DOI: http://dx.doi.org/10.32571/ijct.636581 **E-ISSN:2602-277X**

International Journal of Chemistry and Technology

http://dergipark.org.tr/ijct

Research Article

Quantitative modeling for prediction of boiling points of phenolic compounds

Soumaya KHEROUF [1], Nabil BOUARRA [1,2], Djelloul MESSADI [1,*]

on the last page

[1]*Laboratory of Environmental and Food Safety, Department of Chemistry, Badji Mokhtar- Annaba University, PB12, 23000, Annaba. Algeria*

[2]*Centre de Recherche Scientifique et Technique en Analyses Physico-Chimiques, BP 384, Siège ex-Pasna Zone Industrielle, Bou-Ismail CP 42004, Tipaza, Algeria*

Received: 23 October 2019; Revised: 25 November 2019; Accepted: 27 November 2019

*****Corresponding author e-mail: djelloul.messadi@univ-annaba.org

Citation: Kherouf, S.; Bouarra, N.; Messadi, D. *Int. J. Chem. Technol.* **2019**, 3 (2), 121-128.

ABSTRACT

This work aims to reveal the correlation of the boiling point values of phenolic compounds with their molecular structures using a quantitative structure-property relationship (QSPR) approach. A large number of molecular descriptors have been calculated from molecular structures by the DRAGON software. In this study, all 56 phenolic compounds were divided into two subsets: one for the model formation and the other for external validation, by using the Kennard and Stone algorithm. A four-descriptor model was constructed by applying a multiple linear regression based on the ordinary least squares regression method and genetic algorithm/variables subsets selection. The good of fit and predictive power of the proposed model were evaluated by different approaches, including single or multiple output cross-validations, the Y-scrambling test, and external validation through prediction set. Also, the applicability domain of the developed model was examined using Williams plot. The model shows R^2 = 0.876, Q^2_{LOO} = 0.841, Q^2_{LMO} = 0.831 and Q^2_{EXT} = 0.848. The results obtained demonstrate that the model is reliable with good predictive accuracy.

Keywords: Phenolic compounds, boiling point, QSPR, MLR, prediction set.

Fenolik bileşiklerin kaynama noktalarının belirlenmesi için kantitatif modelleme

ÖZ

Bu çalışma, kantitatif yapı-özellik ilişkisi (QSPR) yaklaşımı kullanarak fenolik bileşiklerin kaynama noktası değerlerinin moleküler yapıları ile korelasyonunu ortaya koymayı amaçlamaktadır. DRAGON yazılımı ile moleküler yapılardan çok sayıda moleküler tanımlayıcı hesaplanmıştır. Bu çalışmada, 56 fenolik bileşik Kennard ve Stone algoritması kullanılarak biri model oluşumu için diğeri dış doğrulama için iki alt gruba ayrılmıştır. Sıradan en küçük kareler regresyon yöntemi ve genetik algoritma / değişken altkümeleri seçimine dayanan çoklu bir doğrusal regresyon uygulanarak dört tanımlayıcı model oluşturulmuştur. Önerilen modelin iyi uyum ve tahmin gücü, tahmin seti aracılığıyla tekli ya da çoklu çıkış çapraz validasyonları, Y-kombinasyon testi ve dış doğrulamayı içeren farklı yaklaşımlarla değerlendirilmiştir. Ayrıca, geliştirilen modelin uygulanabilirlik alanı Williams plot kullanılarak incelenmiştir. Model R^2 = 0.876, Q^2LOO = 0.841, Q^2LMO = 0.831 ve Q^2EXT = 0.848'i göstermektedir. Elde edilen sonuçlar, modelin iyi bir tahmin doğruluğu ile güvenilir olduğunu göstermektedir.

Anahtar Kelimeler: Fenolik bileşikler, kaynama noktası, QSPR, MLR, tahmin seti.

1. INTRODUCTION

Lipi phenolic compounds are aryl alcohols in which the hydroxyl group (-OH) is attached carbon atom that is part of an aromatic ring, in which phenol is the simplest of these compounds. Anthropogenic produced phenols exist in the environment due to the activity of the chemical, petrol or industrial processes. The entry of phenolic compounds into ecosystems results from the industrial sewage drainage or the municipal and agricultural activities to surface water.[1] The transport and fate of phenols in the environment depends, in part, on their physicochemical properties, and their relative distributions between different environmental compartments.

