ACADEMIA ROMÂNĂ

Revue Roumaine de Chimie

http://web.icf.ro/rrch/

Rev. Roum. Chim.,
2012, *57*(6), 615-622

SHUFFLING MULTIVARIATE ADAPTIVE REGRESSION SPLINES FOR QSPR CORRELATION OF THE MELTING POINT OF PYRIDINIUM BROMIDES, POTENTIAL IONIC LIQUIDS

Kamal KOR, Kobra ZAREI* and Morteza ATABATI

School of Chemistry, Damghan University, Damghan, Iran

Received December 9, 2011

A quantitative structure-melting point relationship was developed to predict the melting point of some pyridinum bromides. A set of 1497 zero- to three-dimensional descriptors were used for each molecule in the data set. Multivariate adaptive regression spline (MARS) was successfully used as a descriptor selection method and also for mapping model. The root mean square error and coefficient of determination were obtained as 17.36 and 0.8750, respectively. The results were compared with those obtained from other model, which after selection of descriptors by MARS, multiple linear regression (MLR) was applied for modeling. The results showed MARS can be used as a powerful model for prediction of melting point of pyridinum bromides.

INTRODUCTION

Ionic liquids (ILs) have been used as a media for a wide range of reactions and separation processes.[1-3] They can be considered as a new class potentially "green" solvents due to their specific properties. In particular, the nonvolatile nature of many ILs[4] could eliminate problems associated with the use of traditional volatile organic solvents which may pose health, fire, and environmental hazards.[5]

The melting point is a fundamental physical property of compounds, which has been found wide use in chemical identification, as a criterion of purity and for the calculation of other physicochemical properties such as vapor pressure, aqueous solubility and phase equilibrium properties.[6] For ILs, melting points have a special significance because the solubility of ILs in water or organic solvents is strongly correlated with their melting points.[7] However, basic data of melting points exist only for relatively few ILs. Development of melting point-quantitative structure–property relationship (QSPR) models for ILs will provide great aid in molecular design[8] and help to screen candidate lead compounds in search for new room temperature ionic liquids.

For investigation of correlation between the structural descriptors and melting point of ILs some QSPR methods have been applied.[9-11]

The most important two factors influencing the quality of QSPR model are the selected descriptors and the method to build model, respectively. Therefore, variable selection methods are important for producing a useful predictive model. A suitable variable selection ensures the model stability and the consistency of relationship between the descriptors and property.[12] In this work shuffling cross-validation technique was used for descriptor selection. To this purpose the data set was divided into several subsets, and variable selection process was performed for different combinations of these subsets by MARS. Then the most frequent descriptors in the models were selected as the most important variables describing the melting point. The selected descriptors were then used to design the MARS model. Finally this model was applied to predict melting point of ILs.

[*] Corresponding author: zarei@du.ac.ir

The results of this work were compared with those obtained using MARS for the descriptor selection and multiple linear regression for descriptor mapping. The results were also compared with the previous work on the prediction of melting point of these compounds.[11] The results are very good and indicate the power of the descriptor selection and mapping techniques in developing methods with good prediction ability.

RESULTS AND DISCUSSION

Multivariate adaptive regression splines

MARS is a local modeling technique, dividing the data space in several possible overlapping region and fitting truncated spline functions in each region. A truncated spline function consists of a left-sided, Eq. (1), and a right-sided, Eq. (2), segment, separated by a so called knot location.[13]

$$b_q^-(x-t) = [-(x-t)]_+^q = \begin{cases}(t-x)^q & \text{if x < t} \\ 0 & \text{otherwise}\end{cases} \tag{1}$$

$$b_q^+(x-t) = [+(x-t)]_+^q = \begin{cases}(x-t)^q & \text{if x > t} \\ 0 & \text{otherwise}\end{cases} \tag{2}$$

Where t is called the knot location; $b_q^-(x-t)$ and $b_q^+(x-t)$ are spline functions describing the regions to the left and right of the given t; q indicates the power (>0) to which the spline is raised; the subscript "+" indicates a value of zero for negative values of argument. A spline function is also called a basis function (BF). For each of the explanatory variables MARS selects the pair of splines and the knot location, which best describe the response variable. In a next step, the different basis functions are combined in one multidimensional model, which describes the response as a function of the explanatory variables. The result is a complex non-linear model of the form:

$$\hat{y} = a_0 + \sum_{m=1}^{M} a_m B_m(x) \tag{3}$$

where $\hat{y}$ is the predicted value for the response variable; a_0, the coefficient of the constant basis function; M, the number of basis functions and B_m and a_m the m$^{\text{th}}$ base function and its coefficient.[13-15]

