ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

QSPR Application on Modeling of Boiling Point of Polycyclic Aromatic Hydrocarbons.

N Bouarra[1,2], S Kherouf[1], A Bouakkadia[1,3], and D Messadi[1]*.

[1]Laboratory of Environmental and Food Safety, Department of chemistry, BADJI Mokhtar Annaba University, PB12, 23000, Annaba. Algeria.
[2]Center of Scientific and Technical Research in Physico-Chemical analyzes (CRAPC), BP 384, Siège ex-Pasna Zone Industrielle, Bou-Ismail CP 42004, Tipaza, Algeria.
[3]University Abbes LAGHROUR Khenchela - Algeria -BP 1252 Route de Batna Khenchela 40004

ABSTRACT

Physicochemical properties of organic pollutant play an important key role to understand their behavior in environment. However, the information behind the property-behavior phenomena of chemical compounds is less found in the literature. Therefore, computational methods had to be applied for process optimization. In the present paper the applicability of the (QSPR), Quantitative structure-property relationship models based on molecular descriptors derived from molecular structures have been developed for the prediction of boiling point using a set of 61 polycyclic aromatic hydrocarbons (PAHs). For this purpose multiple linear regression was used to construct QSPR model. Best obtained model had following statistical parameters: R^2= 99.83%, Q^2_{LOO} = 99.80%, s= 4.587, F=11423.29. The statistical quality of the prediction results agree well with the experimental values of this property. The Insubria graph was applied to verify the reliability of the predictions of the full model for another 57 PAHs with unknown experimental data of boiling point, which belong to the model applicability domain.

Keywords: QSPR; multiple linear regression; applicability domain; external validation; EPS0.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Polycyclic aromatic hydrocarbons (PAHs) are a group of persistent pollutants widely present in the environment, constituted by hundreds of compounds containing at least two aromatic rings [1]. Due to the human activities such as oil spills, burning fossil fuels, domestic wastes, transport emissions incinerators and cigarettes smoke the PAHs introduced in water, soils, sediment and atmospheric environment [2]. Investigations of environmental behavior of PAHs have attracted considerable attention [3], due to their toxic, mutagenic and carcinogenic potentials. Particularly benz [a] anthracene, chrysene, dibenz [a,h] anthracene and benzo [a] pyrene [4], included in the priority list of pollutants of US-EPA[5]. The distribution of PAHs among different phases and environmental compartment is an important factor in determining their fate and assessing their risk. Physical-chemical properties, such as boiling point is used to describe the volatility of chemicals (its presence in the atmospheric environment) [6], defined as the temperature at which a substance has a vapor pressure of 760 mmHg [7]. Also boiling point is a function of a number of molecular properties that control the ability of a molecule to escape from the surface of a liquid into the vapor phase. These properties are molecular size (which controls dispersion interactions within the liquid phase, and also dictates the energy required to create a cavity in the liquid), and polar and hydrogen bonding forces [8].Moreover, boiling points can be used to predict or estimate other physical properties [9], such as critical temperatures [10], and flash points [11].

For many PAHs, the values of boiling point are not available in the literature. Their experimental measurement is expensive, consuming-time and it requires pure compounds. Moreover, the compounds of high molecular weight decompose before reaching their boiling points and require measures under reduced pressure and subsequent correction for atmospheric pressure. Therefore, the direct measurement of the boiling point of the organic compound is laborious [12]. Thus, it is valuable to be able to predict these properties by using Quantitative Structure Property Relationships (QSPR) techniques, which have the additional advantage of being quick to use.

A QSPR is a mathematical description of a property in terms of other properties (descriptors) that are of three broad classes—hydrophobic, electronic, and steric. Since the advent of modern QSAR (Quantitative Structure Activity Relationships) investigations in 1962 [13], many attempts to model physicochemical properties have been published. In the case of boiling point of PAHs several studies have modeled this property. White [14] used simple 1-parameter linear correlation, involving first-order valence molecular connectivity, Todeschini [6] used WHIM descriptors, while Ferreira [5] used the molecular volume (Vol), surface area (SArea), enthalpy of formation(ΔH_f), electron affinity (EA), connectivity indices (Xe, Xv) and Wiener index (W) by applying Partial Least-Squares (PLS) regression. Later, Ribeiro and Ferreira [15] presented a 3 descriptor model using the volume (V), molecular weight (MW) and Randic connectivity index (R) by the means of principal components regression (PCR) and PLS regression method.

