DOI: http://dx.doi.org/10.32571/ijct.1321749 E-ISSN: 2602-277X

International Journal of Chemistry and Technology

http://dergipark.org.tr/ijct

Research Article

Modeling of the n-octanol/water partition coefficient of a series of PAHs: QSPR model

Youssouf DRIOUCHE[1]*, Hamza HADDAG[2], Meriem FERFAR [1], Laid BOUCHAALA[1], Amel BOUAKKADIA[3], Rana AMIRI[4], Nabil BOUARRA[5], Samia ALEM[6]

[1] *Environmental Research Center, Alzon, BP.72 A, Menadia, Annaba, Algeria*

[2] *Organic Synthesis and Biocatalysis Laboratory (LSBO), Badji Mokhtar University, PB. 12, 23000, Annaba, Algeria*

[3] *Abbes Laghrour University, Faculty of Sciences and Technology - Khenchela, BP 1252 Route de Batna 40004 Khenchela, Algeria*

[4] *Department of Chemistry, Faculty of Science, University 20 Août 1955 of Skikda, B.P.26 street El-Hadaiek, 21000 Algeria*

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

[6] *Laboratory of Aquatic and Terrestrial Ecology, Department of Biology, Faculty of Sciences, Badji Mokhtar University, BP. 12, 23000 Annaba, Algeria*

Received: 2 July 2023; Revised: 14 June 2024; Accepted: 26 June 2024

***Corresponding author e-mail: y.driouche@cre.dz

Citation: Driouche, Y.; Haddag, H.; Ferfar, M.; Bouchaala, L.; Bouakkadia, A.; Amiri, R.; Bouarra, N.; Alem, S. *Int. J. Chem. Technol.* **2024**, 8(2), 121-127.

ABSTRACT

A simple linear model was used to investigate the correlation between the n-octanol/water partition coefficient (kow) of non-substituted fused polycyclic aromatic hydrocarbons (PAHs). Among (74) 3D-geometrically tested descriptors calculated using the Dragon software, volume V turned out to be the best descriptor to model the considered endpoint (with a squared correlation coefficient (R^2) of 0.9844 and a standard error of estimation (s) of 0.132 log units). The correlation coefficient cross-validation (Q^2) between experimental and predicted log kow for training and test sets was 0.9811 (for training set) and 0.9828 (for test set), respectively.

The reliability of the proposed model was further illustrated using various evaluation techniques: leave-5-out cross-validation, bootstrap, randomization tests, and validation through the test set.

Keywords: Polycyclic aromatic hydrocarbons, log kow, solute bulk, simple linear model, QSPR.

1. INTRODUCTION

The behavior and distribution of an organic chemical compound within the environment are heavily influenced by its physicochemical attributes. To better understand how organic substances are transported and partitioned in the environment, it is useful to examine key physicochemical properties, including but not limited to aqueous solubility, the n-octanol/water coefficients (kow), and vapor pressures.[1-3]

The n-octanol/water partition coefficient represents the equilibrium ratio of a chemical's concentration in n-octanol to its concentration in water within a two-phase

DOI: http://dx.doi.org/10.32571/ijct.1321749 E-ISSN: 2602-277X

system. The logarithm of the partition coefficient, log kow (also known as log P), servers as a crucial indicator of a molecule's lipophilicity. It has been extensively utilized for predicting biological activities and toxicological outcomes.[4-7]

Although there were approximately 30 000 compounds with measured log P values[8], which might appear significant at first glance, this number is relatively small when compared to the continuously growing demand for log P values in numerous compounds where this data is currently unavailable. Moreover, the process of experimentally determining log P values is laborious, time-consuming, and requires a high level of solute purity[9], making it incompatible with high-throughput techniques. Due to these constraints, there is a continuous focus on developing methods for predicting log P values instead.[10]

In the past few decades, a range of approaches (including fragment-based, atom-centric, and conformation-dependent techniques)[11-13] have been devised, with many of them implemented and accessible as computer programs. However, it is not uncommon that these methods lead to appreciate differences in the log P values calculated for a determined compound.
This study aims to develop a Quantitative Structure-Property Relationship model to forecast the n-octanol/water partition coefficient (log k_{ow} or log P) for a distinct set of 30 unmodified fused PAHs. These compounds are of significant concern within the scientific community because of their potential as environmental contaminants.[14]

2. MATERIALS AND METHODS

2.1. Experimental Data

In this work a set of 30 non-substituted PAHs containing from 2 to 7 fused rings with five and six carbon atoms were studied. Their chemical structures are listed in Figure 1.

