ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Modeling and Prediction of Flash Point of Unsaturated Hydrocarbons Using Hybrid Genetic Algorithm/Multiple Linear Regression Approach.

Mabrouka DIDI, Hamza HADDAG, Youcef DRIOUCHE, and Djelloul MESSADI*.

Environmental and Food Safety Laboratory, Badji Mokhtar Universit , Annaba 23000,Algeria

ABSTRACT

A quantitative structure property relationship (QSPR) study is developed using Genetic Algorithm (GA) / Multiple Linear Regression (MLR) for modeling the flash points of 173 unsaturated hydrocarbons, using theoretical molecular descriptors derived from DRAGON software. The studied dataset was randomly separated into two independent subsets: a training set of 139 compounds to build the model and a test set of the removed 34 compounds to validate its predictive ability. The selection of a minimum set of meaningful descriptors was carried out using Genetic Algorithm in the MOBYDIGS Todeschini software. An MLR model of 4 descriptors with a high predictive power was developed for the prediction of the flash points of unsaturated hydrocarbons.The predictive ability of the obtained model was verified using a set of criteria according to Golbraikh and tropsha and its applicability domain was studied using Willians plot.

Keywords: Flash point; Unsaturated hydrocarbons; Multiple linear regression; Quantitative structure-property relationship; Model prediction.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

The flash point (FP) is defined as the lowest temperature, corrected to 101.3 k Pa, at which an application of an ignition source causes the vapors of the specimen to ignite under specific conditions of a test [1-4].

This parameter gives the knowledge necessary for understanding the fundamental physical and chemical processes of combustion. Moreover, it is of importance in practice for safety conditions in the storing, the processing and the handling of a given compound. And it is one of the major flammability characteristics used to assess the fire and explosion hazards of organic compounds [5].

The flash point of most compounds can be measured by two currently accepted experimental methods, which are the closed cup test and the open cup test [6]. However, for many other compounds, the experimental flash point values are scarce and too expensive to obtain. Moreover, it is even more difficult to make the experimental determination of the flash point of toxic, volatile, explosive and radioactive compounds. Hence, the development of estimation methods which are desirably convenient for predicting the flash point is required.

There are many methods for prediction of FP in the literature. Vidal et al. have presented a review of the most important methods for the prediction of the flash point [7]. Mainly, prediction methods for this property can be categorized as the group contribution method (GCM), the principal component analysis (PCA) and the quantitative structure-property relationship (QSPR).

A simple correlation for predicting the flash point of a large data set consisting various types of cyclic and acyclic hydrocarbons including the studied compounds where the proposed method was based on the number of carbons and hydrogen atoms and some specific molecular moieties, which can easily be used for any type of hydrocarbons [8].

Another method was introduced for the prediction of the flash point of different classes of unsaturated hydrocarbons showing that the number of carbons and hydrogen atoms can be used as a core function that may be revised by a correcting function. Correcting function contains two correcting terms that can be determined on the basis of molecular structure that can be determined on the basis of the molecular structure of unsaturated hydrocarbons [9].

The aim of this work is to build a new QSPR model that can be used for predicting flash points of 173 unsaturated hydrocarbons [9] from their molecular structure. In this work, after obtaining the most statistically significant descriptors by means of genetic algorithm (GA) based on variable selection approach, the multiple linear regression behavior of these molecular descriptors for predicting flash point of these compounds was studied.

MATERIALS AND METHODS

The data set

The experimental flash point dataset was taken from literature [9].

The set of the studied compounds is formed of different classes of unsaturated hydrocarbons including alkenes, alkynes and aromatics. Flash point values are in a range from 137 to 451 K. The dataset was randomly divided into two groups, a training set of 139 compound and a test set of 34 compound.

Descriptor generation

The chemical structure of each compound was sketched on a PC using the Hyperchem program [10] and preoptimized using MM+ molecular mechanics method (Polack-Ribiere algorithm). The final geometries of the minimum energy conformation were obtained by the semi-empirical PM3 method at a restricted Hartree-Fock level with no configuration interaction, applying a gradient norm limit of 0.01 k cal $.A^{-1}.mol^{-1}$ as a stopping criterion.

ISSN: 0975-8585

The output files exported from Hyperchem were transferred into Dragon software [11], to calculate a large number of molecular descriptors on the basis of the geometrical and electronic structure of the molecules. Constant values and descriptors found to be correlated pairwise were excluded in a pre-reduction step (when there was more than 95 % pairwise correlation, one variable was deleted), and the genetic algorithm was applied for variables selection to a final set of 269 descriptor.

Data splitting

In order to check the predictive capability of the proposed model, before model generation the data set was randomly split into a training set of 139 compounds from which the model is built and an external test set of 34 compounds on which to evaluate its prediction power as it is shown on Table 1.