DOI: http://dx.doi.org/10.32571/ijct.636581 **E-ISSN:2602-277X**

As is known, the boiling point is an important physical property that has practical value in chemistry, environmental studies and the pharmaceutical industry,[2] defined as the temperature at which the vapor pressure of a pure saturated liquid is 1.013 x 10^5 Pa.[3] It is also an indicator of the physical state of organic chemicals (e.g., liquid or gaseous). Furthermore, critical temperatures,[4] flash points,[5] and enthalpies of vaporization[6] can be predicted or estimated by using boiling point. The boiling point of a molecule depends on two main factors. The first factor includes intermolecular forces, such as Coulomb interactions and dipole dipoles. The second explains the size and structure of the molecule as a whole, that is, how the energy supplied by the heater is distributed in rotation and vibration modes.[7] However, the boiling point data is often not available in the literature and therefore needs to be estimated theoretically. Bp estimation methods have been widely explored[4-13] using the topology of the molecule and/or quantum chemistry parameters calculated for the optimized structure of the molecule.

There has been a remarkable increase in the use of quantitative structure property relationships (QSPR) methodology, which used to predict the physical and chemical properties of organic chemicals.[14] In consequence, the QSPR method attempts to correlate the properties of chemicals with relevant properties and molecular structure descriptors by establishing a simple mathematical relationship.[15]

The purpose of this study is to found a model for the prediction of the boiling point of various phenolic compounds. Many statistical techniques have been used to develop the model to draw attention to the structural requirements for an exact boiling point value. The three objectives of the present paper have been: first, to explore the structure-property relationships of the boiling point; second, to select the best predictive model from among all developed models for the property, and third, verification of the performance and stability of the obtained model. The model obtained shows which descriptors play a significant role in boiling point variation of phenolic compounds.

2. MATERIAL AND METHODS

2.1. Data set

In the present study, the experimental Bp data listed in Table 1 and 2 were received from the Handbook of Physical-Chemical Properties and Environmental Fate for Organic Chemicals.[16] The reported Bp values ranged from 174.9 to 305°C. The database was divided into two molecular subsets by using Kennard and Stone algorithm,[17] the training set consists of 39 compounds and 17 compounds for the prediction set.

2.2. Modeling and molecular descriptor calculation

All numerical calculations were done by a computer with ® Core™ processor and 4Gb RAM. The molecular structure of each compound was sketched using the Hyperchem software[18] and pre-optimized using MM+ molecular mechanics method (Polak-Ribiere algorithm). The minimal energy conformations of molecules were then fully optimized and calculated with the semi-empirical PM3 method at the restricted Hartree-Fock level with no configuration interaction,[19] applying a gradient norm limit of 0.001 kcal $Å^{-1}mol^{-1}$ as a stopping criterion. Lastly, the final geometries with the minimal energy were used as input for the generation of 1664 descriptors using the Dragon software (Version 5.4).[20] Type and information of molecular descriptor calculated are available in Dragon software user's guide.[20]

2.2. Model development and validation

For the model development, by applying multiple linear regression based ordinary least square (OLS), and genetic algorithm-variable subset selection[21] (GA/VSS) implemented in Mobydigs software[22] using the Ordinary Least Squares (OLS) method and GA-VSS (Genetic Algorithm-Variable Subset Selection). This 'variable selection' procedure generates a 'population' of models, ranked according to decreasing Q^2 values. The best models were chosen by using Q^2 leave-one-out (Q^2_{LOO}) as the optimization value and taking into account the parsimony principle regarding the complexity of the models, which should be as small as possible. Furthermore, the correlation between the modeling descriptors and the modeled response was checked by the QUIK rule (Q Under Influence of K), to exclude models with high predictor collinearity and exclude chance correlation.[23]

The goodness of the model was reached by verifying the model fitting and the model robustness, using the squared correlation coefficient R2 (Eq. (1)) and the cross-validation by the leave-one-out technique (Q^2_{LOO}) (Eq. (2)).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

$$Q^2_{LOO} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_{i/i} - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (2)$$

Where y_i is the observed dependent variable (the experimental response), $\hat{y}_i$ is the calculated value by the model, $\bar{y}$ is the mean value of the studied property, n is the number of compounds in the training set, and $\hat{y}_{i/i}$ is the value predicted by the model built without the compound i.

