DOI: http://dx.doi.org/10.32571/ijct.1385432 **E-ISSN: 2602-277X**

International Journal of Chemistry and Technology

http://dergipark.org.tr/ijct

Research Article

QSPR-based prediction model for the melting point of polycyclic aromatic hydrocarbons using MLR and ANN methods

Bouarra NABIL[1*], Kherouf SOUMAYA[2], Messadi DJELLOUL[3]

[1]*Scientific and Technical Research Center in Physico-Chemical Analysis, BP 384-Bou-Ismail, Tipaza, Algeria, 42004, Tipaza, Algeria*

[2]*Faculty of technology, Badji Mokhtar University, Annaba, Algeria*

[3] *Department of Chemistry, Badji Mokhtar University, Annaba, Algeria*

Received: 3 November 2023; Revised: 28 August 2024; Accepted: 4 September2024

*****Corresponding author e-mail: bouarranabil@yahoo.com

Citation: Nabil, B.; Soumaya, K.; Djelloul, M. *Int. J. Chem. Technol.* **2024**, 8(2), 128-136.

ABSTRACT

The melting point is an important property that helps generate specific compounds with desired thermos-physical properties. Much work has been done applying quantitative structure-property relationships to improve the melting-point correlations, but they are unreliable. This gap might come from the melting point's sensitivity for small molecular variations and descriptors, which currently do not fully consider all factors determining melting behavior. In this work, we provide a QSPR model for predicting the melting point of a heterogeneous polycyclic aromatic hydrocarbons dataset. The model was generated using a robust hybrid linear approach (Genetic Algorithm-Multiple Linear Regression) and a nonlinear approach named Artificial Neural Network (ANN). Three descriptors were chosen to explain the influence of molecular weight and symmetry on melting point. The resulting QSPR model can model melting-point behavior with an RMSE of 34.88K, a coefficient correlation value of R^2=0.887, and a prediction coefficient of Q^2_{LOO}= 0.863. This study reveals that the results produced by MLR were appropriate and served to predict melting points. However, compared to the results obtained by the ANN model, we conclude that the latter is more effective and better than the MLR model. Based on the results, our suggested model may be effective in predicting melting points, and the selected descriptors play essential roles in determining melting points.

Keywords: QSPR, Melting point, Molecular descriptors, Genetic algorithm, ANN.

1. INTRODUCTION

Polycyclic aromatic hydrocarbons (PAHs) are a vast family of neutral and stable organic compounds consisting solely of carbon (C) and hydrogen (H) atoms. This family of compounds contains 2-6 fused aromatic rings.[1] PAHs are widespread, persistent, and toxic molecules found in our environment. The major source of PAHs is the pyrolysis or carbonization of organic compounds, including coal, oil, and wood.[2] PAHs are versatile industrial chemical compounds; some of the intermediate products that PAHs contribute include pharmaceuticals, photographic products, lubricating materials, agricultural commodities, and thermosetting plastics.[3] Due to their widespread distribution, comprehending the physicochemical properties of PAHs is important for assessing their environmental impact, health risks, and practical applications. The melting point is among these properties.[4]

DOI: http://dx.doi.org/10.32571/ijct.1385432 E-ISSN: 2602-277X

A compound's melting point (MP) is one of the most studied aspects of chemistry since it is so helpful in determining the compound's identity.[4] The identification and analysis of organic pollutants have significantly advanced[5,6], but more work is still needed, notably in discovering specific Melting Points for these chemicals. Melting point is also one of the most influential factors in environmental transport and destiny processes. Determining a substance's MP allows the analysis of various contaminants. MP, for example, is directly related to solubility; hence, it is crucial for environmental research. Computational modeling can be used to determine the factors responsible for the distribution of chemical contaminants in the environment. Computational modeling can be used to determine the factors responsible for the distribution of chemical contaminants in the environment, obviating the need for costly and time-consuming empirical studies[7]. Therefore, unconventional methods to understand the environmental behavior of PAHs are needed.

An alternative approach to determine the physical-chemical properties of chemicals, such as the melting point of PAHs, is quantitative structure-property relationships (QSPR), which utilizes molecular descriptors derived from the compound's structure to adjust experimental data. QSPR is based on the concept that changes in the numerical values of structural features, known as molecular descriptors, can be associated with alterations in the compound's behavior, which are reflected in its physicochemical properties as measured in experiments.[8,9] The benefit of this method is that it relies on chemical structure knowledge rather than experimental qualities. Once a correlation is developed and verified, it can predict a compound's properties. It has been shown that the QSPR method can accurately predict a wide range of physical and chemical characteristics of molecules.[10] The literature presents several QSPR models for melting point prediction.[11-16] However, their respective predictive abilities vary greatly. While relatively accurate models have been generated for tiny subgroups of compounds, those built from training sets with considerable structural variation tend to perform poorly overall.

The objective of this work is to build an accurate quantitative structure-property relationship (QSPR) model for predicting the melting points of a diverse set of polycyclic aromatic hydrocarbons (PAHs). By employing both a linear method (Multiple Linear Regression) and a nonlinear method (Artificial Neural Network), the study aims to check the predictive power of these models and the impact of key molecular descriptors, such as molecular weight and symmetry, on melting point behavior and enhance the accuracy and reliability of melting point predictions while identifying the factors that most significantly influence the melting behavior of PAHs by giving a mechanistic interpretation of the developed model.