Table 1: Names, structures and FP values of the studied set

N°	Names	FP(K)	N°	Names	FP(K)
1	Benzene	262	88	2-Heptyne	275
2*	Toluene	280	89	3-Heptyne	257
3*	Ethylbenzene	288	90*	3-Methyl-1-hexyne	268
4*	P-Xylene	300	91	1,7-Octadiyne	296
5	O-Xylene	303	92	2,6-Octadiene	307
6	Propylbenzene	303	93	1-Octyne	289
7	Cumene	304	94	2-Octyne	301
8	m-Ethyltoluene	311	95*	4-Octyne	291
9	1,2,3-Trimethylbenzene	324	96*	1,8-Nonadiyne	314
10*	1,2,4-Trimethylbenzene	321	97	1-Nonyne	306
11	1,3,5-Trimethylbenzene	317	98	1-Decyne	323
12	o-Ethyltoluene	312	99	1-Undecyne	338
13*	p-Ethyltoluene	309	100	4-Undecyne	341
14	Naphthalene	360	101	1-Dodecyne	352
15	Butylbenzene	331	102	1-Tridecyne	366
16	1,2,4,5-Tetramethylbenzene	346	103	Cyclobutene	202
17	2-Ethyl-p-xylene	329	104	Cyclopentene	244
18*	3-Ethyl-o-xylene	338	105	Cyclohexene	256
19*	4-Ethyl-m-xylene	330	106	4-Methylcyclopentene	243
20*	tert-Butylbenzene	307	107	Cycloheptene	267
21	P-Cymene	320	108	4-Methylcyclohexene	272
22	o-DiEthylbenzene	322	109*	3-Methylcyclohexene	270
23	m-DiEthylbenzene	324	110	4-Ethylcyclohexene	286
24	p-DiEthylbenzene	328	111	Ethylene	137
25	4-Ethyl-1,2-dimethylbenzene	331	112	Propene	165
26	1-Methylnaphtalene	355	113	Propadiene	177
27	n-Pentylbenzene	339	114	1,2-Butadiene	197
28	IsoPentylbenzene	335	115	1,3-Butadiene	197
29	Pentamethylbenzene	364	116*	Butene	194
30	p-tert-Butyltoluene	321	117	Cis-2-Butene	200
31	2-Phenyl-2methylbutane	338	118	Isobutylene	197
32	1-Ethylnaphtalene	380	119	1,2-Pentadiene	233
33	2-Ethylnaphtalene	377	120	2,3-Pentadiene	235
34	1,3-Dimethylnaphtalene	382	121	Cis-1,3-Pentadiene	232
35*	1,2-Dimethylnaphtalene	374	122*	2-Methylbutadiene	225
36	Hexylbenzene	356	123*	Pentene	229
37	Hexamethylbenzene	377	124	2-Pentene	253
38	3,5-Dimethyl-tert-butylbenzene	357	125	Cis-2-Pentene	227
39	1,2,4-Trimethylbenzene	349	126	trans-2-Pentene	225