A MARS analysis generally consists of three steps. The first step consists of a forward stepwise procedure which selects the best spline functions in order to improve the model and the second step in the MARS methodology consists of a pruning step. A backward elimination procedure is applied in which the basis functions with the lowest contribution to the model are excluded. Eventually, the selection of the optimal model is performed in a third step. The selection is based on an evaluation of the predictive properties of the different models, which often are determined using cross validation or a new independent test set. The MARS models were built using ARESLab toolbox.[16]

Further details on MARS modeling are given elsewhere.[14]

Selection of the best descriptors by Shuffling cross validation MARS

In this technique, the data set was divided into several subsets, and variable selection procedure and model developing was performed for all combinations of the subsets. Then the most frequent descriptors appeared in the developed models would be selected as most important variables in describing the variation in the property. The use of shuffling cross validation technique guarantees that the developed model is robust and reliable and it is not obtained by chance.

In this work, the molecules were divided into twelve groups, six groups of them consisted of ten molecules and six consisted of eleven molecules. Each group was selected in such a way that it consisted of all range of melting point amounts from low to high. In the descriptor selection procedure, ten groups were applied as calibration set and the two remaining subsets were used as validation set for evaluating the selected descriptors. 66 MARS models were made with various calibration and validation sets. These calibration sets contain different molecules; therefore various descriptors are expected to be selected by MARS method.

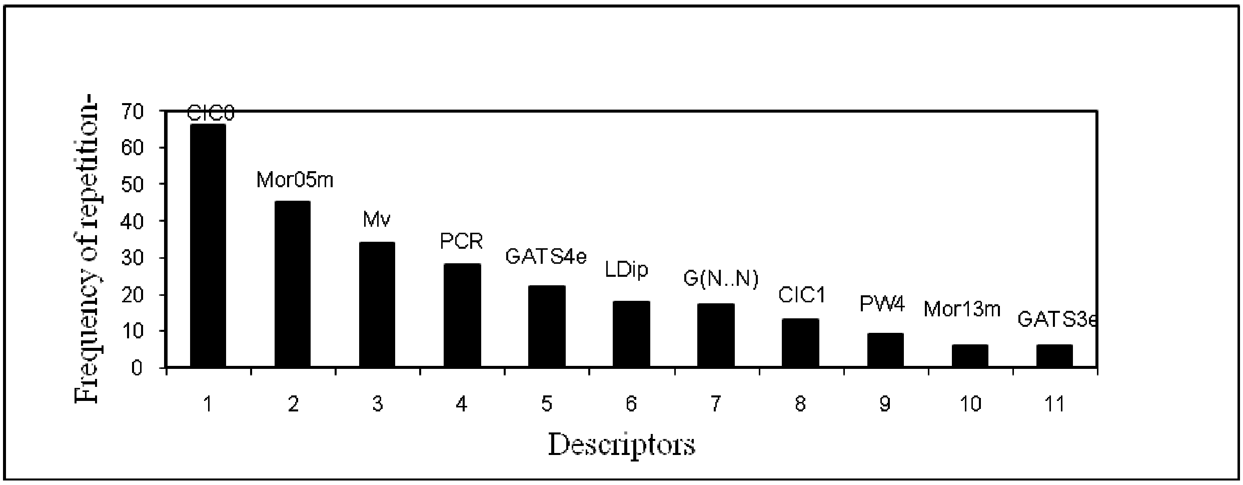

Fig. 1 – The selected descriptors and the frequency of each one in the shuffling-MARS models.