Experimental values of log kow (Table 1) were taken from the handbook by Mackay *et al.*[15]

2.2. Generating Descriptors

Each compound's molecular structure was sketched on a computer through the utilization of the HYPERCHEM software,[16] and then pre-optimized using the MM$^+$ method (Polack-Ribiere algorithm).

The most favorable geometries in their lowest energy conformations were determined through the application of the semi-empirical PM3 method at a restricted Hartree-Fock level, with configuration interaction excluded. A convergence threshold of 0.01 kcal. Å^{-1} was used for the gradient norm. The acquired geometries were employed as input for the creation of 74 3D-geometrical descriptors using DRAGON software, version 5.3.[17]

Geometrical descriptors are generated from the molecule's 3D structure, allowing for the representation of how atoms are positioned relative to each other in three-dimensional space. These descriptors provide valuable insights and the ability to differentiate between similar molecular structures and various molecule conformations.

2.3. CADEX Algorithm

In this study, the CADEX algorithm, originally introduced by Kennard and Stone,[18] was utilized to split the dataset into two separate groups: a training dataset consisting 23 compounds used to build the model, and a test dataset comprising the remaining 7 compounds, exclusively employed for external validation purposes.

2.4. Chemometric Methods

With the MobyDygs software,[19] we constructed single-variable models using the Ordinary Least Square (OLS) regression approach. Specifically, we generated 74 regression models, each associated with one of the 74 Geometry descriptors, were generated. We ranked these models by evaluating their internal predictive performance, measured using Q^2, in descending order. The optimal model was selected from this population of models.

The adequacy of the computed model was evaluated using the standard deviation error in calculation (*SDEC*), and the multiple determination coefficient (R^2).

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

Cross validation techniques allow us to assess both the internal predictive capability (using methods like Q^2_{LMO} cross validation and bootstrap) and the model's robustness (via Q^2_{LOO} cross validation).[20]

Cross-validation techniques involve excluding a certain number of compounds from the training dataset, followed by model reconstruction. This rebuilt model is then utilized to predict the excluded compounds. This process is iterated for each compounds in the training dataset, resulting in predictions for all of them. If the process involves removing one compound at a time, it is referred to as the leave-one-out technique (LOO). Otherwise, if multiple compounds are taken away simultaneously, it is known as the leave-many-out technique (LMO). The leave-one-out or leave-many-out correlation coefficient, commonly denoted as Q^2, is calculated by assessing the accuracy of predicting these "test" compounds prediction.[21,22]

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

In the realm of statistics, the convention of denoting the variable y with a “hat” symbol indicates that it represents an anticipated value of the property under study. The subscript “i/i” is used to specify that these anticipated values are obtained from models that do not incorporate the compound being forecasted. TSS, on the other hand, is an abbreviation for the total sum squares.

The predictive residual sum of squares (PRESS) serves as a metric for measuring the dispersion of predicted values, holding a pivotal role in the determination of Q^2 and the standard deviation error in prediction (*SDEP*).

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

Typically, a Q^2 value greater than 0.5 considered a good result,[23] while a Q^2 value exceeding 0.9 is regarded as excellent.[24,25]

However, research[26,27] has shown that although Q^2 is an essential factor for assessing predictive power in a model, it is not enough on its own. In order to mitigate the risk of overestimating the model’s predictive prowess, a leave-many-out procedure was carried out. This process was repeated 5000 times, with 5 objects being excluded at each iteration, and it is denoted as $Q^2_{L(5)O}$.