ISSN: 0975-8585

N°	Names	FP	N°	Names	FP
40*	1,3,5-Triethylbenzene	354	127	Isopentene	211
41	1,4-Diisopropylbenzene	354	128	1,4,-Hexadiene	248
42	m-Diisopropylbenzene	350	129*	2,4-Hexadiene	264
43	n-Heptylbenzene	368	130	1,5-Hexadiene	246
44	1,2,3,4-Tetraethylbenzene	367	131	2,3-Dimethyl-1,3-butadiene	251
45	2-Phenyloctane	373	132	3-Methyl-1,4-pentadiene	239
46	n-Octylbenzene	380	133	2-Methyl-2,3-pentadiene	255
47*	n-Nonylbenzene	390	134	1-Hexene	253
48	1,3,5-Triisopropylbenzene	359	135	Cis-2-Hexene	252
49	Decylbenzene	380	136*	Cis-3-Hexene	261
50	Pentaethylbenzene	386	137	Trans-3-Hexene	261
51*	n-Undecylbenzene	409	138*	Isohexene	241
52	Dodecylbenzene	418	139	2,3-Dimethyl-1-butene	255
53	1,2,4,5-tetraisopropylbenzene	397	140	2,3-Dimethyl-2-butene	256
54	1,3,5-Tri-tert-butylbenzene	372	141	3,3-Dimethyl-1-butene	244
55	Tridecylbenzene	385	142	2-Methyl-1-pentene	241
56*	1-Methylanthracene	430	143	2-Methyl-2-pentene	246
57	2-Methylanthracene	431	144	4-Methyl-2-pentene	241
58	9-Methylanthracene	431	145	3-Methyl-1-pentene	244
59	1-methylphenanthrene	431	146	Trans-3-Methyl-2-pentene	266
60	7-isopropyl-1-methylphenanthrene	451	147*	2-Ethyl-1-butene	243
61	Phenylacetylene	303	148	1,6-Heptadiene	263
62	Styrene	304	149	1-Heptene	264
63*	2-Vinyltoluene	320	150	Cis-2-Heptene	265
64	3-Vinyltoluene	324	151	Trans-2-Heptene	267
65	3-Phenyl-1-propene	310	152	Trans-3-Heptene	266
66	beta-Methylstyrene	333	153	2-Methyl-1-hexene	267
67*	Cis-1-Propenylbenzene	325	154*	4-Methyl-1-hexene	258
68	Isopropenylbenzene	313	155*	2-Ethyl-1-pentene	263
69*	trans-1-phenyl-1-propene	331	156*	2,4-Dimethyl-2-pentene	264
70	m-Divinylbenzene	338	157	2,3,3-Trimethyl-1butene	256
71	p-Divinylbenzene	337	158*	Cis-5-Methyl-2-Hexene	268
72	1-Butynylbenzene	341	159	Trans-5-Methyl-2-Hexene	268
73	3-Ethylstyrene	333	160	Trans-3-Octene	282
74	4-Ethylstyrene	335	161	Trans-4-Octene	281
75	2,4-dimethyl-1-vinylbenzene	333	162	Cis-4-Octene	294
76	Acetylene	155	163*	1,8-Nonadiene	299
77	Propyne	186	164	2-Ethyl-1-hexene	279
78	1-Pentyne	230	165	1-Nonene	298
79	2-Pentyne	253	166*	1-Undecene	336
80	3-Methyl-1-butyne	221	167	Dodecene	351
81	1-Hexyne	252	168	2-Methyl-1-undecene	345
82*	2-Hexyne	263	169	1-Tridecene	352
83	3-Hexyne	259	170	1-Tetradecene	383
84	3,3-Dimethyl-1-butyne	239	171	1-Pentadecene	386
85	4-Methyl-1-Pentyne	249	172	1-Hexadecene	402
86	1,6-Heptadiyne	282	173	1-Heptadecene	408
87	1-Heptyne	263			

* compounds of the test set .

ISSN: 0975-8585

Model development and validation

Once the molecular descriptors are generated, multiple linear regression (MLR) analysis and variable selection were performed by the software Mobydigs [12] using the Ordinary Least Square (OLS) regression method and Genetic Algorithm _Variable Subset Selection (GA-VSS) [13].

The outcome of the application of the genetic algorithms is a population of 100 regression models, ordered according to their decreasing internal predictive performance, verified by Q2.First of all, models with 1-2 variables were developed by the all-subset-method procedure in order to explore all the low dimension combinations. The number of descriptors was subsequently increased one by one, and the new models were formed. The best models are selected at each rank, and the final model must be chosen from among them. This has to be sufficiently correlated and at the same time, protect against any over parametrisation, which would lead to a loss of predictive power for molecular outside training set. From a statistical view point the ratio of the number of samples (n) to the number of descriptors (m) should not be too low. Usually, it is recommended that $n/m \geq 5$ [14]. The GA was stopped when increasing the model size did not increase the Q2 value to any significant degree.

Particular attention was paid to the collinearity of the selected molecular descriptors by applying the QUIK (Q Under Influence of K) rule [15] a necessary condition for the model validity. Acceptable models are only with a global correlation of [X+Y] block (Kxy) greater than the global correlation of the X block (Kxx) variable, X being the molecular descriptors and Y the response variable. Therefore, when there were models of similar performance, those with higher Δ K (Kxy-Kxx) were selected and further verified.

The goodness of fit of the calculated models were assessed by the means of the multiple coefficient R^2, and the standard deviation error in calculation (SDEC).

$$\mathrm{SDEC} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (1)$$

Cross validation techniques allow the assessment of internal predictivity (bootstrap) in addition to the robustness of model (Q^2_{LOO} cross validation).

Cross validation methods consist in leaving out a given number of compounds from the training set and rebuilding the model, which is then used to predict the compounds left out .This procedure, is repeated for all compounds of the training set, obtaining a prediction for everyone. If each compound is taken away once each time the cross validation procedure is called leave-one-out technique (LOO technique). An LOO correlation coefficient, generally indicated with Q^2, is computed by evaluating the accuracy of these “test” compounds prediction.

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{i/i}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} = 1 - \frac{\mathrm{PRESS}}{\mathrm{TSS}} \qquad (2)$$

The “hat” of the variable y, as is the usual statistical notation, indicates that it is a predicted value of the studied property, and the sub index “i/i” indicates that the predicted values come from the model built without the predicted compound.

The predictive residual of squares (PRESS) measures the dispersion of the predicted values. It is used to define Q^2 and the standard error in prediction (SDEP).

$$SDEP=\sqrt{PRESS/n} \quad (3)$$

A value of $Q^2 \geq 0.5$ is generally considered satisfactory, and a value greater than 0.9 is excellent [16,17]. However, studies have indicated that while Q^2 is a necessary condition for high predictive power of a model, is not sufficient.