The selected descriptors and frequency of each descriptor in shuffling-MARS models are shown in Fig. 1. As can be seen in Fig. 1; eight descriptors have the highest frequency between the others. Among these eight descriptors two descriptors have high correlation with the others; therefore six descriptors were selected by MARS model. The MARS model was designed with six selected descriptors, complementary information content (CIC0, neighborhood symmetry of 0-order), 3D-MoRSE - signal 05 / weighted by atomic masses (Mor05m), Geary autocorrelation - lag 4 / weighted by atomic Sanderson electronegativities (GATS4e), mean atomic van der Waals volume (Mv, scaled on Carbon atom), local dipole index (LDip) and sum of geometrical distances between N..N (G(N..N)). These six descriptors are among topological, 3D-MoRSE, 2D autocorrelations, constitutional, charge and geometrical descriptors. The constructed MARS model with these six descriptors has RMSE (root mean square error) and R^2, 20.52 and 0.8218, respectively. Six data points in this model lie outside the range of 2s (95%confidence limit) from the predicted value (structures 33, 48, 57, 93, 96 and 102 from data set of Katritzly work [11]), therefore these molecules were not used in the future analysis. Finally after elimination of these six data point, the best obtained model has RMSE=17.36 and R^2= 0.8750. The melting temperatures predicted by this model are included in Table 1, and the corresponding correlation chart and residual are given in Figs. 2 and 3, respectively.

Table 1

Experimental and calculated melting points for Pyridinum bromides

Cation		mp (°C)		Cation		mp (ºC)	
N-substituent	Other substituents	Exp.	Calc.	N-substituent	Other substituents	Exp.	Calc.
decyl	3-pentyl	30.0	40.7	2-hydroxyethyl		110.0	129.1
11-propionyloxyundecyl		35.0	32.5	allyl	4-hydroxymethyl	111.0	122.3
benzyl	4-dibenzyl; 3-ethoxycarbonyl	35.0	17.3	5-hexynyl		114.0	98.9
octyl	4-propyl	37.0	50.1	2-cyanoethyl	3-methyl	116.0	134.1
tetradecyl	4-hexyl	37.0	34.9	pyridinyl		118.0	96.2
tetradecyl	3-pentyl	39.0	32.8	isopropyl	4-hydroxymethyl	119.0	101.0
decyl	4-ethoxycarbonyl	40.5	34.3	3-chloropropyl		120.0	132.1
dodecyl	4-propyl	41.0	45.1	ethyl		120.5	107.4
undecyl		41.9	48.0	ethyl	4-cyan	121.0	127.4
octyl	4-ethyl	42.5	52.4	ethyl	4-methyl	121.0	124.0
dodecyl	4-ethyl	43.5	47.1	isopropyl	2-hydroxymethyl	122.0	96.5
decyl		44.5	49.3	2-hydroxyethyl	3-hydroxy	122.5	129.8
dodecyl		45.0	46.6	2-hydroxyethyl	3,4-dimethyl	126.5	104.6
hexyl	2-(2-methyloctyl)	47.0	43.5	3,3-dimethylallyl	4-methyl	127.5	107.2
ethoxycarbonylmethyl	5-butyl; 2-methyl	51.0	68.5	1-methyl-2-oxopropyl	2-methyl	133.0	115.2
2,5-dimethoxyphenethyl		53.8	68.6	2-cyano-ethyl	3,4-dimethyl	133.0	131.3
tridecyl		54.5	48.3	ethoxycarbonylmethyl		137.0	140.1
4-fluoro-benzyl		57.5	81.3	3-bromopropyl	4-methyl	139.5	139.4
tetradecyl		59.0	47.2	(Z)-3-methylpent-2-en-4-inyl		139.5	110.5

Table 1 (continued)