2. MATERIALS AND METHODS

2.1. Dataset and Descriptor Calculation

In the present study, the experimental Mp data listed in Table 1 were collected from the work of Roberto Todeschini *et al*.[17] Table 1 provides a complete list of chemicals and their experimental and predicted melting points using MLR and ANN techniques. The reported Mp values ranged from 251 to 756 K. To validate the QSPR model, the CADEX algorithm[18] divides the dataset into training and prediction subsets. The algorithm ensures that both subsets represent the entire dataset to guarantee external validation significance. The training set comprises 55 compounds, whereas the prediction set comprises 22.

Molecular descriptors were calculated using DRAGON software version 5.5. [19] All families of descriptors (0D-3D) were included in the study. In order to ensure the reliability and accuracy of the modelling process, specific measures were taken during the descriptor selection stage. Firstly, descriptors that contained missing values were excluded from consideration. Then, descriptors that exhibited high pairwise correlation (>95%) or were nearly constant (>80%) were eliminated to prevent redundant information and binary collinearity issues. This pretreatment was done because correlated descriptors can lead to mathematical problems during the modeling phase, and descriptors with limited relevance to most molecules may not help make predictions.[20] Afterward, the remaining descriptors were entered into the QSARINS software, version 2.2.4 [21,22] for further analysis and modeling. Taking these precautions will make the resulting model more reliable and accurate for predicting the melting point.

2.2. Model development

Multiple Linear Regression (MLR)[23] was employed as the modeling strategy in this investigation, with the following equation:

$$\hat{Y} = b_0 + \sum_{j=1}^{J}(b_j . X_j) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_JX_J + \varepsilon \quad (1)$$

The Ordinary Least Square (OLS) approach implemented in QSARINS software[21] minimises the sum of squares between the experimental endpoint and the calculated value. After preparing and dividing the dataset, the descriptor selection process was initiated. In this regard, the training set was utilised to calculate all potential combinations of up to three descriptors. This operation guaranteed that all feasible low-dimensional models were generated before increasing the variables for best results. Based on this foundation, a genetic

DOI: http://dx.doi.org/10.32571/ijct.1385432 E-ISSN: 2602-277X

algorithm was used to find the optimum configuration of the model across a more significant collection of features. Model quality was maximized by the method using Q^2_{LOO}, a fitness function defined by equation (2).

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

Where, $\hat{y}_{i/i}$ is the value of Mp predicted by the generated model according to the LOO method, y_i as the experimental melting point and $\bar{y}$ as the mean of the experimental melting point, n is the total compounds in the training set.

Only models with descriptors that had a *p*-value ≤ 0.05 were retained. The settings of the GA used to provide the best modeling results while consuming as little computational resources as possible were a population size of 500, a generation per size of 500, and a mutation rate of 80%.

2.3. Model Validation

Validation is a crucial stage in QSPR modeling. Therefore, it is essential to evaluate and validate the resulting model extensively. The QSARINS provides several resources for verifying that a model satisfies the OECD's requirements for the creation, validation, acceptance, and use of QSAR models, which boosts confidence in the accuracy of the data predicted by the model. The QSPR-MLR model should satisfy the following conditions to be valid according to these guidelines: A clearly defined property, a straightforward method, a defined applicability domain, adequate measurements of goodness-of-fit, robustness, and predictivity; and a mechanistic description, if possible.[24] We employ cross-validation methods in this work as a form of internal validation. As a first step, we conducted Leave-One-Out (LOO) (Eq. (2)) approach, as disrupting a single molecule in a minimal database provides us with criteria for its resilience. The model's performance when more chemicals are left out was further studied using a Leave-Many-Out (LMO) approach.

External validation was performed to demonstrate the predictability of the model. The prediction set, which was not included in creating the original model, is utilised for this purpose. In order to evaluate the predictive ability of the developed model, numerous statistical parameters were calculated, including (Q^2_{F1}, Q^2_{F2}, Q^2_{F3}, R^2_0, $R^{2\prime}_0$, and CCC_{ext}) (Eq. (3-12)). For more information, these statistical parameters were provided in our previous papers and explained in detail[25-27].

$$Q^2_{F1} = 1 - \frac{PRESS_{EXT}}{SS_{EXT}(\bar{y}_{TR})} \quad (3)$$

$$PRESS_{EXT} = \sum (y_i - \hat{y}_i)^2 \quad (4)$$

$$Q^2_{F2} = 1 - \frac{PRESS_{EXT}}{SS_{EXT}(\bar{y}_{EXT})} \quad (5)$$

$$SS_{EXT}(\bar{y}_{EXT}) = \sum (y_i - \bar{y}_{EXT})^2 \quad (6)$$

$$Q^2_{F3} = 1 - \frac{\left(\frac{PRESS_{EXT}}{n_{EXT}}\right)}{\left(\frac{TSS}{n_{TR}}\right)} \quad (7)$$

$$CCCext = \frac{2\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sum_{i=1}^{n}(y_i - \bar{y})^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 + n(\bar{y} - \bar{\hat{y}})^2} \quad (8)$$

$$R^2_0 = 1 - \frac{\sum_{i=1}^{n}(y_i - k \times \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (9)$$

$$R^{'2}_0 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - k' \times y_i)^2}{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2} \quad (10)$$

$$k = \frac{\sum_{i=1}^{n} y_i \hat{y}_i}{\sum_{i=1}^{n} \hat{y}_i^{\,2}} \quad (11)$$

$$k' = \frac{\sum_{i=1}^{n} y_i \hat{y}_i}{\sum_{i=1}^{n} y_i^{\,2}} \quad (12)$$