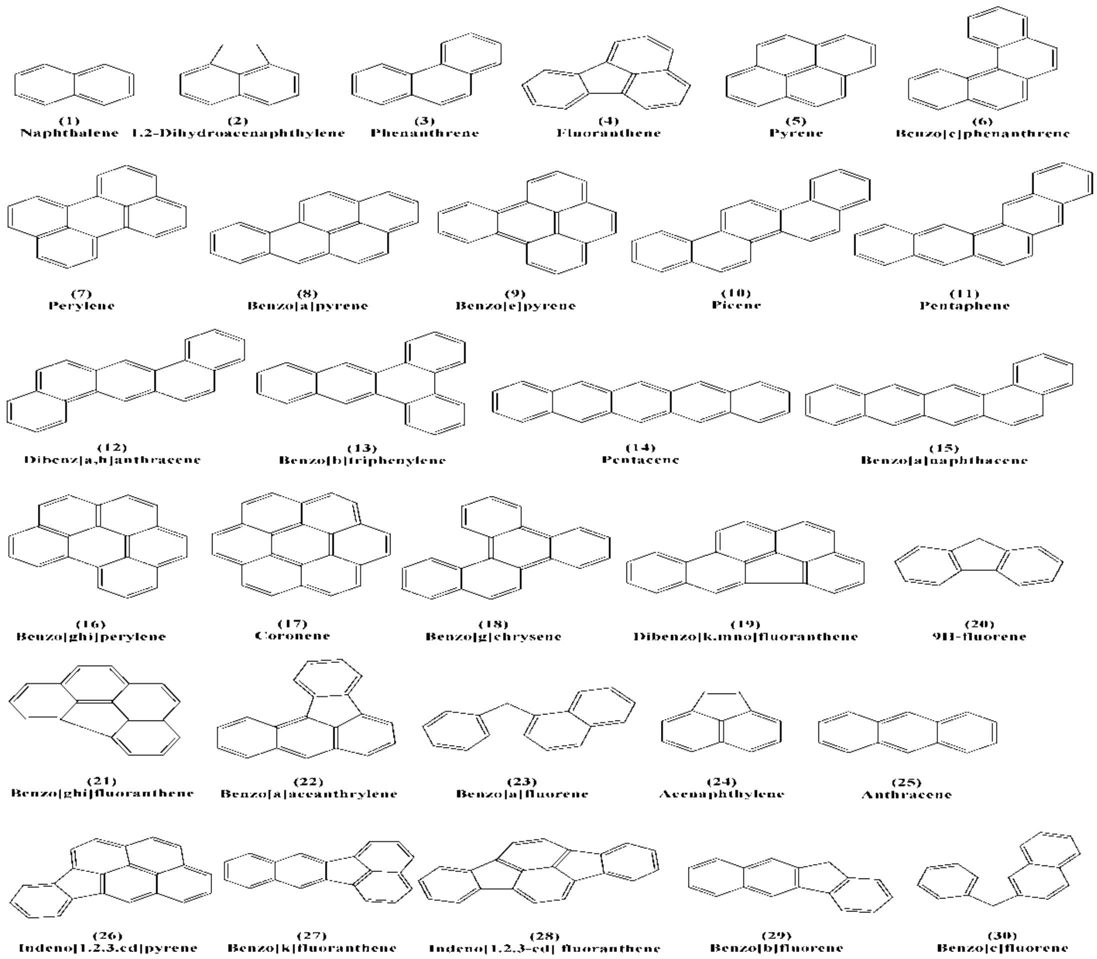

Figure 1. Chemical structures of PAHs used in this work.

The bootstrap validation method involves creating K n-dimensional sets by randomly selecting n-objects from the original dataset with replacement. A model is built using the initially chosen objects, and it is employed to forecast values for the omitted samples. Following that, Q^2 is calculated for each model. This bootstrapping process is repeated 8000 times for every validated model.

Utilizing the selected model, we compute the response values for the test instances and evaluate the accuracy of these predictions with regard to Q^2_{ext}, as defined below:

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

In this context, n_{ext} represents the count of items within the external dataset (or those excluded during

DOI: http://dx.doi.org/10.32571/ijct.1321749 **E-ISSN: 2602-277X**

bootstrap), while n_{tr} signifies the number of objects in the training dataset.

Other important parameters encompass R^2, computed for the validation substances utilizing the model constructed from the training dataset. Another crucial parameter to consider is the external standard deviation error of prediction ($SDEP_{ext}$), which is defined as follows:

$$SDEP_{ext} = \sqrt{\frac{1}{n_{ext}}\sum_{i=1}^{n_{ext}}\left(yi-\bar{y}\right)^{2}} \qquad \textbf{(5)}$$

In this equation, the summation is performed across the elements within the test dataset (n_{ext}).