butyl	4-ethoxycarbonyl	60.0	66.4	methyl	3-methoxycarbonyl	141.0	161.6
2-methylpropyl		64.0	66.1	2-cyanoethyl	3-amino	142.0	170.0
methyl	3-(3-hydroxypropyl)	64.0	78.3	1-methyl-2-oxopropyl		144.0	145.3
benzyl	3-methyl	64.5	73.7	cyanomethyl	3,5-dimethyl	144.0	131.9
butyl	2-benzylsulfanyl	65.0	57.8	methyl	4-methyl-3-hydroxy	145.8	128.0
2-cyclohexyl-2-oxo-ethyl		69.0	51.4	2-cyanoethyl	4-methyl	147.0	167.2
methylpropyl		69.5	66.3	2-cyanoethyl		149.0	145.8
2-pyridinyl		70.4	85.7	methyl		150.0	158.1
2-(ethoxycarbonyl)ethyl		71.0	103.2	methyl	3-hydroxy	153.5	156.0
1-(ethoxycarbonyl)propyl		73.0	78.4	vinyl		154.5	143.2
propyl		75.5	71.5	(E)-3-hydroxyprop-1-en-1-yl		156.5	148.5
ethyl	3-diethylcarbamoyl	77.0	94.5	2-carboxyallyl		156.5	160.4
2-phenoxyethyl		78.0	83.8	2-cyanoethyl	3,5-dimethyl	160.5	123.7
methyl	4-(3-hydroxypropyl)	79.0	98.4	3-carboxypropyl		163.0	147.1
ethyl	2,6-dimethyl	80.0	113.1	methyl	4-methoxycarbonyl	164.0	152.6
methyl	3-pyridinyl	80.0	104.5	ethyl	4-cyano	165.5	145.9
isopropyloxycarbonylmethyl		8.0	114.3	cyanomethyl		167.0	155.4
morpholinomethyl	4-methyl	82.5	77.9	vinyl	4-methyl	169.0	170.7
methyl	4-benzyl	85.5	102.9	isopropyl	4-methoxy	169.5	136.9
2-fluoroethyl	3-ethoxycarbonyl	87.0	108.3	methyl	4-methyl	173.0	147.8
phenethyl	4-methyl	89.9	81.2	methoxycarbonylmethyl		174.5	152.7
butyl	3-carboxy	91.0	110.8	propyl	3-carbamoyl	175.0	142.9
allyl	3-diethylcarbamoyl	92.0	89.0	ethyl	4-dimethylamino	176.0	146.1
bis(ethoxycarbonyl)methyl		93.5	87.7	prop-2-ynyl	4-methyl	178.0	149.2
4-acetoxybutyl	3-hydroxy	95.0	97.7	cyanomethyl	4-methyl	179.0	187.7
benzyloxy		95.0	81.7	2-fluoroethyl		180.0	159.7
allyl		95.5	118.9	methyl	4-acetyl	183.5	161.8
2-hydroxyethyl	3-methyl	97.0	102.8	allyl	4-(hydroxyiminomethyl)	185.0	169.4
ethyl	2-methyl	97.0	108.6	hydrazinocarbonylmethyl		185.5	180.8
isopropyl		97.0	111.5	2-oxopropyl		187.0	181.6
butyl		97.5	63.5	ethyl	4-carbamoyl	188.0	185.7
ethyl	4-(4-pyridyl)	98.0	137.4	(E)-2-carboxy-1-ethyl		188.5	191.6
allyl	3-hydroxy	98.0	138.9	2-propionamido		189.5	176.7
benzyl		99.0	78.0	(E)-2-carboxy-1-ethyl	3-methyl	190.5	184.9
methyl	4-(2-ethoxycarbonylethyl	99.5	92.1	allyl	2-(hydroxyiminomethyl)	192.5	175.6
allyl	3-formyl	100.0	132.8	2-oxopropyl	2-methyl	196.0	175.7
acetonyl	2,6-dimethyl	104.0	149.0	2-hydroxyethyl	2-(hydroxyiminomethyl)	198.5	184.0
ethyl	3-hydroxy	106.0	116.7	cyanomethyl	2,4-dimethyl	199.0	201.4
ethoxy	4-methoxy	108.5	110.3	carboxymethyl		199.0	204.1
propyloxycarbonylmethyl		110.0	96.3	2-carbamoylethyl		199.0	201.8
allyl	2-hydroxymethyl	110.0	120.5	carbamoylethyl		200.0	226.0