In this study, we present a QSPR model for the prediction of the boiling point for polycyclic aromatic hydrocarbons. In aim to develop a simple, fast, accurate, and less expensive method for calculation of boiling point values. The predictive power of resulting model is demonstrated by testing it on unseen data that were not used during model generation. To verify the reliability of the predictions for another 57 PAHs with unknown boiling point values the Insubria graph was applied.

MATERIAL AND METHODS

Developing a QSPR model for boiling point involves six distinct steps: (i) data collection; (ii) molecular geometric optimization; (iii) descriptor calculation; (iv) model development; (v) model performance evaluation; and (vi) applying the model to predict the values of Bp for PAHs, for which the experimentally values of Bp have been unavailable.

Data set

The experimental boiling point values of the present work were obtained from [16, 17]. Boiling point range was from 491 to 869 K. A complete list of the compounds names, CAS number and corresponding experimental boiling points are given in the table 1.

Molecular modeling and descriptors generation

The theoretical molecular descriptors for 61 PAHs were calculated by the following process. Firstly, the molecular structures were pre-optimized by MM+ molecular mechanics force field in HyperChem 6.03 package [18]. The final geometry of the minimum energy conformation was obtained by the semi-empirical method PM3 with a restricted Hartree-Fock level without interaction configuration, applying a standard limit gradient 0.001 Å kcal.mol^{-1} as a stopping criterion. The output files exported from Hyperchem were transferred into Dragon software[19], to calculate a large number of molecular descriptors on the basis of the geometrical en electronic structure of the molecules. Quantum chemical descriptors (dipole moment, energies of the frontier orbitals: Ɛ HOMO, Ɛ LUMO, etc) were generated by HYPERCHEM software. After the generation of 1675 descriptors for each of 61 polycyclic aromatic hydrocarbons, a pre-selection of descriptors was performed with the aim to reduce the pool of descriptors by eliminating those that satisfy the following conditions: (a) the descriptor has a constant or near constant value for all molecules investigated; (b) in the mono parametric correlations with boiling point the descriptors has a squared correlation coefficient lower than 0.1; (c) the descriptors has an inter-correlation coefficient higher than 0.95 with another descriptor. The pre-selection was performed in Dragon software.

Data splitting

It is important to rationally define a training set for which the model is built and external test set on which to evaluate its prediction power. The object of this selection should be to generate two sets with similar molecular diversity, in order to be reciprocally representative and to cover all the main structural and physicochemical characteristics of the global data set. Several procedures can be adopted for the selection of the training and test sets, the latter should contain between 15 and 40% of the compounds in the full data set. In this work three different splitting techniques were applied: (a) random splitting, (b) sorted response splitting and (c) structural similarity ordered by the first axis of Principal Component Analysis (PCA, PC1 score) [20].

QSPR modeling and validation

The best modeling variables were selected by exploring the statistical quality of all the possible combinations of the available experimental descriptors, by applying multiple linear regression based on ordinary least squares (MLR-OLS) in the QSARINS software (version 2.2) [21]. This 'variable selection' procedure generates a 'population' of models, ranked according to decreasing R^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. For this reason, only up to three descriptors were included in the QSARs generated in this study. 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 colinearity and exclude chance correlation [22].The best modeling descriptors were selected using all subset procedure of QSARINS software [21, 22]. The search of the best solution was made by maximizing a selected fitness function, in our case Q^2_{LOO}.

The development of QSPR models is encouraged providing that they respect the five driving principles for the validation of QSPR models drawn up by OECD (2007) [23]: 1) a defined endpoint, 2) an unambiguous algorithm, 3) a defined domain of applicability, 4) appropriate measures of goodness-of-fit, robustness and predictive power, 5) a mechanistic interpretation, when it's possible.

The goodness of the model was reached by verifying the model fitting and the model robustness by calculating respectively the determination coefficient in the training set (R^2) and the cross-validation coefficient (Q^2_{CV}) by the leave-one-out method (LOO), the most known method to evaluate the models robustness [23,24]

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

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

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 dependent variable, n is the number of compounds in the training set, and $\hat{y}_{i/i}$ is the value of BP predicted by the model built without the compound i according to LOO method.

A stronger internal validation is performed by using the leave-many-out (LMO) procedure. By design, model validation by LMO employs smaller training sets than the LOO procedure and can be repeated many more times due to possibility of larger combinations in leaving many compounds out from the training set The premise being 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]. In our case, 30% of chemicals are put aside from training set with 2000 iterations.

To exclude the possibility of chance correlation between the selected modeling descriptors and studied response, Y-scrambling is another internal validation method. In this test, the dependent-variable vector, Y-vector, is randomly shuffled and a new QSPR model is developed using the selected descriptor in the original model [24]. The process is repeated several times (2000 times in our case).Low values of the averaged R^2scrambled (R^2_{ys}) are indicative of a well founded (not by chance) original model.