2.5. Applicability Domain

The applicability domain[25,28] refers to a conceptual region within the space defined by the model's descriptors and the predicted response, where a particular QSPR is anticipated to deliver precise and dependable predictions. In this study, we assessed the structural AD using the leverage approach, where the leverage h_i[29] is defined as follows:

$$h_i = \boldsymbol{x}_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i \qquad \textbf{(6)}$$

In this context, $\boldsymbol{x}_i$ represents the descriptor row vector for the ith compound, $\boldsymbol{x}_i'$ denotes the transpose of $\boldsymbol{x}_i$, $\boldsymbol{X}$ corresponds to the model matrix created from the descriptor values in the calibration set, and $\boldsymbol{X}'$ stands for the transpose of $\boldsymbol{X}$.

Typically, the warning leverage value h^* is set at 3(m+1)/n, with n representing the total number of samples in the training set and m representing the number of descriptors considered in the correlation.

3. RESULTS and DISCUSSION

We derived the optimal one-dimensional equation using the volume index V, which is expressed as follows:

$$\log \text{kow} = -0.752 + 0.00925 \text{ Vol} \qquad \textbf{(7)}$$

$$R^2 = 0.9844 \;;\; Q^2_{LOO} = 0.9811 \;;\; Q^2_{L(5)O} = 0.9829 \;;$$
$$Q^2_{ext} = 0.9828 \;;\; Q^2_{Boot} = 0.9779 \quad SDEC = 0.126 \;;$$
$$SDEP = 0.139 \;;\; SDEP_{ext} = 0.128 \;;\; s = 0.132 \text{ log } unit$$
$$F = 1327.78 \; (p = 0.000)$$

It's important to remember that log kow quantifies the relative attraction between two incompatible liquid phases for a solute. As such, it must reflect the work done to form cavities in the two solvents, and the intermolecular forces, such as hydrogen bonding, that bind solute to solvent. It is therefore not surprising that numerous studies have shown[30,31] that log kow can be factorised into a volume term and one or more electronic terms.

For apolar solutes, log kow reflects only solute bulk, and thus is colinear with many parameters that model bulk. Thus,[32] showed that for such compounds, log kow correlated well with van der Waals Volume V_w.

The outcomes obtained from the randomized models can be compared to the original model by plotting the statistical coefficients R^2 and Q^2, as illustrated in Figure 2. The statistical metrics for the adjusted log kow vectors are considerably lower than those of the original QSPR model, indicating the presence of a genuine structure-property relationship.

The high R^2 value of 0.9844 offers compelling proof of the model's exceptional fit. In general, a greater magnitude of the ratio dignifies superior accuracy in forecasting property values within the training datasets. The substantial F ratio of 1327.78 underscores the model's competence in predicting log kow values. Additionally, the model demonstrates robustness, as there is only a slight difference between R^2 and Q^2(< 0.5%) Figure 3 illustrates the correlation between cross-validation log kow values and experimental log kow values. It's evident that a robust correspondence exists between the predicted and observed log kow values, with very little scattering around a linear pattern. The slope and intercept of the line are both very close to one and zero, respectively.

SDEC closely resembles *SDEP*, indicating that this model possesses internal predictability that is not significantly divergent from its fitting capability. In terms of internal validation, the model demonstrates remarkable stability, with only a marginal difference of 0.18% between Q^2 and $Q^2_{L(5)O}$. Moreover, the bootstrapping process further confirms the model's internal predictivity and stability.

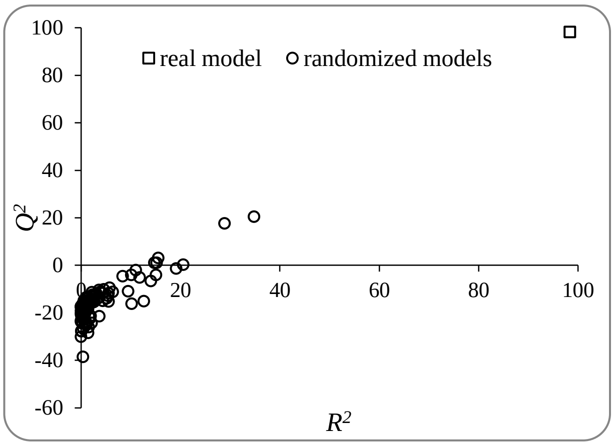

Figure 2. Randomization test associated to previous QSPR model.