Where: $\bar{y}_{TR}$ = average of training observed responses

$\bar{y}_{TR}$ = average of external observed responses

n_{EXT} = number of external objects

n_{TR} = number of training objects

$\bar{\hat{y}}$ is the average of all $\hat{y}_i$

To rule out the possibility of random correlation and to ensure stability and reliability via permutation testing, $Y_{scrambling}$ was used, and new models were rebuilt for randomly reordered data. Since the suggested models rely on a correlation between structure and response, randomized responses should provide models with much lower Q^2 values.[25] The mean values of $R^2_{Yscrambling}$ and $Q^2_{Yscrambling}$ were reported following a data scrambling procedure with a maximum of 200 iterations.

2.4. ANN modeling

Artificial Neural Network (ANN)[28] is an excellent way to find nonlinear correlations because it can make

DOI: http://dx.doi.org/10.32571/ijct.1385432 E-ISSN: 2602-277X

models without requiring a precise analytical method. Many studies have investigated the use of ANN in QSPR research.[28, 29-32]

In this study we used descriptors obtained from MLR as input into a three-layer feed-forward ANN with a back-propagation learning method to predict the melting point of 77 PAHs.[33] The training process involved adjusting the number of hidden neurons through a trial and error technique. The ANN used one output neuron to represent the experimental melting point.

3. RESULTS AND DISCUSSION

3.1. MLR model

Exploring the most effective combinations of molecular descriptors that strongly correlate with the response variable (Mp) resulted in the construction of many models. Therefore, a study of the model's parameters was conducted while considering the concept of parsimony[34] (explaining the highest information with the least number of descriptors). Based on it, a QSPR-MLR model with four variables was created. The best model's equation and statistical parameters are as follows:

$$\boldsymbol{Mp(K)} = -567 + 122\,\boldsymbol{AMW} + 2{,}99\,\boldsymbol{TIC1} - 342\,\boldsymbol{SIC4} + 281\,\boldsymbol{P1m} \quad (13)$$

$\boldsymbol{N_{tr}}$ = 55, $\boldsymbol{R^2}$= 88,72%, $\boldsymbol{Q^2_{LOO}}$ = 86,38%, $\boldsymbol{R^2_{ext}}$ = 82,49%, $\boldsymbol{Q^2_{LMO30\%}}$ = 85,90%, $\boldsymbol{Q^2_{F1}}$ = 89,84%, $\boldsymbol{Q^2_{F2}}$= 77,92%, $\boldsymbol{Q^2_{F3}}$ = 88,43%, $\boldsymbol{CCC_{ext}}$ = 86,64%, $\boldsymbol{RMSE_{tr}}$ =34,88, $\boldsymbol{RMSE_{Val}}$ = 35,32, S = 36,58.

The descriptors included in the model were designed as follows: the AMW represents the average molecular weight, the TIC1 and SIC4 the type of Information indices-; these descriptors represent the Total Information Content index (neighborhood symmetry of 1-order), and the Structural Information Content index (neighborhood symmetry of 4-order) respectively, and P1m for WHIM descriptors- 1st component shape directional WHIM index / weighted by mass.[35-37] The numerical values for the four descriptors used in the final GA-MLR equation (Eq. (13)) are listed in Table 1.

Table 1. Experimental and predicted melting point by MLR and ANN.