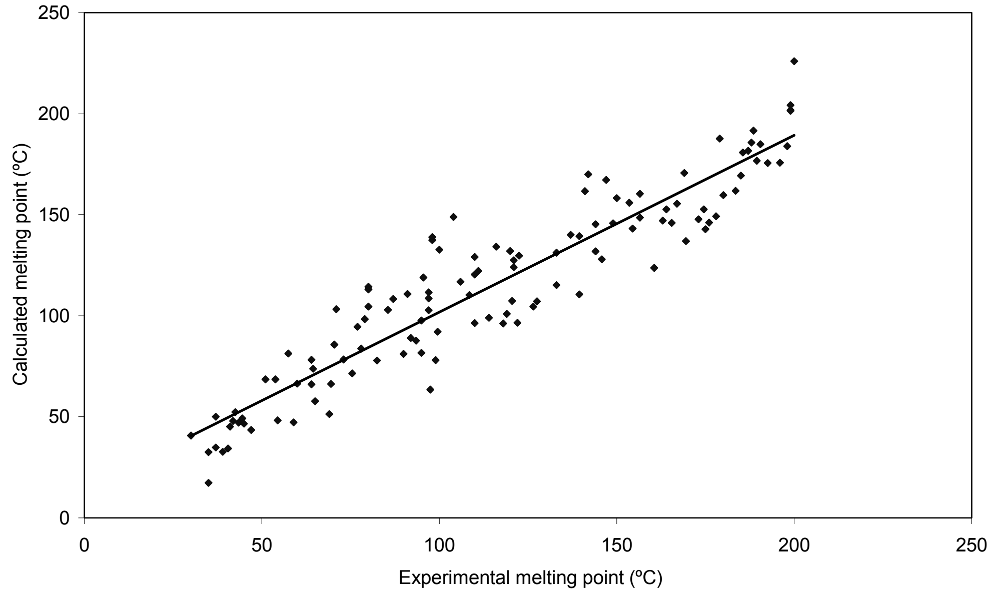

Fig. 2 – Plot of the calculated melting point against the experimental melting point.

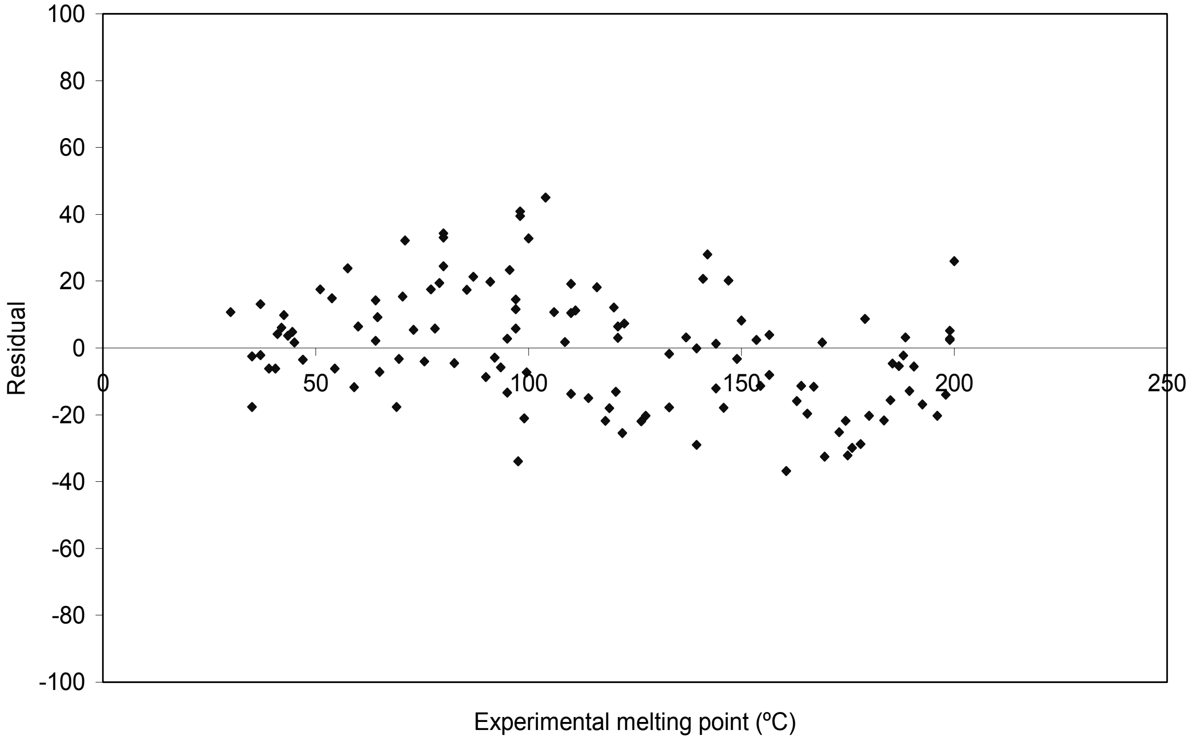

Fig. 3 – Plot of the residuals versus experimental values of melting point.