DOI: http://dx.doi.org/10.32571/ijct.636581 **E-ISSN:2602-277X**

Table 1. Molecular descriptors and Bp values for phenols in training set

No	Name	Exp.BP/°C	PW5	Hy	X5A	R6m	Pred. Bp/°C
1	2-Methylphenol	191.04	0.057	-0.16	0.102	0.001	184.104
2	4-Methylphenol	201.98	0.08	-0.16	0.109	0.016	211.419
3	2,6-Dimethylphenol	201.07	0.054	-0.21	0.089	0.005	203.393
4	3,4-Dimethylphenol	227	0.071	-0.21	0.095	0.015	221.586
5	2,4,6-Trimethylphenol	220	0.065	-0.26	0.082	0.018	236.418
6	2-Propylphenol	220	0.082	-0.26	0.104	0.048	226.094
7	4-Propylphenol	232.6	0.085	-0.26	0.1	0.051	240.189
8	2-Isopropylpheonol	213.5	0.062	-0.26	0.096	0.022	203.141
9	4-Isopropylphenol	230	0.072	-0.26	0.091	0.073	242.831
10	4-Butylphenol	248	0.088	-0.29	0.1	0.068	246.757
11	2-sec-Butylphenol	228	0.073	-0.29	0.096	0.068	230.179
12	2-tert-Butylphenol	223	0.059	-0.29	0.091	0.029	207.716
13	3-tert-Butylphenol	240	0.076	-0.29	0.093	0.058	238.970
14	4-sec-Butylphenol	241	0.074	-0.29	0.09	0.092	250.322
15	4-tert-Butylphenol	237	0.066	-0.29	0.084	0.095	250.308
16	4-tert-Octylphenol	279	0.076	-0.4	0.085	0.161	274.000
17	1-Naphthol	288	0.094	-0.29	0.084	0.013	276.601
18	2-Naphthol	285	0.101	-0.29	0.084	0.036	293.926
19	2-Phenylphenol	286	0.095	-0.35	0.087	0.099	289.135
20	4-Phenylphenol	305	0.095	-0.35	0.087	0.093	287.652
21	2-Allylphenol	220	0.082	-0.26	0.104	0.037	223.375
22	4-Chlorophenol	220	0.08	-0.04	0.109	0.01	218.370
23	2,3-Dichlorophenol	206	0.054	-0	0.089	0	217.182
24	3,4,5-Trichlorophenol	275	0.065	0.031	0.082	0.01	254.852
25	4-Nitrophenol	279	0.072	0.031	0.091	0.103	270.657
26	1,2-Dihydroxybenzenene	245	0.057	0.846	0.102	0.001	255.259
27	1,3-Dihydroxybenzene	276.5	0.06	0.846	0.097	0.001	270.704
28	Hydroquinone	285	0.08	0.846	0.109	0.007	280.349
29	2-Methoxyphenol	205	0.063	-0.11	0.101	0.015	203.248
30	3-Methoxyphenol	244	0.074	-0.11	0.102	0.028	222.661
31	2,6-Dimethoxyphenol	261	0.075	-0.12	0.09	0.051	254.253
32	3-Methylphenol	202.27	0.06	-0.16	0.097	0.002	199.797
33	3,5-Dimethylphenol	221.74	0.058	-0.21	0.082	0.007	225.178
34	2,3,5-Trimethylphenol	233	0.065	-0.26	0.082	0.016	235.924
35	3,4,5-Trimethylphenol	248.5	0.065	-0.26	0.082	0.016	235.924
36	2-Ethylphenol	204.5	0.063	-0.21	0.101	0.013	195.241
37	4-Ethylphenol	217.9	0.078	-0.21	0.098	0.04	233.131
38	2,5-Dichlorophenol	211	0.071	-0	0.095	0	232.903
39	2,4,6-Trichlorophenol	246	0.065	0.031	0.082	0.01	254.852