DOI: http://dx.doi.org/10.32571/ijct.1321749 E-ISSN: 2602-277X

The data derived from Q^2_{ext} seems to exhibit a certain level of optimism, especially when working with small datasets comprising 20–30 compounds. In such cases, the validation of external predictability for entirely novel chemicals can only be confirmed in a subsequent, individualized manner.

The applicability domain of the linear model was assessed through the Williams plot (not reported here), which involves plotting standardized residuals against leverage values.

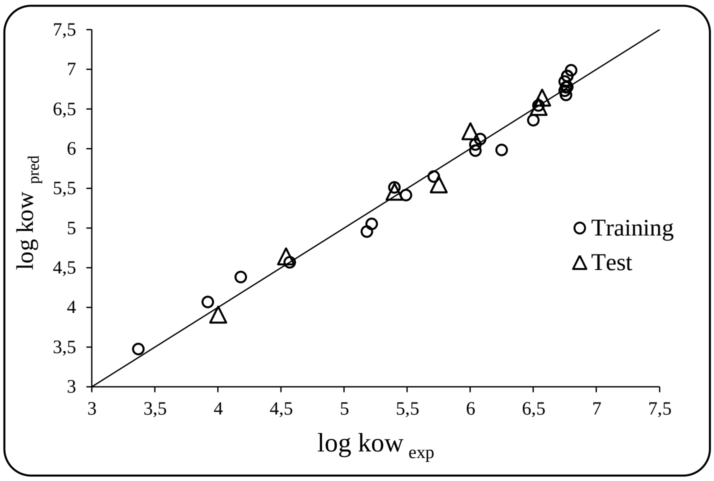

Figure 3. Cross-validation versus Experimental log kow

In Table 1 (column 6), samples with absolute standardized residual values less than 2.5 standard deviation units are listed, indicating that no Y- outliers were detected. Conversely, observation 1 (Naphthalene) within the training dataset (column 6), exhibiting a leverage exceeding the cautionary threshold of 0.261, can be identified as a structurally influential compound, often referred to as an X- outlier. The removal of observation 1 (Naphthalene) may result in a slight change in R^2, yielding a value of 0.9793 (with Q^2= 0.9735), and it would also lead to a decrease in the *F* ration to 1327.78.

Table. 1 Data for the studied PAHs.

ID	Object	Log kow$_{Exp}$	Log kow$_{Pred}$	Hat (h_i)	Std.Err.Pred ($e_{i\ std}$)
1	Naphthalene	3.37	3.5101	0.29	1.2622
2	1.2-Dihydroacenaphthylene	3.92	4.0932	0.183	1.4546
3	Phenanthrene	4.57	4.5631	0.116	-0.0560
4	Fluoranthene	5.22	5.0315	0.072	-1.4853
5	Pyrene	5.18	4.9281	0.079	-1.9929
6	Benzo[c]phenanthrene	5.71	5.6414	0.045	-0.5326
7	Perylene	6.25	5.9675	0.044	-2.1937
8	Benzo[a]pyrene	6.04	6.0492	0.045	0.0712
9	Benzo[e]pyrene	6.04	5.9688	0.044	-0.5530
10	Picene	6.77	6.7735	0.08	0.0281
11	Pentaphene	6.77	6.924	0.092	1.2269
12	Dibenz[a,h]anthracene	6.75	6.8506	0.086	0.7987
13	Benzo[b]triphenylene	6.76	6.7698	0.08	0.0778
14	Pentacene	6.80	7.0033	0.099	1.6261
15	Benzo[a]naphthacene	6.77	6.9241	0.092	1.2278
16	Benzo[ghi]perylene	6.50	6.3464	0.054	-1.1988
17	Coronene	6.75	6.7247	0.077	-0.2000
18	Benzo[g]chrysene	6.76	6.6676	0.073	-0.7285
19	Dibenzo[k.mno]fluoranthene	6.54	6.5382	0.064	-0.0143
20	9H-fluorene	4.18	4.4083	0.138	1.8666
21	Benzo[ghi]fluoranthene	5.49	5.4037	0.052	-0.6729
22	Benzo[a]aceanthrylene	6.08	6.1164	0.046	0.2833
23	Benzo[a]fluorene	5.40	5.5114	0.049	0.8670
24*	Acenaphthylene	4.00	3.9042	0.201	-0.8135