In bootstrap validation technique K n-dimensional groups are generated by a randomly repeated selection of n-objects from the original data set. The model obtained on the first selected objects is used to predict the values for the excluded sample, and then Q^2 is calculated for each model. The bootstrapping was repeated 5000 times for each validated model [18].

Obtaining a robust model does not give real information about its prediction power .This is evaluated by predicting the compounds included in the test set.

The external Q2ext for the test set is determined [19] with the equation (4):

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{n_{ext}} (\hat{y}_{i/i} - y_i)^2 / n_{ext}}{\sum_{i=1}^{nt_r} (y_i - \bar{y}_{tr})^2 / n_{tr}} \quad (4)$$

Where y_i and $\hat{y}_{i/i}$ are, respectively, the measured and predicted (over the prediction set) values of the dependent variable, and $\bar{y}$ the averaged value of the dependent variable for the training set. n_{tr} and n_{ext} are the number of objects in the external set, respectively.

Other useful parameters are R^2, calculated for the validation chemicals by applying the model developed on the training set, and an external standard deviation error of prediction ($SDEP_{ext}$), defined as:

$$SDEP_{ext} = \sqrt{\frac{1}{n_{ext}} \sum_{i=1}^{n_{ext}} (y_i - \bar{y})^2} \quad (5)$$

Where the sum runs over the test set objects (n_{ext}).

According to [20] a QSPR model is successful if it satisfies several criteria as follows:

$$R^2_{CV_{ext}} > 0.5 \quad (6)$$

$$r^2 > 0.6 \quad (7)$$

$$\left(r^2 - r^2 0\right) / r^2 < \left(r^2 - r^{2'} 0\right) / r^2 < 0.1 \quad (8)$$

$$0.85 \leq k \leq 1.15 \text{ or } 0.85 \leq k' \leq 1.15 \quad (9)$$

$$Ab = \left|r^2 - r_0'^2\right| < 0.3 \quad (10)$$

Here:

ISSN: 0975-8585

$$r = \frac{\sum (y_i - \tilde{y}_i)(\tilde{y}_i - \bar{\tilde{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\tilde{y}_i - \bar{\tilde{y}}_i)^2}} \quad (11)$$

$$r_0^2 = 1 - \frac{\sum (\tilde{y}_i - \tilde{y}_i^{r_0})}{\sum (\tilde{y}_i - \bar{\tilde{y}}_i)^2} \quad (12)$$

$$r_0'^2 = 1 - \frac{\sum (y_i - y_i^{r_0})^2}{\sum (y_i - \bar{y}_i)^2} \quad (13)$$

$$k = \frac{\sum y_i \tilde{y}_i}{\sum y_i^2} \quad (14)$$

$$k' = \frac{\sum y_i \tilde{y}_i}{\sum \tilde{y}_i^2} \quad (15)$$

$$T1 = \frac{(r^2 - r_0^2)}{r^2} \quad (16)$$

$$T2 = \frac{(r^2 - r_0'^2)}{r^2} \quad (17)$$

Where r^2 is the correlation between the calculated and the experimental values in the test set; r^2 (calculated versus observed values) and r^2 (observed versus calculated values)are the coefficients of determination; k and k' are slopes of regression lines through the origin of calculated versus observed end observed versus calculated, respectively $y_i^{r_0}$, $\tilde{y}_i^{r_0}$ are defined as $y_i^{r_0} = k\tilde{y}_i$ and $\tilde{y}_i^{r_0} = k'y$ and the summations run over the test set.

QSPR model Applicability Domain (AD)

The applicability domain ability (AD) [19,17] is a theoretical region in the space defined by the descriptors of the model and the method response, for which a given QSPR should make reliable predictions. In this work, the structural AD was verified by the leverage (hii) approach [21].

The warning leverage h^* is, generally, fixed at 3(m+1)/n, where n is the total number of samples in the training set and m is the number of descriptors involved in the correlation.

The presence of both the response outliers (Y outliers) and the structurally influential compounds (X outliers) was verified by the Williams plot [22].The plot of standardized residuals versus leverage values.

ISSN: 0975-8585

RESULTS AND DISCUSSION

Several acceptable MLR models of different dimensions, based on various descriptors, were obtained. The best one, taking into account the parsimony principal regarding the complexity of the models, is a model of 4 descriptors with a high predictive power.

The equation (18) the optimal model is given as:

$$FP = -235 + 12.5\,nSK + 416\,FDI - 83.3\,Mor26v + 19.4\,R5u \qquad (18)$$

Here, nsK is a constitutional descriptor (block1); representing the number of non-hydrogen atoms [11].