DOI: http://dx.doi.org/10.32571/ijct.636581 **E-ISSN:2602-277X**

Table 2. Molecular descriptors and Bp values for phenols in prediction set

No	Name	Exp.BP/°C	PW5	Hy	X5A	R6m	Pred. Bp/°C
1	3-Ethylphenol	218.4	0.074	-0.213	0.102	0.023	213.913
2	Phenol	181.87	0.062	-0.088	0.113	0	174.129
3	2-Chlorophenol	174.9	0.057	-0.039	0.102	0	192.291
4	3,4-Dichlorophenol	253	0.071	-0.001	0.095	0.01	235.375
5	4-Chloro-3-methyphenol	235	0.071	-0.107	0.095	0.011	228.110
6	2-Nitrophenol	216	0.062	0.031	0.096	0.008	220.091
7	4-Methoxyphenol	243	0.078	-0.107	0.098	0.036	239.655
8	4-Hydroxy-3-methoxybenzaldehyde	285	0.086	-0.119	0.09	0.088	281.690
9	2,3-Dimethylphenol	216.9	0.054	-0.213	0.089	0.003	202.899
10	2,4-Dimethylphenol	210.98	0.071	-0.213	0.095	0.016	221.833
11	2,5-Dimethylphenol	211.1	0.071	-0.213	0.095	0.014	221.339
12	2,4,5-Trimethylphenol	232	0.076	-0.257	0.088	0.025	243.892
13	3-Chlorophenol	214	0.06	-0.039	0.097	0	207.736
14	2,4-Dichlorophenol	210	0.071	-0.001	0.095	0.01	235.375
15	2,6-Dichlorophenol	220	0.054	-0.001	0.089	0	217.182
16	3,5-Dichlorophenol	233	0.058	-0.001	0.082	0	238.472
17	2,4,5-Trichlorophenol	247	0.076	0.031	0.088	0.01	260.596

Besides, the Root Mean Squared of Error (RMSE) that resume the overall error of the model, which used to measure and compare prediction accuracy in the training ($RMSE_{tr}$) and the prediction ($RMSE_{p}$) sets defined in Eq. (3).

$$RMSE_{tr\,(p)} = \sqrt{\frac{1}{n_{tr(ext)}} \sum_{i=1}^{n_{tr\,(ext)}} (y_i - \hat{y}_i)^2} \qquad (3)$$

A stronger internal validation is performed by using the LMO (leave-many-out) procedure. By design, model validation by LMO employs smaller training sets than the LOO procedure and can be repeated many more times due to the possibility of larger combinations in leaving many compounds out from the training set, it is common to choose 5-40% of the entire number molecules in the training set to be left . The premise is that if a QSPR model has a high average in Q^2_{LMO} validation, we can reasonably conclude that the obtained model is robust.[24]

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 for the test set is determined by using Eq. (4):

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

Here n_{ext} and n_{tr} are the number of objects in the external set and the number of training set objects, respectively.

In order to exclude the possibility of a chance correlation between the selected descriptors and the studied response, Y-scrambling as an internal validation method was used. In this method, the dependent-variable vector, Y-vector, are randomly permuted and a new QSPR model is developed using the selected descriptor in the model.[25]

A successful QSPR model should be validated with the test set and satisfies criteria in Equations (5-9).[24, 25]

$$Q^2_{LOO} > 0.5 \qquad (5)$$

$$Q^2_{ext} > 0.5 \qquad (6)$$

$$R^2 > 0.6 \qquad (7)$$

$$(R^2 - R^2_0) / R^2 < 0.1 \text{ or } (R^2 - R'^2_0) / R^2 < 0.1 \qquad (8)$$

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

Where R^2_0 (predicted versus observed values) and R'^2_0 (observed versus predicted values) are coefficients of determination, k and k' are slopes of regression lines through the origin of predicted versus observed and observed versus predicted respectively. Mathematical definitions of parameters (R^2_0, R'^2_0, k, and k') can be found in the literature.[24,26]

DOI: http://dx.doi.org/10.32571/ijct.636581 **E-ISSN:2602-277X**

2.3. Applicability Domain (AD)

Williams plot, the plot of standardized residuals versus the leverage (h_{ii})[27,28] is always used to verify the applicability domain (AD) of the developed QSPR model. The structural AD is quantified by applying the leverage approach, this approach is based on the calculation of the Hat matrix for the structural domain. Leverage indicates a compounds distance from the centroid of X. The leverage of a compound in the original variable space is defined as in Eq. (10):[27]

$$h_{ii} = x_i^T (X^T X)^{-1} x_i \quad (10)$$

Where x_i is the descriptor vector of the considered compound and X is the descriptor matrix derived from the training set descriptor values. The warning leverage h^* was calculated according to Eq. (11)[29]

$$h^* = 3(p + 1) / n \quad (11)$$

Where p is the number of independent variables used and n is the number of compounds in the training set.