DOI: http://dx.doi.org/10.32571/ijct.1321749 E-ISSN: 2602-277X

25*	Anthracene	4.54	4.6389	0.105	0.7937
26*	Indeno[1.2.3.cd]pyrene	6.54	6.5232	0.062	-0.1318
27*	Benzo [k] fluoranthene	6.00	6.2148	0.049	1.6719
28*	Indeno[1.2.3-cd] fluoranthene	6.57	6.642	0.069	0.5668
29*	Benzo[b]fluorene	5.75	5.5468	0.047	-1.5804
30*	Benzo[c]fluorene	5.40	5.4522	0.050	0.4070

* Test compounds

3.1. Comparison to other studies

The results of this study were compared with those from previous publications that employed different modeling techniques (Dadfar *et al.*[33]*;* Mebarki *et al.*[34]). These comparisons are detailed with statistics in Table 2. The R^2, R^2 adj and RMSE values from this study were measured against those from other studies, demonstrating that the MLR model presented here offers more accurate predictions than most other methods, primarily due to the use of fewer descriptors. The difference between this work and the work in publications mentioned in the Table 2 is that the dataset is different and the number of compounds in the training set was not the same. Additionally, my study includes both a training set and a validation set, whereas the other publications do not include a validation set. This confirms that the presented model is superior to the other models.

Table. 2. Comparison of the presented results with those obtained from the other studies.

Reference	Method	Test set	Training Set	Descriptors	R^2 %	R^2 adj %	RMSE
This study	MLR	7	23	1	98.44	98.37	0.1320
Dadfar *et al.*[33]	MLR	–	43	3	31.20	25.90	–
	BP-ANN	–			98.40	–	0.1873
Mebarki *et al.*[34]	MLR	–	16	3	91.80	89.30	0.4316

4.CONCLUSION

In this research, a highly accurate QSPR model was developed for the prediction of the n-octanol/water partition coefficient (log kow, log P) for a set of thirty non-substituted PAHs. The findings demonstrated that the volume parameter effectively characterizes the molecular structure of these substances. The derived one-dimensional equation, which relies on the volume of these compounds, can reliably estimate the log kow values for both novel compounds and existing compounds with unknown experimental data.

Acknowledgements

The authors would like to extend their appreciation to the Environmental Research Center in Annaba for their valuable support and guidance.

Conflict of Interest

The authors do not have any competing interests related to this study.

REFERENCES

1. Touhami, I.; Messadi, D. *Enrgy Proced*. **2019,** 157, 522–532.
2. Xu, H.Y.; Zhang, J.Y.; Zou, J.W.; Chen, X.S. *J. Mol. Graph. Model.* **2008,** 26 (7), 1076-1081.
3. Touhami, I.; Haddag, H.; Didi, M.; Messadi D. *Chromatographia*. 2016, 79, 1023–1032.
4. Amiri, R.; Messadi, D.; Bouakkadia, A.; lourici, L. *Egypt.J.Chem.* **2019**, 62(9), 1563-1574.
5. Katritzky, A.R.; Kuanar, M.; Slavov, S.; Hall, C.D. *Chem. Rev*. **2010,** 110 (10), 5714-5789.
6. Li, W.R.; Song, G.B.; Ding G.H.; Gao, H. *IOP Conf. Ser.: Earth Environ. Sci*. **2020,** 612(1), 012044.
7. Amiri, R.; Messadi, D.; Bouakkadia, A. *J. Serb. Chem. Soc.***2020,** 85(4), 467-480.
8. Mannhold, R.; Van de Waterbeemd, H. *J. Comput. Aid. Mol. Des*. **2001,** 15(4), 337-354.
9. Mannhold, R.; Rekker, R.F. *Perspect. Drug. Discov*. **2000,** 18(1), 1-18.
10. Ziani, N. ; Amirat, K.; Messadi, D. *Rev. Sci. Technol. Synthèses*. 2014, 29, 51-58.
11. Benfenati, E.; Gini, G.; Piclin, N.; Roncaglioni, A.; Vari, M.R. *Chemosphere*. **2003,** 53(9), 1155-1164.
12. Mannhold, R.; Petrauskas, A. *QSAR. Comb. Sci*. **2003,** 22(4), 466-475.
13. Klopman, G.; Li, J.K.; Wang, S.; Dimayuge, M. *J. Chem. Inf. Comp. Sci*. **1994,** 34(4), 752-781.
14. Netto, A.D.P.; Moreira, J.C.; Dias, A.E.X.O.; Arbilla. G.; Ferreira, L.F.V.; Oliveira, A.S.; Barek, J. *Quim. Nova*. **2000,** 23(6), 765-773.