This model contains a constant B_1 (=1) and nine basis function. These nine BFs represented by B_2 to B_{10} as well as their coefficients a_i are shown in Table 2. As an example of a basis function in the model, consider B_2:

$$3.988 - CIC0 = \begin{cases} 3.988\text{-CIC0} & \text{if CIC0} < 3.988 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

this means that, when CIC0<3.988, the second term of equation of $LogBCF = \sum_{i=1}^{38} a_i B_i$ is 56.942(3.988-CIC0), otherwise it is 0. Fig. 4 also shows variations of melting point versus CIC0. These variations are for CIC0<3.988 and this is in agreement with MARS equation.

Table 2

List of basis function B_i of the MARS model and their coefficients, a_i

Bi	Definition	a_i
B1	1	91.576
B2	3.988 - CIC0	56.942
B3	LDip -0.194	474.5
B4	Mv -0.62	-770.49
B5	0.812 - GATS4e	97.281
B6	4.35 - G(N..N)	-3.4328
B7	Mor05m -1.984	36.158
B8	-1.984 - Mor05m	-17.271
B9	0.6 - Mv	-3728.8
B10	0.59 - Mv	4047.3

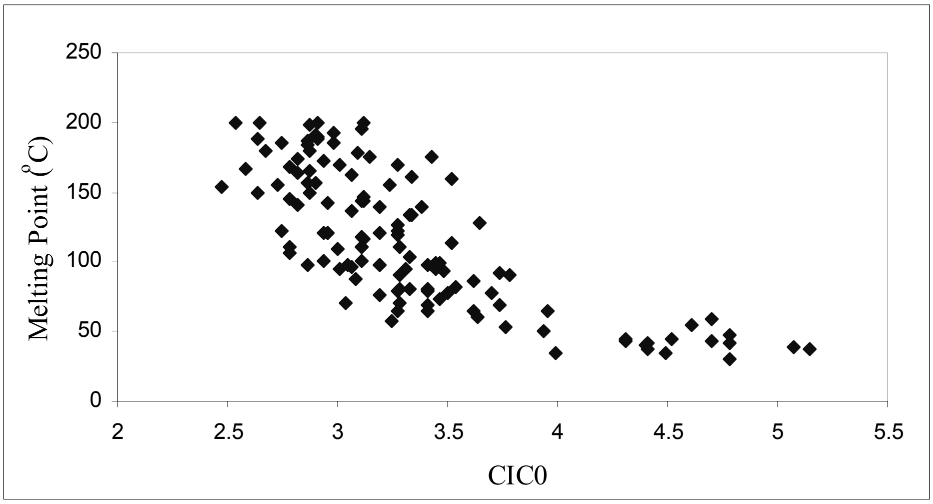

Fig. 4 – Variations of melting point versus descriptor of CIC0.

CIC0 is the first descriptor, complementary information content (neighborhood symmetry of 0-order). CIC0 is from topological descriptors. This descriptor describes the connectivity and branching in a molecule and can be related to molecular shape and symmetry. The decreasing in melting point with increasing in CIC0 reflects the fact that cations with lower symmetry have weaker coordination ability that leads to lower melting temperatures. The second descriptor is local dipole index (LDip). This descriptor is a molecular descriptor calculated as the average of the charge differences over all *i-j* bonded atom pairs. The MARS equation shows the melting point increases with increasing in LDip. This result is an accepted fact which larger charge density (charge/volume ratio) led to stronger bonding and higher melting point. The third descriptor is mean atomic van der Waals volume (Mv) which it is calculated by dividing the sum of the van der Waals volumes by the number of atoms. The fourth descriptor is Geary autocorrelation - lag 4 / weighted by atomic Sanderson electronegativities (GATS4e). Geary coefficient *(c(d))* is general index of spatial autocorrelation that, if applied to a molecular graph, can be defined as: [17]

$$c(d) = \frac{\frac{1}{2\Delta}\cdot\sum_{i=1}^{A}\sum_{j=1}^{A}\delta_{ij}\cdot\left(w_i - w_j\right)^2}{\frac{1}{(A-1)}\cdot\sum_{i=1}^{A}\left(w_i - \bar{w}\right)^2} \qquad (5)$$