The leverages approach was used to estimate the degree of extrapolation for the predictions obtained in the training and prediction sets and for compounds without experimental data. The chemicals with a leverage *hii* > h^* are outside the structural domain of the training set. therefore, their predictions are extrapolations and could be less reliable.

3. RESULTS AND DISCUSSION

The hybrid method genetic algorithm/Multiple linear Regression (GA/MLR) included in MOBYDIGS software was used to select the best descriptors able to explicating property variation in the training set. Finally, a 4-variables model was chosen as the best model. The regression equation of the developed model defined as follows:

$$BP = 314 + 1663\,\boldsymbol{PW5} + 70.9\,\boldsymbol{Hy} - 2091\,\boldsymbol{X5A} + 247\,\boldsymbol{R6m} \quad (12)$$

N_{tr} = 39, N_{pr} = 17, S = 11.111°C, RMSEtr = 10.375, RMSEpr = 11.4928. F = 60.4552

R^2 = 0.876, Q^2_{LOO} = 0.841, Q^2_{LMO} = 0.848, Q^2_{ext} = 0.848.

The statistical parameters that evaluate the model are listed in Table 3, from this table we can conclude that the statistical parameters of the developed model have very good predictive performance and that the descriptors in which it is involved describe well the boiling point. The developed model satisfies the above accept conditions (Equations (5-9)).

Statistical parameters and the meanings of descriptors are grouped in Table 4. The probability (P) that the descriptor is there by chance, should usually be less than 0.05 (i.e., 5%) to be considered statistically significant; otherwise this descriptor should be thrown out, the P-values suggests that all the descriptors in each model are significant. The high absolute t-values means that the regression coefficients of the descriptors introduced in the model are significantly larger than the standard deviation, values of VIF less than 5 indicate that the descriptors are not strongly correlated with each other.[30]

Another proof of the model quality is the strong correlation between observed and predicted Bp values for both training and prediction sets. Figure 1 illustrates the predicted values of the boiling point versus the experimental values, the correlation coefficient (R^2 = 0.876) of this plot indicates the good agreement between these values, which prove the reliability of the model.

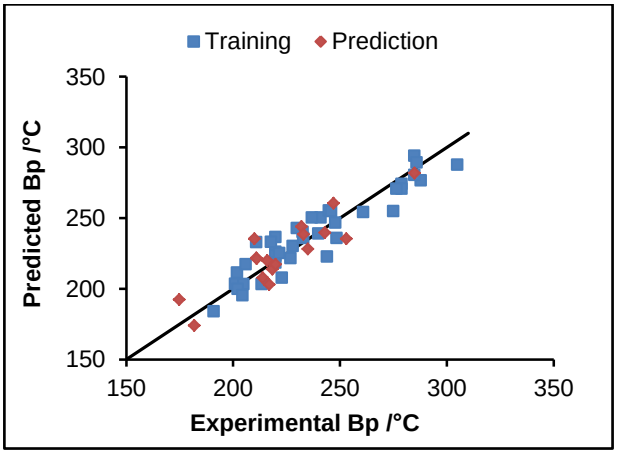

Figure 1. Scatter plot of experimental and predicted boiling point.

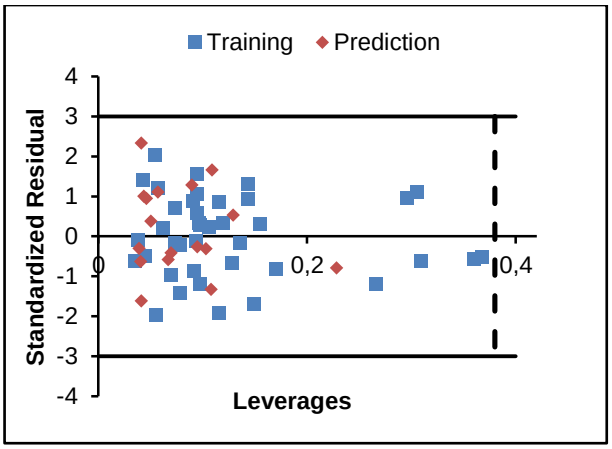

Figure 2. The respective Williams plot of standardized residual versus leverages for the model.

