR model, Int. J. Mol. Sci. 8 (2007) 407–432.
[23] F. Gharagheizi, M. Mehrpooya, Prediction of some important physical properties of sulfur compounds using QSPR models, Mol. Divers. 22 (2008) 143–155.
[24] F. Gharagheizi, Quantitative structure–property relationship for prediction of lower flammability limit of pure compounds, Energy Fuels 22 (2008) 3037–3039.
[25] F. Gharagheizi, B. Tirandazi, R. Barzin, Estimation of aniline point temperature of pure hydrocarbons: a quantitative structure–property relationship approach, Ind. Eng. Chem. Res., in press, doi:10.1021/ie801212a.