Moreover the predictive power of the developed model was evaluated by the external validation set on a series of coefficients: R^2_{EXT}(characterizing the correlation between predicted and experimental values in the validation set) the coefficients Q^2_{F1} [25], Q^2_{F2} [26], Q^2_{F3} [27, 28] and CCC [29-32].

In addition, the Root Mean Squared Error ($RMSE$) that summarizes the overall error of model, was used to measure and compare prediction accuracy in the training ($RMSE_{tr}$) and in the prediction ($RMSE_{pr}$) sets. Defined as follow:

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

And the mean absolute error, or MAE, which describe the difference between the model predictions and observations in the units of the variable, given by

$$E = \frac{1}{n} \sqrt{|y_i - \hat{y}_i|} \quad (4)$$

Applicability domain (AD)

The third OECD principal requires a defined applicability domain. Any QSPR model must be verified for its applicability domain for defined if the model is to be used for screening new chemicals [24]. The Williams plot, the plot of standardized residuals versus the diagonal values of hat matrix (hi leverage), was used to visualize respective applicability domain [23]. The leverage h_i [31] and warning leverage h^* [33] are defined with the following expressions:

$$h_i = x_i^T (X^T X)^{-1} x_i \ (i = 1, \ldots n) \quad (5)$$

$$h^* = \frac{3(p+1)}{n_{tr}} \quad (6)$$

Where x_i is the descriptor vector of the considered compound, X is the model matrix derived from the training set descriptor values, n_{tr} is the number of training compounds and p is the number of model parameters.

Williams plot was used to detect outlier's responses (i.e. chemicals with absolute standardized residuals greater than 3 standard deviation units) and the structurally influential chemicals (i.e. chemicals with leverage values greater than the threshold value, $h_i > h^*$). In case of chemicals without experimental data, the use of the Insubria Graph, which is a plot of h_i diagonal values versus predicted values, can provide a visualization of interpolated and extrapolated predictions [34].

ISSN: 0975-8585

RESULTS AND DISCUSSION

All Subset-Multiple linear Regression procedure, included in QSARINS software, was used to select optimal descriptors capable of explaining property variation among the training set. For the study and the prediction of boiling point of PAHs, we selected, as the best, the models with the most significant descriptors according to the GA-MLR algorithm are given in the equation of the selected model (random splitting) defined as:

$$Bp(K) = 509.51(\pm 28.04) + 28.213(\pm 3.091)\,HOMO + 34.101(\pm 0.283)\,EPS0 \quad (7)$$

N_{tr} = 42, R^2= 99.83%, Q^2_{LOO} = 99.80%, R^2_{ext} = 99.77%, Q^2_{LMO} = 99.80%, Q^2_{F1} = 99.79%, Q^2_{F2}= 99.72%, Q^2_{F3} = 99.74%, CCC_{ext} = 99.86%, $RMSE_{tr}$ = 4.420, $RMSE_{cv}$ = 4.770, $RMSE_{pr}$ =5.425, S = 4.587.

Statistical parameters shows that the model (Eq.7) established a strong correlation between the 2 selected variables and the studied property, characterized by excellent parameters, in addition to a very large value of the Fisher F (=11423.299), which indicates the excellence of the model in prediction of boiling point values, and a good standard error (s = 4.587). Equation.7 presented a value of R^2_{adj} = 99.82% indicating excellent agreement between correlation and variation of the data, also the low value of R^2ys indicating that the obtained model has no chance correlation. All statistical parameters of the model are satisfying and prove that the model is stable, robust and predictive. Then, the built model was used to predict the test set data. Table.1 showed the prediction results.

Table 1: CAS numbers of studied compounds and their experimental and predicted boiling points.