FDI is folding degree index; belongs to the list of geometrical descriptors calculated by dragon (block12). Geometrical descriptors are defined in several different ways but always derived from the three-dimensional structure of the molecule. Generally, geometrical descriptors are calculated either on some optimized molecular geometry obtained by the methods of the computational chemistry or on crystallographic coordinates. The folding degree index is the largest eigen value of the distance/distance matrix, normalised dividing it by the number of atoms nAT. This index tends to one for linear molecules (of infinite length) and decreases in correspondence with the folding of the molecule. Thus, it can be thought of as a measure of the folding degree of the molecule because it indicates the degree of departure of a molecule from strict linearity [11].

Mor26v is a 3D-Morse descriptor (block 14); 3D-Molecule Representation of Structures based on Electron diffraction) descriptors are based on the idea of obtaining information from the 3D atomic coordinates by the transform used in electron diffraction studies for preparing theoretical scattering curves.
The following expression is used for 3D-MoRSE descriptor calculation:

$$Morsw = \sum_{i=1}^{nAT-1} \sum_{j=i+1}^{nAT} W_i\, W_j\, \frac{\sin(s.r_{ij})}{s.r_{ij}} \qquad (19)$$

Where Mor*sw* is the scattered electron intensity, w is an atomic property, r_{ij} are the interatomic distances and nAT is the number of atoms. The term *s* represents the scattering in various directions by a collection of nAT atoms.

In order to obtain uniform length descriptors, the intensity distribution is made discrete, calculating its value at a sequence of evenly distributed values; in particular, in DRAGON, it is assumed that *s* takes integer values in the range 0 – 31 [11].

R5u (R autocorrelation of lag 5 / unweighted), is a GETAWAY descriptor (block16). GETAWAY descriptors have recently been proposed as chemical structure descriptors derived from a new representation of molecular structure [11].

The obtained statistical parameters are reported in table 2.

Table 2: Statistical parameters of the developed model

n_{tr}	n_{ext}	Q^2_{LOO} (%)	R^2 (%)	R^2_{adj} (%)	Q^2_{ext}	Q^2_{boot}
139	34	97.11	97.41	97.34	97.71	96.96
F	SDEC	SDEP	$SDEP_{ext}$	Kxy	Kxx	s
1261.62	10.09	10.66	9.50	49.77	34.57	10.28

ISSN: 0975-8585

The adjusted $R^2 (R^2_{adj})$ is a better measure of the proportion of variance in the data explained by the correlation than R^2, because R^2 is somewhat sensitive to changes in the number of samples of the training set and the number of descriptors involved in the correlation.

Statistical parameters show that the model (Eq.18) established a strong correlation between the selected variables and the studied property, characterized by an excellent coefficient of determination ($R^2 = 97.41\%$) that explains around 97.41% of data variation, in addition to a very large value of the Fisher F (F=1261.62), which indicates the excellence ability of the model in the prediction of FP values, and a good standard error (s=10.28). Equation (18) presents an R^2_{adj} (%) =97.34 indicating excellent agreement between correlation and variation of the data.

The small difference between R^2 and Q^2_{LOO} informs about the robustness of the model. The cross-validation prediction coefficient illustrates the reliability towards the elimination of the model focusing on the sensitivity towards the elimination of any 5 data. The value of Q^2_{boot} (%) =96.96) confirms both the internal predictability and stability of the model.

A visual comparison of the predicted results of the new correlation with the experimental data is also shown in the plot of observed *versus* predicted values of FP (Figure 1) for the training and test sets confirmed that a linear model with very good fitting can be used to predict our studied property .