DOI: http://dx.doi.org/10.32571/ijct.1321749 **E-ISSN: 2602-277X**

15. Mackay, D.; Shiu, W.V.; Ma, K.C. Illustrated Handbook of Physical-Chemical Properties and Environmental Fate for Organic Chemicals, Vol.3, Lewis, London, **1998**.

16. HyperchemTM. Release 6.02 for windows. Molecular Modeling system, (**2000**). (http://www.hyper.com/).

17. Todeschini, R.; Consonni, V.; Mauri, A.; Paven, M. (2005) DRAGON Software for the calculation of Molecular Descriptors version 5.3 for Windows, Talete s. r. l., Milano, Italy. (http://www.talete.mi.it/).

18. Kennard, R.W.; Stone, L.A. *Technometrics*. **1996,** 11(1), 137-148.

19. Todeschini, R.; Ballabio, D.; Consonni, V.; Mauri, A. ; Paven, M. (2004) MobyDigs - version 1.1 - **2009** Copyright TALETE srl. (http://www.talete.mi.it/).

20. Bouakkadia, A.; Driouche, Y.; Kertiou, N.; Messadi, D. *Int. J. Saf. Secur. Eng*. **2020,** 10(3), 389-396.

21. Driouche, Y.; Messadi, D. *J. Serb. Chem. Soc*. **2019,** 84(4), 405-416.

22. Didi, M.; Haddag, H.; Driouche, Y.; Messadi, D. *Res. J. Pharm. Biol. Chem. Sci*. **2017,** 8(4):379-390.

23. Bouakkadia, A.; Kertiou, N.; Amiri, R.; Driouche, Y.; Messadi, D. *J. Serb. Chem. Soc*. **2021,** 86(7-8), 673–684.

24. Eriksson, L.; Jaworska, J.; Worth, A.; Cronin, M.Mc.; Dowell, R.M.; Gramatica, P. *Environ Health Persp.* **2003,** 111(10):1361-1375.

25. Tropscha, A.; Gramatica, P.; Grombar, V.K. *QSAR. Comb. Sci*. **2003,** 22(1), 69-77.

26. Kubinyi, H.; Hamprecht, F.A.; Mietzner, T. *J. Med. Chem*. **1998,** 41(14), 2553-2564.

27. Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Model*. **2002,** 20(4), 269-276.

28. Shen, M.; Béguin, C.; Golbraikh, A.; Stables, J.P.; Kohn, H.; Tropsha, A. *J. Med. Chem*. **2004,** 47(9), 2356-2364.

29. Weisberg, S. Applied Linear Regression, 3rd edition. John wiley and sons, Inc., Hoboken, New Jersey, **2005**.

30. Van de Waterbeemd, H.; Testa, B. The parametrization of lipophilicity and other structural properties in drug design. in: Testa B (Ed.) Advances in Drug Research, vol.16, Academic Press, New York, pp 85-225, **1987**.

31. Leahy, D.E. *J. Pharm. Sci*. **1986,** 75(7), 629-636.

32. Moriguchi, I.; Kanada, Y.; kawatsu, K. *Chem. Pharm. Bull*. **1976,** 24(8), 1799-1806.

33. Dadfar, E.; Shafiei, F.; Isfahani, T.M. *Curr Comput Aided Drug Des*. **2020**, 16(3), 207-221.

34. Mebarki, F.; Meneceur, S.; Abderrhmane Bouafia, A. *Asian J. Research Chem.* **2022**, 15(6), 443-8.