DOI: http://dx.doi.org/10.32571/ijct.636581 **E-ISSN:2602-277X**

Table 3. Evaluation results of the developed model

Training set			Prediction set				
R^2	Q^2_{LOO}	Q^2_{LMO}	Q^2_{ext}	$(R^2-R^2_0)/R^2$	$(R^2-R'^2_0)/R^2$	k	k'
0.8767	0.841	0.831	0.848	0.0059	0.0244	0.9907	1.0068

Table 4. Characteristics of the selected descriptors in the model

Predictor	Descriptors signification	Coefficient	SE Coef	T	P	VIF
Constant		313.58	22.68	13.83	0.000	
PW5	Path/walk 5 - Randic shape index	1662.9	173.3	9.59	0.000	1.358
Hy	Hydrophilic factor	70.871	6.910	10.26	0.000	1.364
X5A	Average connectivity index of order 5	-2091.3	235.5	-8.88	0.000	1.222
R6m	R autocorrelation of lag 6 / weighted by mass	247.16	57.83	4.27	0.000	1.471

Figure 2 shows the Williams plot. As can be seen in this figure, all residuals were situated on the range of ±3 standard deviations (horizontal lines) and also there is no structural influential compound both for training and prediction sets (here, the leverage value hi of all data sets are lower than the warning value h^* = 0.38), which means that the model has a good predictive ability. So, the developed model could be used to predict the boiling point of phenolic compounds just from their molecular structure.

The Y-scrambling tests between the original and permuted response data indicate the robustness of the model. The significant low values of R^2_{Yscr} and Q^2_{Yscr} (black circle) obtained for 100 iterations confirm the robustness of the developed model (asterisk). Figure 3 shows that in the case of all randomized models, the values of R^2Y_{scr} and Q^2Y_{scr} were < 0.5 this insure that the good results of the original model have a real basis, not due to chance correlation.

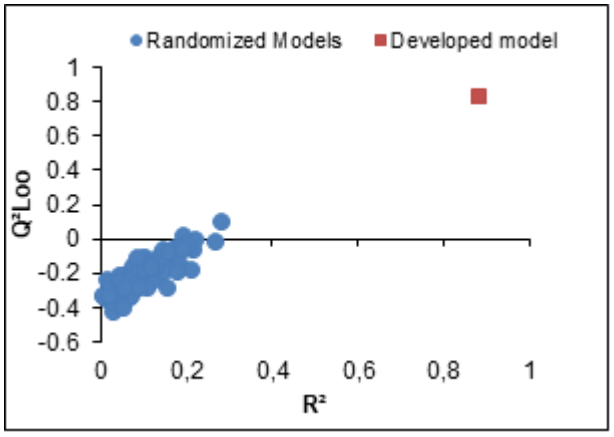

Figure 3. Randomization test: R^2 and Q^2 of Y-scrambled models compared with the original model.

3.1. Descriptor contribution and interpretation

In order to verify the relative contributions of the four descriptors in the developed model, they are illustrated in Figure 4. The importance of the descriptors included in the model decreases in the following order: *Hy* (28.155%) > *PW5* (27.476%) > *X5A* (26.844%) > *R6m* (17.524%).

The most significant descriptor is a hydrophilic factor (*Hy*), which gives a correlation with Bp of 0.196 and explains 28.155% of the contributions. The hydrophilic factor is a simple empirical index related to the hydrophilicity of compounds based on count descriptors.[31] It is a measure of the number of hydrophilic functional groups (-OH,-SH,-NH).[32] The Molecular property descriptor here digitizes the hydrophilic properties of the molecules caused by the group "OH".[31]

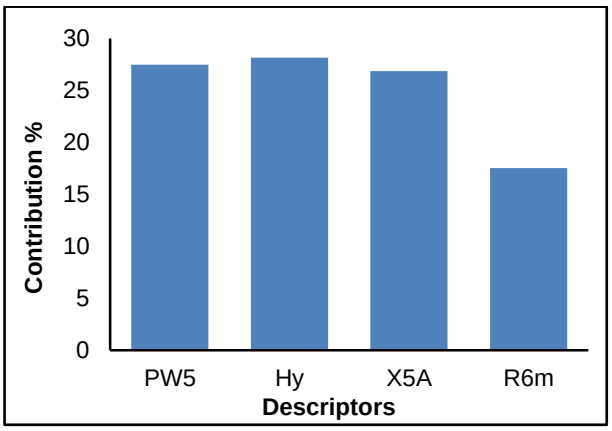

Figure 4. Relative contributions of the descriptors in the developed model.