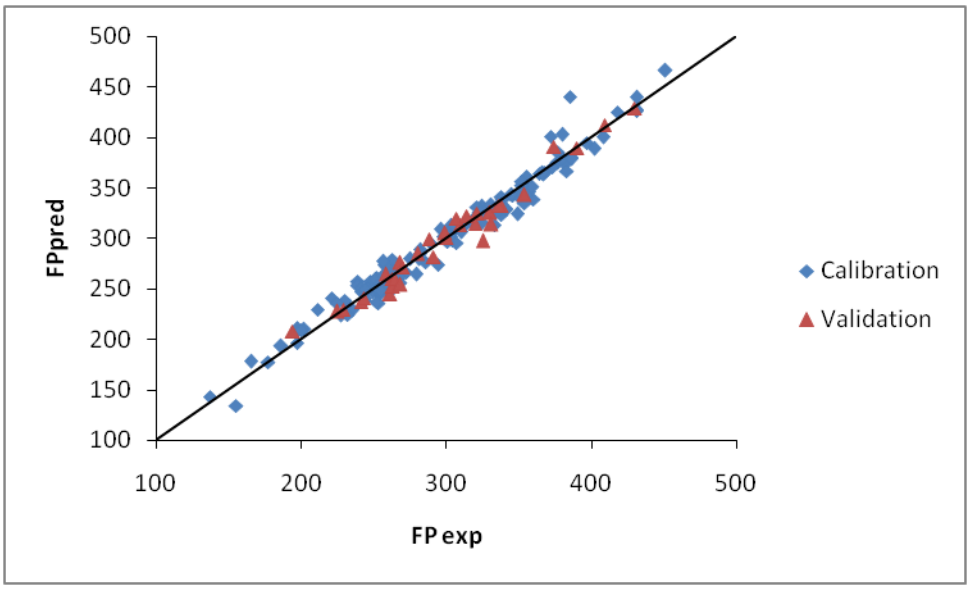

Figure 1: Experimental versus Predicted FP for the training and test sets.

Figure 2 represents the graph of statistical coefficients Q^2 and R^2 which allows comparing the results for randomized patterns (sign +) to the starting model (triangle) which is the real model. It is clear that the flash points statistics obtained for the modified vectors are smaller than those of the real QSPR model, to ensure that a real structure / property (FP) relationship has been established (Figure 2).

ISSN: 0975-8585

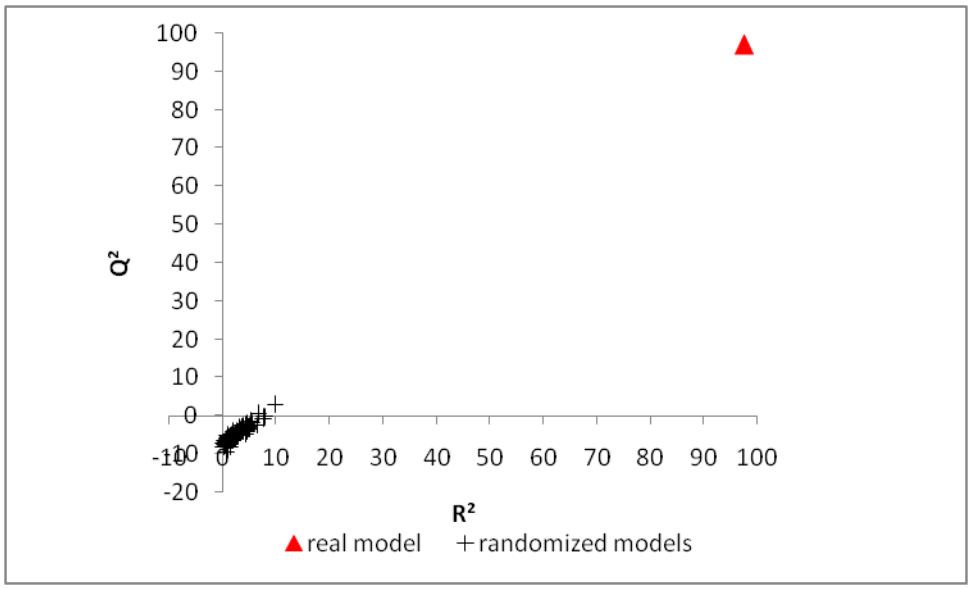

Figure 2: Randomization test associated to previous QSPR model

Signs + represent the randomly ordered flash points, and the triangle corresponds to the real flash point. The statistics for the modified FP vectors are clearly lower than the real QSPR model. Q^2 values are lower than 10 %, and for the major part one obtains even $Q^2 < 0$ for random models symbolized by sign +. This ensures that a real structure –property relationship has been found out.

Based on a previously described procedure [23], the relative contribution of the four descriptors to the model were determined as follow: nsk (45.82%) > FDI (18.55%) > Mor26v (18.34) > R5u (17.29). As it is seen the nsk contribution is greater than FDI , Mor26v and R5u contributions , while the difference in the descriptor contribution is not significant, indicating that the nsk (number of carbon atoms) descriptor is more necessary in generating the predictive model than the other descriptors as seen on figure 3 .