Name	Mp (k) exp	Predicted by MLR	Predicted by ANN	Name	Mp (k) exp	Predicted by MLR	Predicted by ANN
1-methylnaphthalene	251	261.500	281.602	6-methylchrysene	530	451.123	497.805
1-ethylnaphthalene	259	248.379	275.695	Dibenzo[def,mno]chrysene	534	525.748	531.358
2,3,5-trimethylnaphthalene	298	310.052	297.063	Dibenz[a,h]anthracene	543	529.866	541.921
1-phenylnaphthalene	318	355.686	318.100	Pentacene	544	590.803	553.300
9-methylfluorene	320	386.263	343.257	Perylene	551	511.603	513.312
4-methylphenanthrene	323	361.380	323.160	Benzo[ghi]perylene	556	519.535	567.616
1,5-dimethylnaphthalene	353	335.152	324.430	Benzo[b]chrysene	567	507.705	552.000
1-methylfluorene	360	365.315	363.547	Phenalene	358	307.466	364.225
9-methylphenanthrene	364	364.367	315.191	2,6-dimethylanthracene	523	479.026	518.558
Acenaphthylene	366	366.500	368.434	Dibenzo[a,i]anthracene	537	517.676	550.481
2,7-dimethylnaphthalene	370	374.163	355.830	Hexaphene	581	575.254	579.994
Azulene	373	309.701	357.000	Coronene	633	632.265	636.199
2-phenylnaphthalene	377	392.271	418.956	Indene	271	260.737	275.157
2,6-dimethylnaphthalene	383	389.644	391.497	Ovalene	746	687.364	745.564
Fluoranthene	384	426.377	399.785	Quaterrylene	756	773.746	754.259
4H-cyclopenta[def] phenanthrene	389	442.627	395.330	Dibenzo[a,e]pyrene	507	470.443	492.133
Fluorene	390	413.629	389.196	1,7-dimethylnaphthalene*	259	287.000	281.709
2-methylpyrene	417	459.662	445.047	1,3,7-trimethylnaphthalene*	287	311.387	301.515
4-methylpyrene	421	395.527	402.579	2-ethylnaphthalene*	266	327.560	325.722
Benzo[ghi]fluoranthene	422	479.084	449.675	1,2-dimethylnaphthalene*	269	295.989	289.426
Pyrene	429	468.186	424.528	2-methylphenanthrene*	329	387.713	407.037
1-methylchrysene	434	451.460	476.668	3-methylphenanthrene*	338	375.144	357.132
Benz[a]anthracene	435	450.735	443.544	1-methylpyrene*	343	418.840	384.976
Indeno[1,2,3-cd]pyrene	436	489.075	472.464	Naphthalene*	354	388.671	378.065
Benzo[j]fluoranthene	439	454.561	423.649	1-methylanthracene*	359	392.409	398.433
Benzo[b]fluoranthene	441	468.656	437.920	Acenaphthene*	369	338.456	362.181
Benzo[a]pyrene	450	460.460	443.864	2,3,6-trimethylnaphthalene*	374	347.672	347.286

DOI: http://dx.doi.org/10.32571/ijct.1385432 E-ISSN: 2602-277X

Benzo[e]pyrene	452	449.209	469.414	Phenanthrene*	374	409.561	399.207
3-methylcholanthrene	453	505.530	459.006	2-methylfluorene*	377	380.572	354.213
9,10-dimethylanthracene	456	475.786	471.462	3,6-dimethylphenanthrene*	414	432.118	418.587
Benzo[a]fluorene	463	440.433	456.185	2,7-dimethylanthracene*	514	471.494	513.516
Triphenylene	472	474.468	467.344	Pentaphene*	536	527.848	532.542
Dibenz[a,c]anthracene	478	495.968	496.672	4-methylfluorene*	344	353.237	343.081
2-methylanthracene	482	416.050	464.652	3,4-benzofluorene*	398	414.310	436.605
Benzo[b]fluorene	482	457.084	475.477	2-methylnaphthalene*	308	310.028	295.382
Anthracene	489	470.404	504.474	1-methylphenanthrene*	396	374.231	366.675
Aenzo[k]fluoranthene	490	498.875	497.879	2,3-dimethylanthracene*	525	473.408	520.169
Chrysene	529	473.547	471.783	3-methylfluorene*	361	368.775	359.065
Naphthacene	530	536.877	542.158				

Compounds with asterisks (*) are prediction set.

Analyzing the data in Table 1, it is evident that there are certain PAH compounds for which the melting points predicted by both MLR and ANN models are quite accurate and close to the actual values. In contrast, others show a wide deviation from the real values, and these variations can be attributed to several factors. One key consideration is the quality of the experimental melting point data used for training the models. However, if the experimental data is less accurate, it will affect the predictive accuracy of the models. On the other hand, PAHs with available experimental data correspond to more accurate predictions of the QSPR model. Another factor that has to be considered is the molecular size of PAHs and the fact that they are chemically diverse and can have different structures. Significant variations in the melting point tendency can be observed depending on the number and connection position of fused aromatic rings and the type of diverse substituent groups. Thus, PAHs with fewer rings are easier to model and have fewer errors in the predicted values. However, the range of molecular weights for larger or more structurally diverse PAHs is considered in the increased variability to predict the experimental values.

Figure 1 (right) shows the scatter plot of the Predicted versus experimental values of Mp for the training and validation set obtained by MLR modeling. An agreement between the experimental and predicted Mp for each set is observed. Furthermore, the data show a low scattering around the first bisector, suggesting that the predictions are correct and reasonable, that the model yields good performance in both training and validation and that the differences between predicted and actual melting points are fairly small. This is evident considering the conformity of the predictions by the MLR model with the actual observed melting points within the dataset.