where w_i is any atomic property such as electronegativities, $\overline{W}$, is its average value on the molecule, *A* is the atom number, d is the considered topological distance (i.e. the lag in the autocorrelation terms), δ_{ij}, is a Kronecker delta (δ_{ij}, = 1 if $d_{ij} = d$, zero otherwise). Δ is the sum of the Kronecker deltas. Strong autocorrelation produces low values of this index. Therefore it seems with increasing at difference of electronegativity, this coefficient is decreased and according to MARS model, melting point is increased. G(N..N) (sum of geometrical distances between N..N) from geometrical descriptors is the fifth descriptor and the last descriptor is Mor05m from MoRSE class descriptors. MoRSE (molecular representation of structures based on electron diffraction)- signal 05 / weighted by atomic masses encodes structural features such as mass and amount of branching.[18] The negative coefficient in MARS equation shows that with increasing in mass and branching, melting point is decreased.

Validation

To validate the developed MARS model for the prediction of melting points of ionic liquids, data set was divided into three subsets A, B, and C. The MARS models were obtained for the subsets A+B, A+C, and B+C with six selected descriptors. The resulting new MARS models were used in turn to predict the melting points for subsets C, B, and A, respectively. Square of correlation coefficients and root mean square error of calibration and prediction are presented in Table 3. The correlation chart of the validation showing the summary of all three predictions is given in Fig. 5.

Table 3

Validation of MARS model with six descriptors

Cal= A+B		Pre=C		Cal=A+C		Pre=B		Cal=B+C		Pre=A	
RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
16.34	0.8898	24.82	0.7422	17.79	0.8708	22.56	0.7818	17.96	0.8634	20.88	0.8263

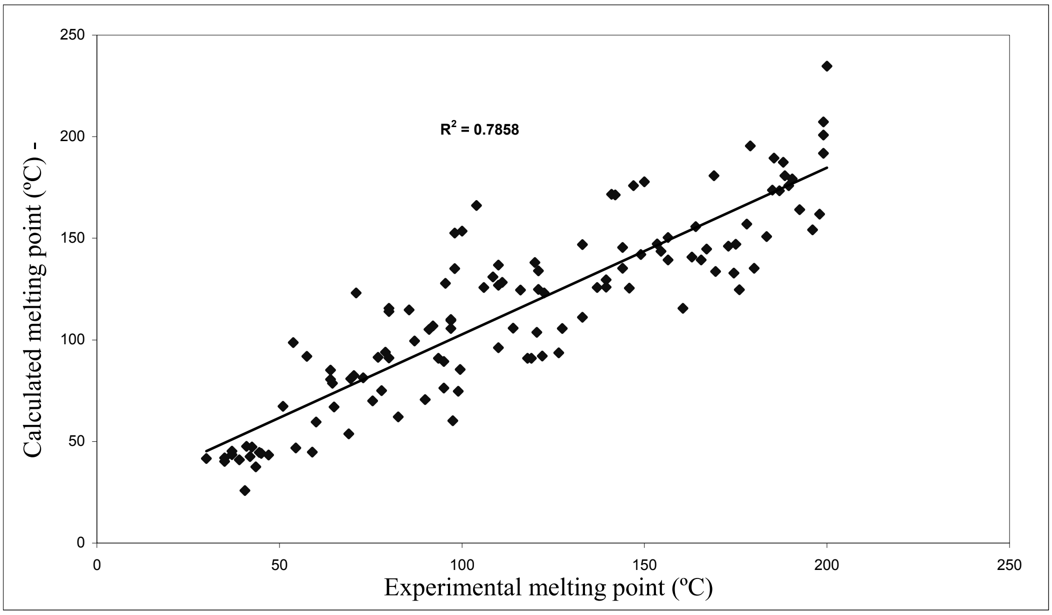

Fig. 5 – Cross validation for MARS model.

Comparison with other models

The other model was also constructed for the sake of comparison. In this model, the selected descriptors by MARS were used for building of multiple linear regression (MLR) model. In other words, MARS was used for descriptors selection and then MLR was applied for model development. MLR model was also made with the same six selected descriptors in the previous step. Initially 126 molecules were applied to obtain the MLR model. Thus the equation MLR model was obtained as below with RMSE=32.96 and R^2=0.5401:

$$M.P. = -54.553 + 7.89\text{CIC0} + 17.821\text{Mor05m} - 4.008\text{GATS4e} + 285.039\text{Mv} \\ + 13.803\text{LDip} - 2.607\text{G(N..N)} \qquad (6)$$

It can be seen that MARS model is superior over the MLR model which applied the same descriptors and shows improvements for R^2 and RMSE.