DOI: http://dx.doi.org/10.32571/ijct.636581 **E-ISSN:2602-277X**

The second significant descriptor is *PW5*, which gives a correlation with Bp of 0.566 and explains 27.47% of the contributions. *PW5* is a topological index that considers the shape of molecules as molecular properties in the variations of compounds.[33] The shape of molecules with a specific kind of branching was also selected as a significant descriptor in the QSAR model developed by Mitra and co-workers.[34] The variation in branching structural features has also been considered by Ray and co-workers[35] to develop the specific QSAR model. *PW5* refers to the proportions of path/walk-in length 5 from the molecular Randic shape index. Randic[36] characterizes shape index for a molecular graph by considering both paths and walks of different lengths within a graph and then making the proportions of the number of path and the number of walks the same length. The third descriptor is *X5A*, this descriptor belongs to Connectivity indices descriptors; *X5A* is the fifth-order average connectivity index appearing in the MLR model mainly shows the topological characteristics. Topological indices are numerical quantifiers of molecular topology and an H-depleted molecular graph. They involve one or more structural features of the molecule such as the size, shape, symmetry, and branching and can also codify chemical information about atom type and bond multiplicity.[37] The last descriptor is *R6m* belongs to GETAWAY descriptor, and provides information on the molecular leverage autocorrelation of lag 6 weighted by atomic mass. *R6m* is a geometrical descriptor encoding information on the real position of substituent and fragments in the molecular space.[38,39]

4. CONCLUSIONS

In this work, quantitative relationships between the boiling point and some phenolic compounds, and their molecular descriptors were investigated by using multiple linear regression techniques. Genetic algorithm is a powerful method used to reduce the number of descriptors in the development of the models. The selected descriptor in this study gives a good estimate for boiling point, indicated by several calculated metrics, R^2 for the goodness of fit, Q^2_{LOO} and Q^2_{LMO} for robustness, and Q^2ext for the predictive power of the model. Also, the applicability domain of the MLR model is verified by the leverage approach. In conclusion, the model proposed in this work provides a feasible, effective and practical tool to predict the boiling point of phenolic compounds.

Conflict of interests

Authors declare that there is no a conflict of interest with any person, institute, company, etc.

REFERENCES

1. Michałowicz, J.; Duda, R.O.W. *Water. Air. Soil. Pollut*. **2005**,16, 205-222.

2. Todeschini, R.; Gramatica, P.; Provenazi, R.; Marengo, E.; *Chemom. Intell. Lab. Syst*. **1995**, 27, 221-229.

3. Gharagheizi, F.; Mirkhani, S. A.; Ilani-Kashkouli, P.; Mohammadi, A. H.; Ramjugernath, D. Richon, D.; *Fluid Phase Equilib*. **2013**, 354, 250-258.

4. Katritzky, A. R.; Mu, L.; Lobanov, V. S. *J. Phys. Chem*. **1996**, 100, 10400-10407.

5. Yi-min, D.; Zhi-ping, Z.; Zhong, C.; Yue-fei, Z.; Ju-lan, Z.; Xun L. *J. Mol. Graphics. Modell.* **2013**, 44, 113-119.

6. White, C.M. *J. Chem. Eng. Data*. **1986**, 31, 198-203.

7. Smeeks, F.C.; Jurs, P.C. *Anal. Chim. Acta*. **1990**, 233, 111-119.

8. Admire, B.; Lian, B.; Yalkowsky, S. H. *Chemosphere*. **2015**, 119, 1436–1440.

9. Shuai, D. ;Wen, S.; Li, Zhao. *Int. J. Refrig*. **2016**, 63, 63–71.

10. Liangjie, J.; Peng B. *Chemom. Intell. Lab. Syst*. **2016**, 157, 127–132.

11. Ramane, H. S.;· Yalnaik, A. S. *J. Appl. Math. Comput*. **2017**, 55, 1–2, 609–627.

12. Varamesh, A.; Hemmati-Sarapardeh, A.; Dabir, B.; Mohammadi, A.H. *J. Mol. Liq*, **2017**, 242, 59-69.

13. Arjmand, F.; Shafiei, F. *J. Struc. Chem*. **2018**, 59, 3, 748-754.

14. Katritzky,A. R.; Maran, U.; Lobanov,V. S.; Karelson, M. *J. Chem. Inf. Comput. Sci*. **2000**, 40, 1-18.

15. Katritzky, A. R.; Lobanov,V. S.; Karelson, M. *Chem. Soc. Rev*. **1995**, 24, 279-287.

16. Mackay, D.; Shiu,W.Y.; Ma, K.C.; Lee, S.C. Handbook of Physical-Chemical Properties and Environmental Fate for Organic Chemicals, CRC Press Taylor & Francis Group 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742.Vol III, 2006.