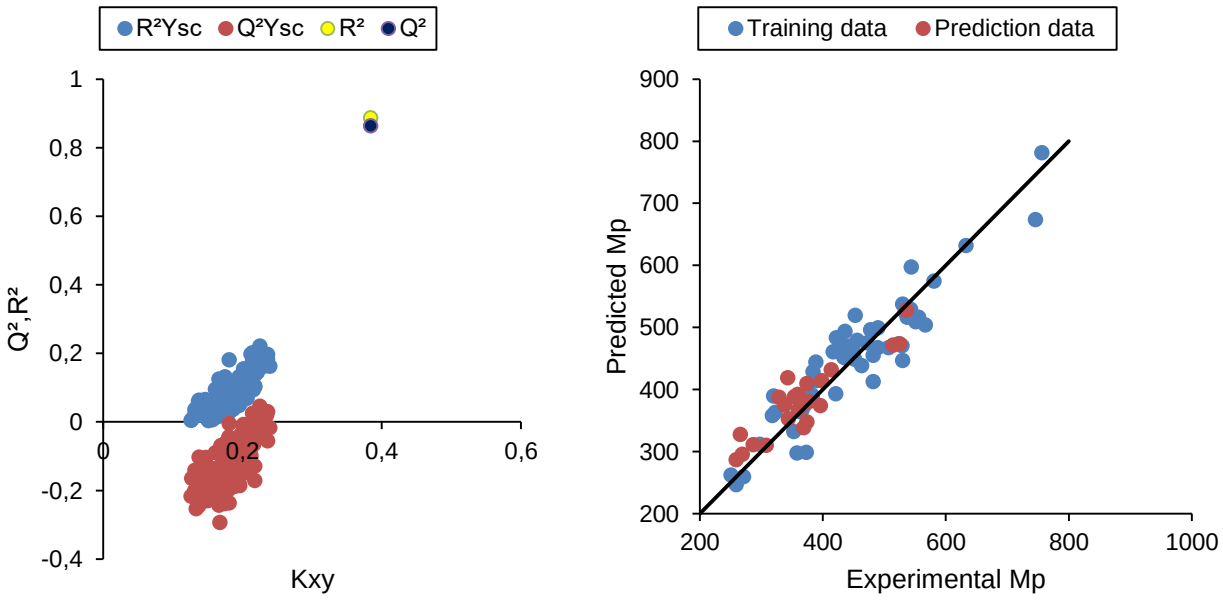

Figure 1. Right) Predicted versus experimental values of Mp; Left) Y-Scramble plot of R^2 and Q^2 vs. Kxy for random models.

DOI: http://dx.doi.org/10.32571/ijct.1385432 E-ISSN: 2602-277X

The final experiment of the internal validation was the Y-scrambling technique[28], as we previously stated; it was done to prove that the model is not the consequence of a fortuitous correlation. Here, the response variable (Mp) was inserted randomly, so there is no association with the descriptors. As a result of this, the model's performance decreases drastically. The R^2 and Q^2 of each iteration and their averages (R^2_{Yscr} and Q^2_{Yscr}) supply the criterion that the model is excellent, as these parameters are ever lower with relation to the values of the model (R^2_{Yscr} = 0.073 and Q^2_{Yscr} = -0.127). The R^2_{Yscr} and Q^2_{Yscr} values versus R^2 and Q^2 of the model are represented in Figure 1 (left). Note that the values of R^2 and Q^2 in the model are distant from the values obtained for those parameters in the Y-scrambling experiment, which suggests that the model is not developed due to a random correlation.

3.2 Correlation matrix

Table 2 presents the intercorrelation coefficient matrix of the molecular descriptors from the MLR model, which showed low Pearson's correlation values between them (all less than 0. 80). This means the descriptors were pretty independent. Also, we calculated the Variable Inflation Factor (VIF) values[29] for three of the descriptors. The VIF values for the four descriptors were all between 1.072-1.195, which is less than five, so the descriptors aren't too correlated, and there's no multicollinearity. Thus, the MLR model made with these four descriptors is a good regression equation with statistical significance and stability.

3.3. Results of the ANN model

The adaptability of the Artificial Neural Network (ANN) method in mathematics makes it a valuable tool for constructing predictive models. One advantage of using ANN is its capability to integrate nonlinear interdependencies between dependent and independent variables without requiring a specific mathematical function. In this investigation, the back-propagation algorithm (BP-ANN)[38] was utilized to create a nonlinear model, with four descriptors derived from the MLR model used as inputs.

Figure 2 displays the statistical parameter values, which vary with the number of neurons in the hidden layer. The RMSE values for the training, validation, and test sets are close to each other and reach their lowest point when five neurons are used in the hidden layer.

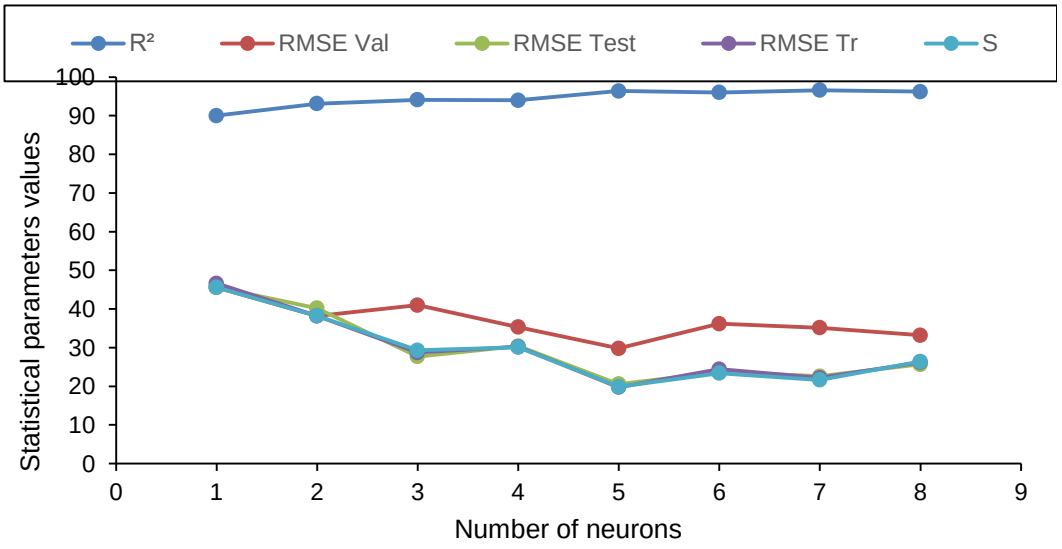

Figure 2. Statistical parameters vs the number of neurons in the hidden layer.