CAS	Exp. Bp (K)	Pred.by model eq.	ei (1)	Predicted by EPISUITE	ei (2)	CAS	Exp. Bp (K)	Pred. by model eq.	ei (1)	Predicted by EPISUITE	ei (2)
000091-57-6	514	518.5673	4.567	522.6	8.6	000205-99-2	754	751.9209	2.079	715.75	38.25
000090-12-0	518	521.8923	3.892	522.6	4.6	000207-08-9	754	757.5619	3.561	715.75	38.25
000581-42-0	535	536.66	1.66	539.66	4.66	000192-97-2	769	761.1749	7.825	715.75	53.25
000582-16-1	535	535.3904	0.390	539.66	4.66	000213-46-7	792	803.5801	11.58	743.09	48.91
000575-37-1	536	539.1162	3.116	539.66	3.66	000050-32-8	769	767.6623	1.337	715.75	53.25
111495-85-3	538	539.4266	1.426	539.66	1.66	000198-55-0	770	770.9932	0.993	715.75	54.25
000575-43-9	539	538.9752	0.024	539.66	0.66	000053-70-3	808	804.6564	3.343	743.09	64.91
000581-40-8	541	537.7338	3.266	539.66	1.34	000191-24-2	815	816.0489	1.048	759.31	55.69
000571-58-4	541	543.1184	2.118	539.66	1.34	000191-07-1	863	861.1951	1.804	802.87	60.13
000571-61-9	542	541.877	0.123	539.66	2.34	000192-65-4	865	864.2395	0.760	786.65	78.35
000575-41-7	544	540.9741	3.025	539.66	4.34	000189-55-9	867	866.7829	0.217	786.65	80.35
002245-38-7	558	557.8031	0.196	555.81	2.19	000191-30-0	868	865.9605	2.039	786.65	81.35
000829-26-5	559	555.3469	3.653	555.81	3.19	000189-64-0	869	871.8613	2.861	786.65	82.35
001430-97-3	591	593.1494	2.149	580.25	10.75	000205-12-9	679	683.403	4.403	643.44	35.56
000832-71-3	625	619.0926	5.907	612.75	12.25	000191-26-4	820	824.9121	4.912	759.31	60.69
002531-84-2	628	619.2055	8.794	612.75	15.25	000224-41-9	804	803.9793	0.020	743.09	60.91
000883-20a-5	628	618.6993	9.300	612.75	15.25	000193-43-1	804	803.0724	0.927	759.31	44.69
000613-12-7	632	630.0154	1.984	612.75	19.25	000193-39-5	807	813.2558	6.255	759.31	47.69
000832-69-9	632	620.7307	11.26	612.75	19.25	000215-58-7	808	805.7525	2.247	743.09	64.91
000610-48-0	636	632.1555	3.844	612.75	23.25	000205-82-3	753	758.2407	5.240	715.75	37.25
001576-67-6	636	636.7962	0.796	624.49	11.51	027208-37-3	712	710.1049	1.895	673.85	38.15
003353-12-6	683	682.7427	0.257	656.45	26.55	000091-20-3	491	500.7226	9.722	504.64	13.64
003442-78-2	683	680.0102	2.989	656.45	26.55	000208-96-8	543	543.8939	0.893	547.85	4.85
002381-21-7	683	684.0123	1.012	656.45	26.55	000083-32-9	552	557.0131	5.013	545.72	6.28
000243-17-4	675	678.0972	3.097	643.44	31.56	000086-73-7	567	574.0128	7.012	565.57	1.43
000238-84-6	680	681.0613	1.061	643.44	36.56	000085-01-8	611	600.9999	10.00	600.31	10.69
000203-12-3	705	700.5363	4.463	688.41	16.59	000120-12-7	613	613.0736	0.073	600.31	12.69
000056-55-3	708	708.4418	0.441	672.19	35.81	000203-64-5	632	628.6046	3.395	616.02	15.98
000217-59-4	712	699.4674	12.53	672.19	39.81	000206-44-0	656	652.5746	3.425	644.85	11.15
000218-01-9	714	705.4811	8.518	672.19	41.81	000129-00-0	666	664.1969	1.803	644.85	21.15
000092-24-0	723	719.5621	3.437	672.19	50.81						

ISSN: 0975-8585

$MAE_{QSPR\ model}$= 3.541 MAE_{EPIWIN} =29.173
$RMSE_{QSPR\ model}$ =4.756 $RMSE_{EPIWIN}$ =37.647

ei (1): Exp. Bp - predicted by model equation, ei (2): Exp. Bp - predicted by EPISUITE

Goodness-of-fit, robustness, and predictive power have been confirmed by the values of R^2, Q^2_{LOO}, R^2_{ext} relatively low values of the errors, $RMSE_{tr}$, $RMSE_{cv}$, $RMSE_{pr}$. Moreover, the scatter plot of the predicted boiling point versus experimental shows a visual correlation between observed and predicted boiling point values for the training (tr) and predicted (pr) confirmed the good quality of the model (fig.1). Since the error values ($RMSE_{tr}$, $RMSE_{pr}$) were close and there were no significantly large residual values for the validation set displayed in fig.1, we can conclude that the model has not over fitted. This means that the model predict correctly not only for the training compounds but also for other (external) compounds.