DOI: http://dx.doi.org/10.32571/ijct.636581 **E-ISSN:2602-277X**

17. Kennard, R.; Stone, L.A. *Technometrics*. **1969**, 11, 137-148.

18. HyperChem 6.03 Package. Hypercube, Inc., Gainesville, Florida, USA, 1999, software available at: http://www.hyper.com.

19. Gaber, M.M. Scientific Data Mining and Knowledge Discovery: Principles and Foundations; Springer Heidelberg Dordrecht London, Berlin, 2009.

20. Talete Srl. *Dragon for Windows (Software for Molecular Descriptor Calculation)* Version 5.5 Milano, Italy, 2007, software available at: http://www.talete.mi.it.

21. Leardi, R.; Boggia, R,; Terrile, M. *J. Chemom*. **1992**, 6, 267-281.

22. Todeschni, R.; Ballabio, D.; Consonni, V.; Mauri, A.; Pavan, M. 2009. Mobydigs – version 1.1 – Copyright TALETE Srl.

23. Todeschini R.; Maiocchi A.; Consonni, V. *Chemom. Int. Lab. Syst*. **1999**, 46, 13-29.

24. Tropsha, A.; Gramatica, P.; Gombar, V. K. *QSAR Comb. Sci*. **2003**, 22, 70-77.

25. Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Model*. **2002**, 20, 269-276.

26. Yu, X. L.; Yi, B.; Yu,W. H.; Wang, X. Y. *Chem. Pap*. **2008**, 62, 623-229.

27. Netzeva, T.I.; Worth, A.P.; Aldenberg, T.; Benigni, R.; Cronin, M.T.D.; Gramatica, P.; Jaworska, J.S.; Kahn, S.; Klopman, G.; Marchant, C.A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G.Y.; Perkins, R.; Roberts, D.W.; Schultz, T.W.; Stanton, D.T.; vande Sandt, J.J.M.; Tong, W.;Veith, G.; Yang, C. *Altern. Lab. Anim*. **2005**, 33, 155–173.

28. Gramatica, P.; Cassani, S. Roy, P. P. Kovarich, S.; Yap. C. W. Papa, E. *Mol. Inform*. **2012**, 31, 817-835.

29. Eriksson, L.; Jaworska, J.; Worth, A.P.; Cronin, M.T.D.; McDowell, R.M.; Gramatica, P. *Environ. Health. Perspect*. **2003**, 111, 1361-1375.

30. Jaiswal, M.; Khadikar, P.V.; Scozzafava,A.; Supuran, C.T. *Bioorg. Med. Chem. Lett*. **2004**, 14, 3283-3290.

31. Todeschini, R.; Consonni, V. Molecular Descriptors for Chemoinformatics; Wiley-VCH: Weinheim, Germany, 2009.

32. Gramatica, P.; Navas, N.; Todeschini, R. *Trends Anal. Chem*. **1999b**, 18, 461-471.

33. Randic, M.; Razinger, M. *J. Chem. Inf. Model*. **1995**, 35, 140-147.

34. Mitra I.; Saha A.; Roy K. *Mol. Simul*. **2010**, 36, 1067-1079.

35. Ray, S.; Sengupta, C.; Roy K. *Cent. Eur. J. Chem*. **2007**, 5, 1094-1113.

36. Randic, M. *J. Chem. Inf. Model*. **2001**, 41, 607-613.

37. Abbasi, M.; Sadeghi-Aliabadi, H.; Amanlou, M. *J. Pharm. Sci*. **2017**, 25, 1-17.

38. Consonni,V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci*. **2002**, 42, 682-692.

39. Consonni,V.; Todeschini, R.; Pavan, M.; Gramatica, P. *J. Chem. Inf. Comput. Sci*. **2002**, 42, 693-705.

ORCID

https://orcid.org/0000-0001-9797-3746 (S. Kherouf)

https://orcid.org/0000-0001-5438-8678 (N. Bouarra)

https://orcid.org/0000-0003-3257-9590 (D. Messadi)