The performance of the ANN models is greatly influenced by the number of neurons in the hidden layer. This study determined that the number of hidden neurons should be at most 8 when working with a training set of 55 samples.[39] After optimizing the network architecture concerning the number of hidden neurons, better results were achieved by using five hidden neurons. Thus, the selected architecture was (4-5-1), yielding the following statistical results for the training set: R^2 = 96.387%, $RMSE_{val}$ = 29.808, $RMSE_{test}$ = 20.559, $RMSE_{tr}$ = 19.742, and s = 19.878.

In order to verify the goodness of fit, the predicted values for melting temperature were plotted against the experimental values. The resulting plot, depicted in Figure 3, exhibited a slight dispersion of points around the first bisector, suggesting that the values were in good agreement with each other. The statistical results were: R^2_{tr} = 0.964, R^2_{val} = 0.87, and R^2_{test} = 0.951.

DOI: http://dx.doi.org/10.32571/ijct.1385432 E-ISSN: 2602-277X

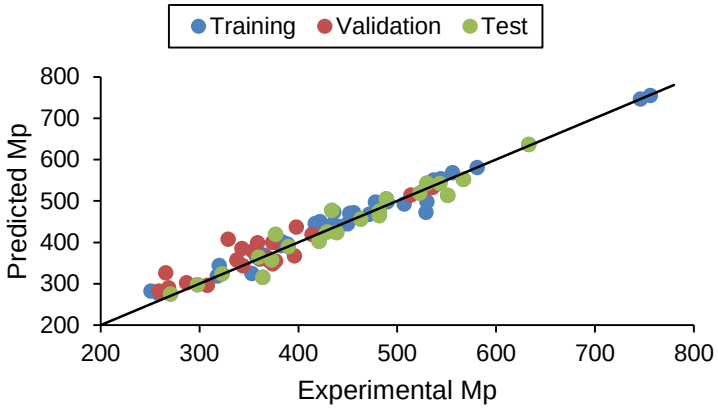

Figure 3. The plot of predicted values for the training, validation and test sets against the experimental values.

Table 1 and Figure 3 present the dataset's prediction results for the ANN model. These results demonstrate some divergence from those obtained via the MLR model, further confirming a nonlinear relationship between structural information and the melting temperature values for compounds. Furthermore, the ANN model proposed in this study demonstrated solid predictive ability under test set conditions, with statistical parameters including:

r^2 = 0.8713 r_0^2 = 0.9765 $r_0'^2$ = 0.9755

$(r^2 - r_0^2)/r^2$ = -0,1208 $(r^2 - r_0'^2)/r^2$ = -0.1196

Q^2_{ext} = 0.8471 0,85 ≤ k = 1.0274 ≤ 1,15 0.85 ≤ k' = 0.9682 ≤1.15

3.4.MLR and ANN Comparison

We performed a comparison between the results obtained by the two methods. Figure 4 establishes that the performance of both approaches is generally good but with an advantage for the nonlinear model. Based on the results obtained for both models, the Artificial Neural Network technique gives better results than MLR.

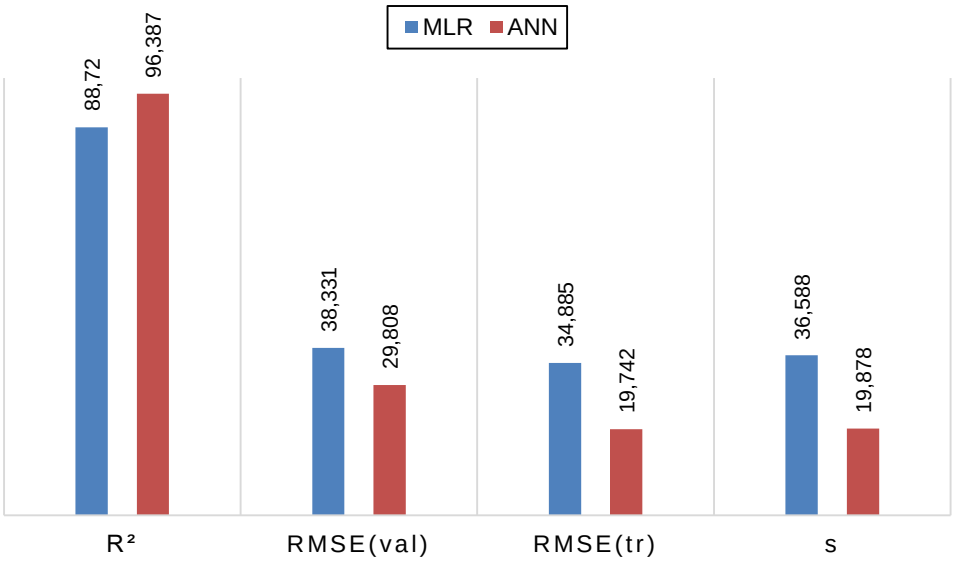

Figure4. Comparison of the performance of MLR and ANN models.