In comparison with previous reported work on these 126 ionic liquid[11] (the same data set in this work) with R^2= 0.7883 and s= 23.0 K, our model has better statistical results with RMSE=20.52 and R^2=0.8218 for 126 compounds and RMSE=17.36 and R^2= 0.8750 for 120 compounds.

EXPERIMENTAL

Data set

The known experimental melting point values of the 126 pyridinium bromides were taken from the literature[11] and shown in Table 1.

Calculation of descriptors

The 3-D structures of these compounds were optimized using Hyper Chem software (version 7.0) with semi empirical

AM1 optimization method. After optimization a total of, 1497 0-, 1-, 2-, and 3-D descriptors were generated using Dragon software (version 3.0).

CONCLUSION

The main aim of the present work was the development of a QSPR method using multivariate adaptive regression spline methodology for both descriptor selection and for feature mapping of melting points of ILs. It is shown in this work that MARS as feature selection method generates very predictive descriptors and also it is a powerful mapping tool. The most significant descriptors appeared in the model are: complementary information content reflecting the coordination ability of a cation and atomic van der Waals volume (Mv) and local dipole index (LDip) which show larger charge density (charge/volume ratio) led to stronger bonding and higher melting point. This seems the descriptors used in this model are in consistence with the suggested experimental factors to affect the melting point of ionic compounds.

Because of the dominating role of the cation in determining the properties of ionic liquids, the developed models can help to suggest compounds in search for new potential ILs.

Acknowledgements: The authors acknowledge the partial support of this work from the Research Council of Damghan University.

REFERENCES

1. J. Dupont, R. F. Souza and P. A. Z. Suarez, *J. Chem. Rev.,* **2002**, *102*, 3667-3692.
2. P. Wasserscheid and W. Keim, *Angew. Chem.Int. Ed.,* **2000**, *39*, 3772-3789.
3. T. Welton, *Chem. Rev.,* **1999**, *99*, 2071-2083.
4. P. T. Anastas, *Critical ReV. Anal. Chem.,* **1999**, *29*, 267-268.
5. M. Koel, “Ionic Liquids in Chemical Analysis”, Talor & Francis: USA, 2009.
6. A.R. Katritzky, R. Jain, A. Lomaka, R. Petrukhin and M. Karelson, *Crystal Growth Design,* **2001**, *1*, 261-265.
7. R. Abramowitz and S. H. Yalkowsky, *Pharm. Res.,* **1990**, *7*, 942-947.
8. S. Trohalaki, R. Pachter, G. W. Drake and T. Hawkins, *Energy & Fuels,* **2005**, *19*, 279-284.
9. D. M. Eike, J. F. Brennecke and E. J. Maginn, *Green Chem.,* **2003**, *5* 323-328.
10. A. R. Katritzky, R. Jain, A. Lomaka, R. Petrukhin , M. Karelson, A. E. Visser and R. D. Rogers, *J. Chem. Inf. Comput. Sci.,* **2002**, *42*, 225-231.
11. A. R. Katritzky, A. Lomaka, R. Petrukhin, R. Jain, Karelson; M., A. E. Visser and R. D. Rogers, *J. Chem. Inf. Comput. Sci.,* **2002**, *42* 71-74.
12. T. Ghafourian and M. T. D. Cronin, *Sar Qsar Environ. Res.,* **2005**, *16*, 171-190.
13. J. H. Friedman, *Annals of Statistics,* **1991**, *19*, 1-67.
14. E. Deconinck, Q. S. Xu, R. Put, D. Coomans, D. L. Massart and Y. V. Heyden, *J. Pharm. Biomed. Anal.,* **2005**, *39*, 1021-1030.
15. S. Sekulic and B. R. Kowalski, *Journal of Chemometrics,* **1992**, *6*, 199-216.
16. Jekabsons G (2009) ARESLab: Adaptive Regression Splines toolbox for Matlab, http://www.cs.rtu.lv/jekabsons/
17. R. Todeschini and V. Consonni, “Handbook of Molecular Descriptors”, Wiley-VCH: Weinheim, Germany, 2000.
18. H. Modarresi, J. C. Dearden and H. Modarress, *J. Chem. Inf. Model.,* **2006**, *46*, 930-936.