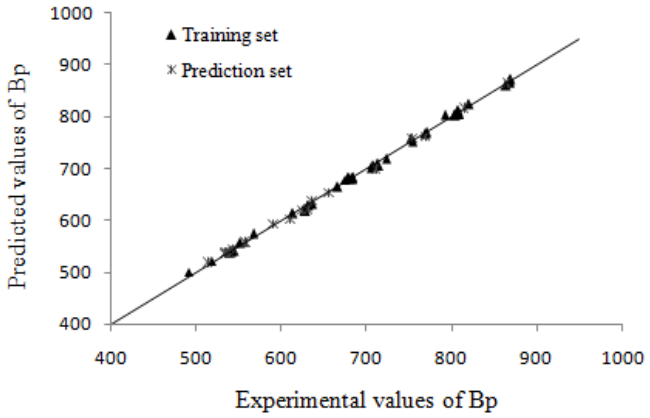

Figure 1: Scatter plot of predicted versus experimental Bp for the training and predicted sets.

Analyzing the model applicability domain is another stage of validation. So called Williams plot (fig. 2) presents the relationship between leverage values (expressing similarity of a given compound to training set and standardized residuals). Analyzing the plot, all residuals were located within the range of ±3 SD (horizontal lines), there were not outlying predictions observed, and there is no structural influential compound both for training or prediction sets (vertical dashed line) represent the formal leverage (similarity) threshold value h^* equal to 0.214), which means that the model has a good external predictivity. Due to its high predictive ability, the proposed model could be used to screen existing databases or virtual chemical structures to identify boiling point of PAHs. In this case, the applicability domain will serve as a valuable tool to filter out "dissimilar" chemical structures.

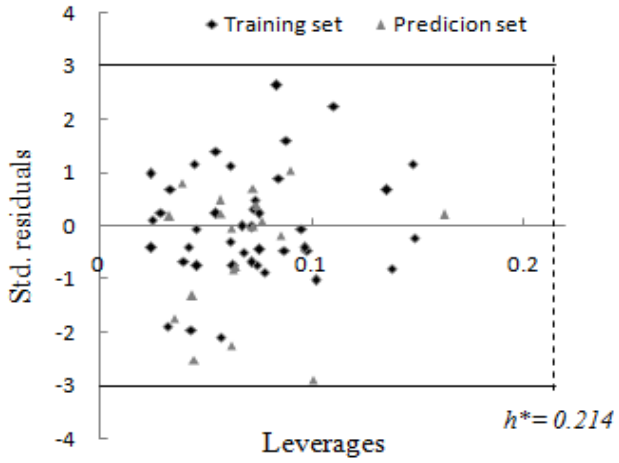

Fig 2: Williams plot: standardized residuals versus leverages. Solid lines indicate ±3SD units, dash line indicate the threshold value h^*= 0.214.

ISSN: 0975-8585

The results (R^2ys and Q^2ys versus Kxy [35]) are plotted in (fig.3), where Kxy is the total correlation in the model variables (y included). The lower values of R^2ys (=0.0492) and Q^2ys (= -1.1309) indicate the good results of the original model are not due to chance correlation or structural dependency of the training set.

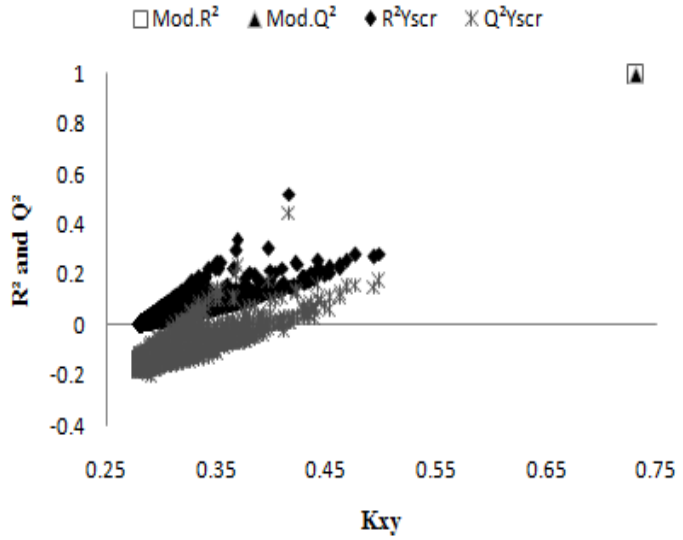

Figure 3: Obtained R^2 and Q^2 using permuted response data versus Kxy.