DOI: http://dx.doi.org/10.32571/ijct.1385432 E-ISSN: 2602-277X

4. Mechanistic Interpretation

On the mechanistic interpretations of QSRR, we start with the number and type of molecular descriptors used in the model. To interpret the mechanism, we tried to use the four descriptors included in the developed model. This affirms that the study is in agreement with the OECD's[40] fifth principle.

The Average Molecular Weight (AMW)[37] significantly influences the melting point (Mp) in a favourable manner, as larger molecules generally have greater melting points. The fact that the melting point tends to increase as the molecular weight of the PAH increases is indicated by the positive coefficient of +122 for the AMW descriptor. The increased number of atoms in larger molecules generally results in more substantial Van der Waals forces, which in turn leads to stronger intermolecular interactions. As a result, the melting point is elevated as a result of the increased energy (in the form of heat) necessary to surmount these forces during the phase transition from solid to liquid.[41]

TIC1 (Total Information Content, Neighbourhood Symmetry of 1st-order) and SIC4 (Structural Information Content,[37] Neighbourhood Symmetry of 4th-order): The symmetry of the molecular neighbourhood is reflected in these descriptors at different orders (1st-order and 4th-order, respectively). The melting point is slightly elevated as a result of an increase in the 1st-order neighbourhood symmetry, as indicated by the positive coefficient for TIC1 (+2.99). The degree of symmetry in the immediate molecular environment is quantified by TIC1. The melting point is elevated as a result of the uniformity in intermolecular interactions, which contributes to the lattice's stability and the more regular clustering of molecules in the crystal lattice often resulting from increased symmetry.[41] Increased structural symmetries in the 4th-order neighbourhood is indicative of a decrease in the melting point, as indicated by the substantial negative coefficient for SIC4 (-342). This could imply that high-order symmetry results in a more flexible or less compact molecular packing, which facilitates the molecules' ability to surmount the forces that hold them together in the solid state. This results in a lower melting point, as less thermal energy is required for dissolving.[42]

P1m (WHIM Descriptor: First Component Shape Directional WHIM Index Weighted by Mass)[35-36]. P1m describes the shape and mass distribution of the molecules. It aids in capturing the molecular three-dimensional properties, which influence the melting point. P1m's positive coefficient (+281) implies that as the form directional index grows, so does the melting point. The P1m descriptor represents the shape and mass distribution of the molecule.[43] Higher P1m values often indicate more elongated or anisotropic molecules with a specific mass distribution, which might increase molecular rigidity or influence how molecules pack together in a crystal. This greater rigidity and shape-induced stability would necessitate more energy to break the molecular arrangement, resulting in a higher melting temperature.[44]

5.CONCLUSION

The MLR and ANN Methods (linear and nonlinear) were exploited in this work to develop models of the melting temperature of a series of PAH. Both methods appear helpful, although their comparison is advantageous to ANN. The superiority of the ANN results indicates that the PAH melting temperature has some nonlinear characteristics. The MLR method is suitable for selecting inputs for the ANN modeling and more potent in choosing the critical parameters. The results of this work show that the introduction of neural network improves the quality of melting point prediction.

ACKNOWLEDGEMENTS

The authors are grateful to Prof. Paola Gramatica for the free license of QSARINS. We are thankful to the Algerian Directorate-General for Scientific Research and Technological Development (DGRSDT) for providing financial assistance for this research.

Conflict of interest

Authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

1. Pogorzelec, M.; Piekarska, K. *Sci. Total Environ*. **2018**, 631, 1431-1439.

2.Abdel-Shafy, H. I.; Mansour, M. S. M. *Egypt. J. Petrol*. **2016**, 25, 107-123.

3.Kaminski, N. E.; Faubert Kaplan, B. L.; Holsapple, M. P. Casarett and Doull's Toxicology, the basic science of poisons, C. D. Klaassen (Ed.), Mc-Graw Hill, Inc., New York, 2008.

4.Katritzky, AR.; Maran, U.; Lobanov, VS.; Karelson, M. *J Chem. Inf. Comput. Sci.* **2000**, 40,1–18.

5.Ding, G.; Chen, J.; Qiao, X.; Huang, L.; Lin, J.; Chen, X. *Chemosphere*. **2006**, 62,1057–1063.
6.Xu, HY.; Zou, J.W.; Yu, Q.S.; Wang, Y.H.; Zhang, J.Y.; Jin, H.X. *Chemosphere*. **2007**, 66,1998–2010.

DOI: http://dx.doi.org/10.32571/ijct.1385432 **E-ISSN: 2602-277X**

7.Watkins, M.; Sizochenko, N.; Rasulev, B.; Leszczynski, J. *J. Mol. Model*. **2016**, 22, 1-14.

8.Devillers, J.; Balaban. A.T. Topological Indices and Related Descriptors in QSAR and QSPR, 1st Ed.; Gordon and Breach: Amsterdam, Netherlands, **1999**.

9.Afantitis, A.; Melagraki, G.; Sarimveis, H.; Koutentis, P.A.; Igglessi-Markopoulou, O.; Kollias, G. *Mol. Diversity*. **2010**,14, 225–235.