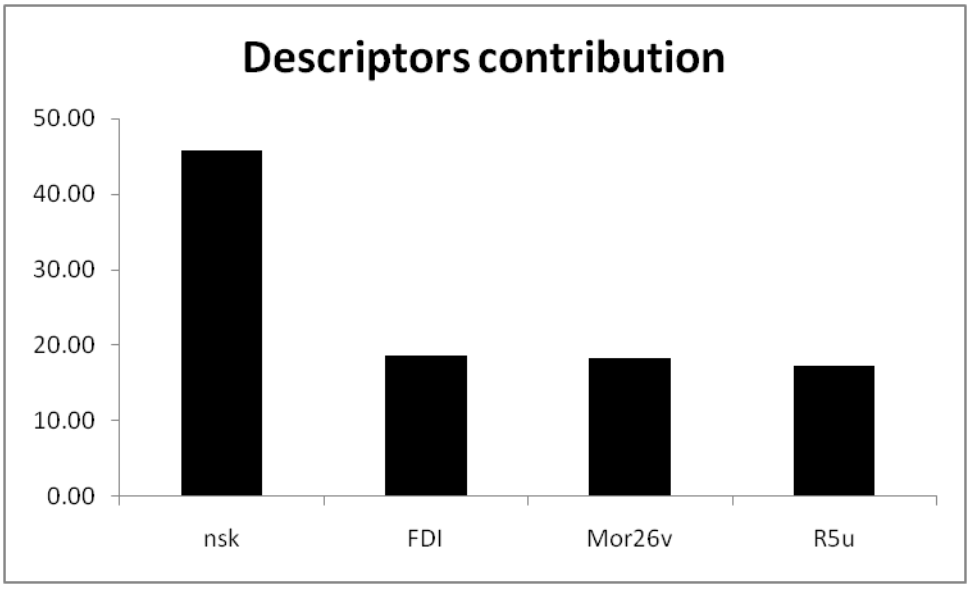

Figure 3: Relative contribution of the selected descriptors in the MLR model

ISSN: 0975-8585

The following statistical parameters according to Tropsha et al. reported in table 3, obtained for the external test set, check the generality accepted conditions which demonstrate the prediction power of the present model.

Table 3: Statistical parameters of Tropsha et al.

$R^2_{CV_{ext}}$	r^2	r_0^2	$r_0'^2$	T1
0.9668	0.968	0.9996	0.9993	-0.0326
T2	k	k′	Ab	
-0.0324	1.0035	0.9955	0.0313	

$$R^2_{CV_{ext}} = 0.9668 > 0.5 \quad ; r^2 = 0.968 > 0.6 \; ; \; r_0^2 = 0.9996$$
$$r_0'^2 = 0.9993 \quad ; \; T1 = -0.0326 \quad < 0.1 \; ; \; T2 = -0.0324 < 0.1$$
$$0.85 < k = 1.0035 < 1.15 \; ; \; 0.85 < k' = 0.9955 < 1.15$$
$$\left| r^2 - r_0'^2 \right| = 0.0313 < 0.3$$

The applicability domain is analyzed using the Williams plot, presented in figure 4 shows standardized residuals in prediction plotted against leverage (Hat diagonal) values of each compound used to evaluate the applicability domain (AD) of a QSPR model suggested by [24] .

The plot makes possible to verify the presence of the outliers objects which are compounds with standardized residual greater than 3 standard deviation units and 9 compounds very influential in the determination of the model parameters which is the compounds with leverage greater than $h^*=3(m+1)/n_{tr} = 0.1079$, where h^* is the warning leverage or the critical value (Figure 4) .

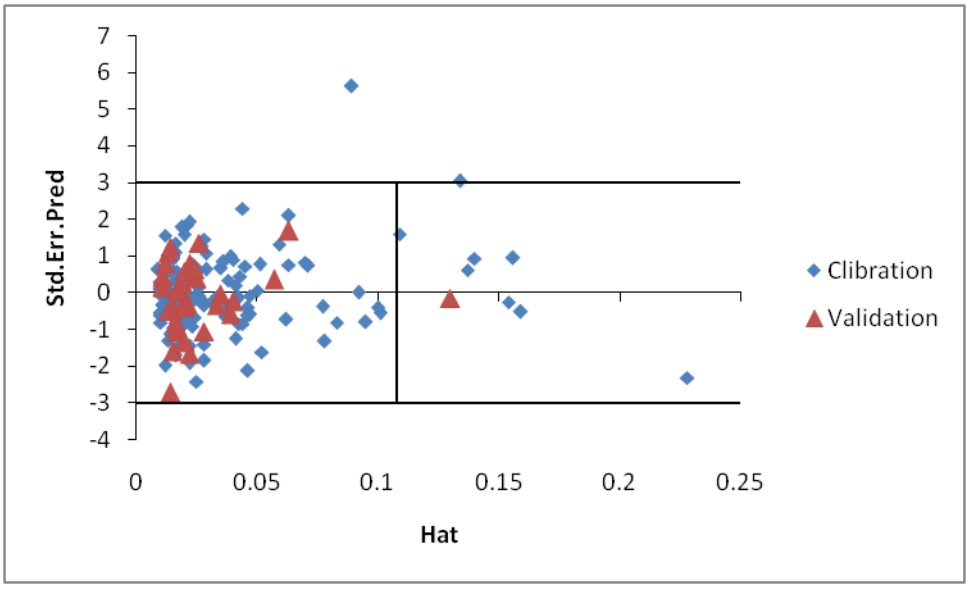

Figure 4: The Williams plot

As it is seen on figure 4 the only outlier object is Tridecylbenzene from the training set with a high FP value and considered as flammable substances. This compound is out of the AD of the QSPR model.