In order to verify, whether all chemicals from the prediction set (chemicals, for which experimentally values of boiling point have been unviable) are inside the model domain, we applied insubria graph. The graph (fig.4) plots the leverage for prediction set versus predicted values. With insubria graph, we defined the reliable prediction zone of the model based on structural similarity to the training compounds (leverage value) and the predicted value of Bp. We supposed that the predicted results are reliable, if both condition: $h_i < h^*$ and $Y_{min} < Y_{pred} < Y_{max}$ (Y_{min} and Y_{max} are the minimal and the maximal value of Bp in the training set). We found that all compounds from the prediction set were located within the model's applicability domain. This demonstrates that the model obtained in this work has a high applicability to new PAHs, and we can apply in order to screen and prioritize them for future experiments or for filling the data gap.

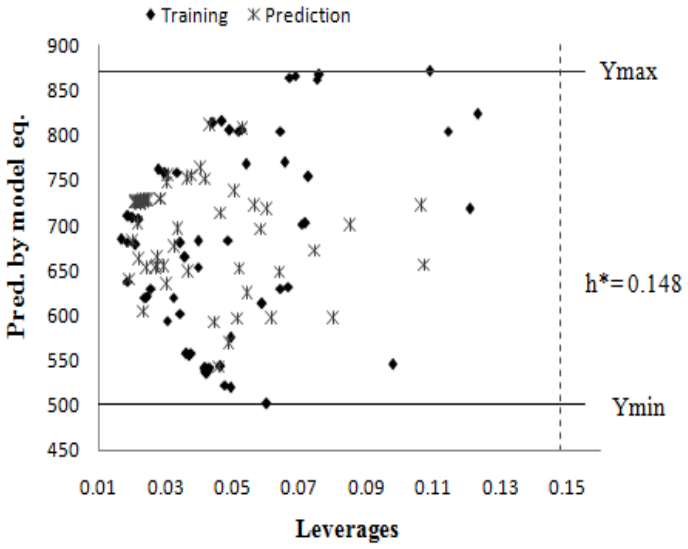

Figure 4: Insubria graph (plot of leverages values versus predicted values for the whole set of PAHs)

ISSN: 0975-8585

Descriptor contribution analysis and interpretation

The relative model contributions of the two descriptors were determined and plotted in (fig. 5). The significance of the descriptors involved in the model decreases in the following order: EPS0 (edge connectivity indices order 0) (91.56 %) >HOMO (higher occupied molecular orbitals) (8.43 %). It should be noted that the difference in the descriptor contribution in the model is significant, we found that edge connectivity indices descriptor itself yielded a one-variable equation **Bp= 252.530+35.759EPS0** with R^2 = 99.71% and standard deviation s = 7.1002 K for n_{tr} = 42. This means that the edge connectivity indices descriptor used is an important descriptor for the influence of molecular structure on boiling point behavior for PAHs. However, a major drawback of edge connectivity indices is its degeneracy, i.e., isomers obtains identical numerical values. Hence, the model developed only by employing edge connectivity indices, is not accurate enough for PAHs. To improve the description a second regressor, HOMO, was added as shown above in Eq. (7).

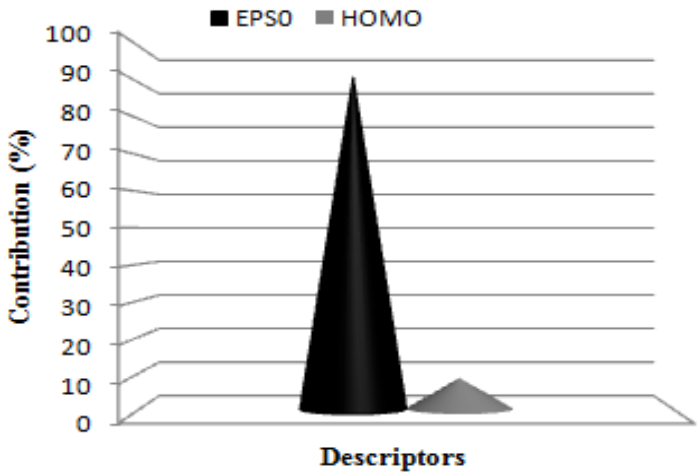

Figure 5: Relative contribution of selected descriptor to the MLR model

According to White [14] the properties of PAH are a direct function of their size and topology. Size is a function of the number of π electrons, while topology is related to whether the ring systems are kata-annellated or peri-condensed. Topology is also a function of linear and angular ring annelation. The PAHs size and topology affect the energy of the highest occupied molecular orbital (HOMO), which in turn is a reasonable predictor of their properties.