10.Katritzky, A.R.; Kuanar, M.; Slavov, S.; Hall, C.D.; Karelson, Kahn, M. I.; Dobchev, D.A. *Chem. Rev*. **2010**,110, 5714–5789.

11.Guendouzi, A.; Mekelleche, S.M. *Chem. Phys*. Lipids. **2012**, 165, 1–6.

12.Eike, D.M.; Brennecke, J.F.; Maginn, E.J. *Green. Chem*. **2003**, 5 ,323–328.

13.Karthikeyan, M.; Glen, R.C.; Bender, A. J. *Chem. Inf. Comput. Sci.* **2005**, 45, 581–590.

14.Godavarthy, S.S.; Robinson, R.L.; Gasem, K.A.M. *Ind. Eng. Chem. Res*. **2006**, 45, 5117–5126.

15.Habibi-Yangjeh, A.; Pourbasheer, E.; Danandeh-Jenagharad, M. *Bull. Korean Chem. Soc*. **2008**, 29, 833–841.

16.Deeb, O.; Goodarzi, M.; Alfalah, S.; *Mol. Phys*. **2011**, 109,507–516.

17.Todeschini, R.; Gramatica, P.; Provenzani, R.; Marengo, E.; *Chemometr. Intell. Lab*. **1995**, 27, 221-229.

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

19.Talete Srl. Dragon for Windows (Software for Molecular Descriptor Calculation) Version 5.5 Milano, Italy, **2007**.

20.Gramatica, P. *Comput. Toxicol*. **2013**, 2, 499–526,

21.Gramatica, P.; Chirico, N.; Papa, E.; Cassani, S.; Kovarich, S. J. Comput. Chem. **2013**, 34, 2121–2132.

22.Gramatica, P.; Cassani, S.; Chirico, N.; *J. Comput. Chem*. **2014**, 35, 1036–1044.

23.Katritzky, A.R.; Lobanov, V.S.; Karelson, M. Chem. Soc. Rev. **1995**, 24, 279-287.
24.Worth, A.P.; Bassan, A.; De Bruijn, J.; Gallegos Saliner, A.; Netzeva, T.; Patlewicz, G.; Tsakovska, I.; Eisenreich, S. *SAR. QSAR. Environ. Res*. **2007**, 18, 111-125.

25.Kherouf, S.; Bouarra, N.; Messadi, D. *Int. J. Chem. Technol.* **2019**, 3, 121-128.

26.Bouarra, N.; Nadji, N.; Nouri, L.; Boudjemaa, A.; Bachari, K.; Messadi, D. *J. Serb. Chem. Soc*. **2021**, *86*, 63-75.

27.Bouarra N.; Nadji N.; Kherouf S.; Nouri L.; Boudjemaa A.; Bachari K.; Messadi D. *J. Turk. Chem. Soc. A: Chem*. **2022**, 9, 709-720.

28.Gramatica, P.; Giani, E.; Papa, E. *J. Mol. Graph. Model*. **2007**, 25, 755-766.

29.Bouarra, N.; Nadji, N.; Nouri, L.; Boudjemaa, A.; Bachari, K.; Messadi, D. *Alg. J. Env. Sc. Tech*, **2021**, 7, 2013-2023.

30.Fissa, M. R.; Lahiouel, Y.; Khaouane, L.; Hanini, S. J. *Mol. Graph. Model*. **2019**, 87, 109-120.

31.Quang, N. M.; Mau, T. X.; Nhung, N. T. A.; An, T. N. M.; Van Tat, P. *J. Mol.Struct*. **2019**, 1195, 95-109.

32.Moshayedi, S.; Shafiei, F.; Momeni Isfahani, T. *Int. J. Quantum. Chem*. **2022**, *122*, e27003.

33.Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. *Nature*, **1986**, 323, 533-536.

34.Carbó-Dorca, R. ;Gallegos, A., & Sánchez, Á. J. *J. comp. Chem*. **2009**,30, 1146-1159.

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

36.Todeschini, R.; Gramatica, P. *Quant. Struct.-Act. Relat*. **1997**, 16, 113-119.

37.Todeschini, R.; Consonni, V. Molecular Descriptors for Chemoinformatics, Wiley-VCH, New York, **2009**.

38.Bocaz-Beneventi, G.; Latorre, R.; Farková, M.; Havel, J. *Anal. Chim. Acta*. **2002**, 452, 47–63.

39.Sheela, K. G.; Deepa, S. N. *Math.prob.eng*. **2013**, 2013,1-11.

40.OECD. Principles for the validation, for regulatory purposes, of (quantitative) structure activity relationship models. In: 37th joint meeting of the chemicals committee and working party on chemicals, pesticides and biotechnology. Paris, France: Organisation for Economic Cooperation and Development, OECD; 2007.
41.Katritzky, R.; Jain, R.; Lomaka, A.; Petrukhin, R.; Maran, U.; Karelson, M.; *Cryst. Growth Des*. 2001, 1, 261-265.

42.Dearden, J. C. *Sci. Total Environ*. **1991**, 109/110, 59-68.

43.Kitaigorodsky, A. I. In Molecular Crystals and Molecules; Loebl, E. M., Ed.; Academic Press: New York, **1973.**

44.Steinstrasser, R.; Pohl, L. *Angew. Chem., Int. Ed. Engl*. **1973**, 12, 617-630.