Nine compounds (Hexamethylbenzene , 1,2,4,5-Tetraisopropylbenzene, Ethylene , Acetylene, 2-Methylanthracene, 9-Methylanthracene, 7-Isopropyl-1-methylphenanthrene, 1,3,5-Tri-tert-butylbenzene)

from the training set and one object (1-Methylanthracene) from the test set, are influential objects with Hat values greater than the critical Hat value , but they belongs to the AD of the model.

CONCLUSION

A QSPR model for predicting flash points for 173 unsaturated hydrocarbons was established after applying successive steps beginning from the molecular structure generation to the model generation and the statistical analysis.

The obtained results ensure that the four molecular descriptors explain successfully the studied property which is the flash point. High correlation coefficient $R^2 = 0.9741$, high $Q^2_{ext} = 0.9771$ and the low values of the prediction error (SDEP = 10.66 and $SDEP_{ext}$ = 9.50).confirm the predictive ability of the obtained model.

The results showed that the predicted values of flash points agreed with the experimental values satisfactorily which can sometimes approach the accuracy of experimental flash point determination .Thus this QSPR model using MLR can be successfully used to estimate flash points for new organic compounds or for other unsaturated hydrocarbons for which experimental values are unknown. Furthermore, this work is of assistance to the further study on other flammability characteristics, such as auto ignition temperature and flammability limits, in order to predict the risks of environmental pollution.

REFERENCES

[1] Evlanov SF, KhimZh, Prikl. J ApplChem-USSR (Engl. Transl.) 1991; 64: 747-752.
[2] Tetteh J, Takahiro S, Metcalfe E, Howells S. J ChemInf Comp Sci 1999; 39: 491.
[3] Katritzky AR, Petrukhin R, Jain R, Karelson M. J ChemInf Comp Sci 2001; 41: 1521-1530.
[4] ASTM International, General test Method. 2004;14.02, (ASTM, West. Conshohocken, PA, 2004).
[5] Liaw HJ, Lee YH, Tang CL, Hsu HH, Liu JH. J LOSS PREVENT PROC2002; 15: 429-438.
[6] Lyman J, Reehl WF, Rosenblatt DH. Handbook of Chemical Property Estimation Methods. New York: Mc Graw Hill 1982: 751-752.
[7] Vidal M, Rogers WJ, Holste JC, Mannan MS. A review of estimations method for flash points and flammability limits. Process Saf Prog 2004; 23: 47-55.
[8] Keshavars MH. Indian J Eng Mater S 2012; 19: 269-278.
[9] Keshavarz MH, Ghanbarzadeh M. J Hazard mater 2011; 193: 335- 341.
[10] Hyperchem™, Release 6.03 for windows, Molecular Modeling system. 2000.
[11] Todeschini R, Consonni V, Mauri A, and Pavan M. DRAGON Software for the Calculation of Molecular Descriptor. Version. 5.3 for Windows, Talete S. r. l., Milan. Italy. 2005.
[12] Todeschini R, Ballabio D, Consonni V, Mauri A, and Pavan M. MOBYDIGS Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm. Release 1.1 for windows, Milano. 2009.
[13] Leardi R, Boggia R,Torrile M. J Chemometr1992; 6: 267-281.
[14] Xu J, Zang H, Lei Wang, Liang G, Wang L, Shen X, Xu W. SPECTROCHIM ACTA A 2010; 76: 239-247.
[15] Todeschini R, Maiocchi A, Consonni V. ChemomIntell Lab Syst. 1999; 46: 13-29.
[16] Eriksson L, Jaworska J, Worth A, Cronin M, Mc Dowell RM, Gramatica P. Environ Health Perspect 2003; 111: 1361-1375.
[17] Tropsha A, Gramatica P, Grombar VK. QSAR Comb Sci 2003; 22: 69-76.
[18] Efron B. The jackknife, the Bootstrap and Other Resampling Planes, Society for Industrial and Applied Mathematics, Philadelphia, PA. 1994
[19] Shi LM, Fang H, Tong W, Wu J, Perkias R, Blair RM, Branham WS, Dial SL, Moland CL, Sheehan DM. J ChemInf Comp Sci 2001; 41: 186-195.
[20] Golbraikh A, Tropsha A. J Mol Graph Model2002; 20: 269-276.
[21] Weiberg S. Applied Linear Regression, 3rd edition. (John Wiley and sous, Inc.,New Jersey); 2005.
[22] SCAN-Software for Chememometric Analysis. Version1.1-for Windows, Minitab USA; 1995.
[23] Zheng F, Bayram E, Sumithran SP, Ayers JT, Zhen CG, Schmitt JD, Dwoskim LP, Crooks PA. Bioorg Med Chem 2006; 14: 3017-3037.
[24] Gramatica P.QSAR Comb Sci2007; 26: 694-701.