The importance of EPS0 [36]: on the boiling point values is apparent, since the EPS0 descriptor explains 91.568 % of the contributions (8.432 % of HOMO). The EPS0 descriptor is highly correlated with the experimental boiling point values (R = 0.997). The positive coefficient of EPS0 indicates that the PAHs with larger values for this descriptor would have larger Boiling Points values. Thus, this descriptor could be an indicator for the PAHs that have a big Bp value. This kind of molecular descriptor is defined as follows:

$$\varepsilon(G) = \sum_S [\delta(e_i).\delta(e_j)]_\delta^{-1/2} \quad (8)$$

Where $\delta(e_i)$ is the degree of the edge (e_i), and the summation is carried out over all pairs of adjacent edges in the graph. Since the number of connections is sensitive to different features of molecular structure such as size, branching, cyclicity, and multiple bonds; this molecular parameter, EPS0, is known as molecular complexity index [37]. The more complicated graphs show the larger values of EPS0 and explain why the boiling points are large.

Comparisons to literature models

The comparison between the most important published models [5,6,14,15] for boiling point prediction of PAHs, can be achieved using the table 2, It is easy possible to verify if one model is better than other, taking into account the size of the studied data set, the statistical parameters taken into considerations and the complexity of model (i.e. number of descriptors involved and the statistical modeling method), the model developed in this work includes more compounds and fewer descriptors, in addition, our model(eq.7) was evaluated by using different statistical parameters compared with other models in the literature [5,6,14,15], the present MLR model shows better statistical quality and performance than previous works.

Table2: Comparison between previous works and this work for the boiling point

Works	N	Size of model	Type of descriptors	R^2 (%)	R	Q^2_{LOO} (%)	$RMSE_{tr}$ $RMSE_{cv}$ ($RMSE_{pr}$)	S	Q^2_{F1}(%) Q^2_{F2}(%) Q^2_{F3}(%)
C.M.White[14]	47	1	The first-order valence molecular connectivity ^1X_v,	-	0.994	-	-	8.59	-
Todeschini R, et al[6]	53	4	WHIM descriptors	95.9	-	95.0	(17.4)	-	-
M. Márcia, and C. Ferreira[5]	23	3	EA, Xv, Log W	-	0.999	-	-	6.68	-
		3	EA, Xe, Log W	-	0.999	-	-	5.52	-
		4	EA, Xe, Xv, Log W	-	0.999	-	-	5.70	-
		4	EA, Xe, SArea, Log W	-	0.999	-	-	4.40	-
Ferreira et al[15]	36	3	Volume (V), molecular weight (MW) and Randic connectivity index(R)	PLS model: 99.5	-	99.42	(7.756)	-	-
				PCR Model: 99.5	-	99.38	(8.474)	-	-
This work	61	2	EPS0 HOMO	99.83	0.998	99.80	4.420 4.770 (5.425)	4.587	99.79 99.72 99.74

Table.1 compare calculated values of Bp using our model and that calculated by EPIWIN model [38], giving the mean absolute error (MAE) and the root mean square error (RMSE) of the two models. The comparison in both cases is in favor of our QSPR model, due $MAE_{QSPR\ model}$ (3.541) << MAE_{EPIWIN} (29.173) and $RMSE_{QSPR\ model}$ (4.756) << $RMSE_{EPIWIN}$ (37.647). This allows us to say that our model is a specialized model for the calculation of PAHs boiling point.

CONCLUSION

In this study, boiling point of 61 polycyclic aromatic hydrocarbons was correlated with their molecular structure by QSPR technique using QSARINS software. A tow dimensional model was calculated by all subset based multiple linear regression, from the training sets obtained with different splitting procedures. The very high values of statistical parameters, related both to internal and external predictivity (Q^2_{LOO}, Q^2_{LMO}, Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, CCC_{ext}) proved that the proposed model possesses good predictive ability and robustness, and thus it can be used to estimate the boiling point for PAHs without experimental data available in the literature. The validity of the model predictions is further guaranteed by the verification on 57 PAHs without experimental values, considering those belonging to the applicability domain with the leverage approach. The QSAR model presented in our work showed better statistical parameter values and better predictability results compared with those obtained previously by different authors.

ACKNOWLEDGMENTS

The authors wish to thank and are grateful to Professor Paola Gramatica and Stefano Cassani for their precious help in use of QSARINS software.

ISSN: 0975-8585

REFERENCES

[1] Armstrong G, Hutchinson E, Unwin J, Fletcher T. Environ. Health. Perspect 2004; 112: 970-978.
[2] Gramatica P, Papa E, Marrocchi A, Minuti L, Taticchib A, Ecotox Environ Safe 2007; 66 : 353–361.
[3] Tham YW, Sakugawa H. Bull. Environ. Contam. Toxicol 2007; 79: 670-673.
[4] Kuo C-Y, Cheng Y-W, Chen Y-W, Lee H. Environ. Res. Sec A 1998 ; 78 : 43-49
[5] Ferreira M M C. Chemosphere 2001; 44: 125-146.
[6] Todeschini R, Gramatica P, Provenazi R, Marengo E. Chemometr Intell Lab 1995; 27: 221-229.
[7] Gharagheizi F, Mirkhani S A, Ilani-Kashkouli P, Mohammadi A H, Ramjugernath D. Richon D. Fluid Phase Equilibr 2013; 354: 250–258.
[8] Katritzky A R, Mu L, Lobanov V S. J Phys Chem 1996; 100: 10400–10407.
[9] Cao D, Liang Y, Xu Q, Yun Y. Li H. J Comput Aided Mol. Des 2011; 25: 67–80.
[10] Fisher, C. H. Chem. Eng 1989,96, 157.
[11] Satyanarayana K, Kakati M C. Fire Mater 1991; 15: 97–100.
[12] Yi-min D, Zhi-ping Z, Zhong C, Yue-fei Z, Ju-lan Z, Xun L. J Mol Graphics Modell 2013 44: 113–119.
[13] Hansh C, Maloney PP, Fujita T, Muir RM. 1962. Correlation of biological activity of phenoxyatic acids with Hammett substituent constants and partition coefficient s. Nature 194: 178-180.
[14] White, C.M. J Chem Eng Data 1986; 31: 198–203.
[15] Ribeiro F A L, Ferreira M M C. J Mol Struc-Theochem 2003; 663 : 109–126.
[16] Bjorseth, A., (Ed.), Handbook of Polycyclic Aromatic Hydrocarbons; Marcel Dekker: New York, 1983; Appendix.
[17] Karcher W, Fordham R J, Dubois J J, Glaude P G J M, Lighart, J A M. Spectral Atlas of Polycyclic Aromatic Compounds, Kluwer Academic Publishers, Dordrecht, 1988.
[18] HyperChem 6.03 Package. Hypercube, Inc., Gainesville, Florida, USA, 1999; software available at: http://www.hyper.com.
[19] Talete Srl. Dragon for windows (Software for Molecular Descriptor Calculation) Version 5.5 Milano, Italy, 2007; software available at: http://www.talete.mi.it.
[20] Jackson J E, A User's Guide to Principal Component. Wiley, New York, United States. 1991.
[21] Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S. QSARINS, Software for the Development and validation of QSAR MLR Models, available on request in http://www.qsar.it.
[22] Gramatica P, Chirico N, Papa E, Kovarich S, Cassani, S. J Comput Chem Software news and updates 2013; 34: 2121–2132.
[23] OECD. Guidance Document on the Validation of (Quantitative) Structure–Activity Relationships [(Q)SAR] Models. Organisation for Economic Co-Operation and Development, Paris, France, 2007.
[24] Tropsha A, Gramatica P, Gombar V K. QSAR Comb Sci 2003; 22: 70–77.
[25] Gramatica P. Mol Inf 2014; 33: 311–314.
[26] Schüürmann G, Ebert R, Chen J, Wang B, Kühne R. J Chem Inf Model 2008 ; 48: 2140–2145.
[27] Consonni, V, Ballabio, D, Todeschini R. J Chem Inf Model 2009 ; 49: 1669−1678.
[28] Consonni V, Ballabio D, Todeschini R. J Chemom 2010 ; 24: 194−201.
[29] Lin L I Biometrics 1989; 45: 255–268.
[30] Chirico N, Gramatica P. J Chem Inf Model 2011; 51: 2320−2335.
[31] 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 GY, Perkins R, Roberts D W, Schultz T W, Stanton D T, van de Sandt J J M, Tong W, Veith G, Yang C. ATLA Altern Lab Anim 2005; 33:155−173.
[32] Chirico N, Gramatica P. J Chem Inf Model 2012; 52: 2044–2058.
[33] Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T.D., McDowell, R.M. and Gramatica, P. Environ Health Persp 2003;111: 1361–1375.
[34] Gramatica P, Cassani S, Roy P P, Kovarich S, Yap C W, Papa E. Mol Inf 2012; 31: 817 – 835.
[35] Todeschini R, Consonni V, Maiocchi A. Chemometr Intell Lab 1999; 46: 13–29.
[36] Estrada E, Guevara N, Gutman I. J Chem Inform Comput Sci 1998; 38: 428–431.
[37] Bagheri M, Borhani T N G, Zahedi G. Fluid Phase Equilibr 2013; 337: 183–190.
[38] https://www.epa.gov/tsca-screening-tools/epi-suite^{tm}-estimation-program-interface