Prediction of Molecular Lipophilicity for Aromatic Aldehydes to Tetrahymena Pyriformis Using QSPR Approach

B. Souyei[1*], A. Boukhari[2], A. Khechekhouche[2], A. Miloudi[3], O. Mostefaoui[4], N. Zair[3], N. Kerttiou[5,6], A. Rebiai[1], N. Smakdji[7], A. Khelassi-Sefaoui[8]

[1]Applied chemistry and the environment laboratory, University of El Oued, Algeria

[2]Faculty of Technology, University of El Oued, Algeria

[3]New Technology and local Development Laboratory, University of El Oued, Algeria

[4]Faculty of Economic, Commerce and Management Sciences, University of El Oued, Algeria

[5]Environmental and Food Safety Laboratory (08 EFSL), University of Annaba, Algeria

[6]Faculty of Sciences and Technology - Khenchela, Abbes Laghrour University, Algeria

[7]Laboratory of Applied Energetics and Materials, University of Jijel, Algeria

[8]Hydraulic department, Institute of Sciences and Technology, University Center of Maghnia, Algeria

Received 03/10/ 2023; Accepted 05/01/ 2024; Published 19/01/2024

Abstract

In this study, a predictive QSPR (quantitative structure–property relationship) model was developed using Dragon descriptors to estimate the lipophilicity (LogKow) of aromatic aldehydes. The model was constructed with a dataset of 77 compounds and utilized multiple linear regression analysis, along with the combination of the ordinary least square regression method and genetic algorithm-based variable subset selection. The resulting model exhibited a high correlation coefficient (R2) of 88.71% and a standard error of estimation (s) of 0.324 log unit, indicating its reliability. Further validation was performed on an independent test set of 23 compounds, demonstrating the model's effectiveness in predicting the lipophilicity of new aromatic aldehydes. This valuable information can aid in drug design and optimization efforts, potentially facilitating the development of novel pharmaceuticals.

Keywords: Molecular Lipophilicity, QSPR, Molecular Descriptors, MLR, physicochemical property.

Tob Regul Sci.™ 2024;10(1): 2074 - 2092

DOI: doi.org/10.18001/TRS.10.1.132

1. Introduction

Aromatic aldehydes are derivatives of aromatic compounds that contain a carbonyl group attached to an aromatic ring. They can be selectively labeled at the formyl position through

various methods, such as reduction of esters with LiAlD4 followed by oxidation, reaction of amides with deuterated Schwartz's reagent, reductive carbonylation of aryl halides via Pd/Rh-cocatalyzed reactions, or deoxygenative deuteration of carboxylic acids with synergistic photoredox and organic catalysis [1-4]. Electron-withdrawing groups attached to the aromatic ring of aldehydes can enhance their reactivity when compared to electron-donating groups [5]. This property makes them valuable for synthesizing a diverse range of compounds. For instance, researchers have discovered a method to synthesize benzothiazoles by condensing 2-aminobenzenethiol and aromatic aldehydes in refluxing toluene at 110 °C [6]. Furthermore, the photophysical properties of aromatic aldehydes are summarized in the study, highlighting their potential application as photoinitiators in polymerization reactions. These organic synthesis applications were discussed in a prior work [7].

In drug discovery and development, lipophilicity represents a fundamental physicochemical property that holds significant importance [8,9]. The logarithmic partition coefficient (logP) is a key measure of lipophilicity and plays a crucial role in medicinal chemistry [10,11]. It serves as a highly informative and effective parameter in this field, affecting various aspects of drug behavior, including solubility, membrane permeability, potency, selectivity, promiscuity, metabolism, pharmacokinetics, pharmacodynamics, and toxicological profile [12, 13]. As a result, lipophilicity significantly influences the ADMET (absorption, distribution, metabolism, excretion, and toxicity) characteristics of drugs.

Lipophilicity serves as a crucial parameter in various fields, including pharmacology, medicine, food science, chemical industry, fragrance development, and environmental protection [16]. To assess this property, the octanol-water partition coefficient (K_{ow}) is commonly used, representing a substance's solubility in both aqueous and organic phases [14]. Specifically, the logarithmic partition coefficient ($logK_{ow}$) or n-octanol/water partition ratio plays a vital role in environmental risk assessment of chemicals, enabling estimation of environmental fate, bioavailability, exposure, and toxicity of compounds [15].

Given the significance of $logK_{ow}$, exploring its relationship with the molecular structure of compounds becomes imperative. This investigation is facilitated by employing the quantitative structure-rroperty relationship (QSPR) approach. QSPR allows for the study of the quantitative relationship between molecular descriptors and various properties or characteristics of compounds, such as $logK_{ow}$.

In recent years, the application of Quantitative Structure-Activity Relationship (QSAR) and QSPR approaches has seen a rising trend across diverse disciplines, particularly in drug design [17]. These quantitative modeling techniques have become invaluable tools for predicting a wide array of properties and activities of chemical compounds. Their efficacy has been demonstrated in predicting physicochemical properties, biological activity, toxicity, chemical reactivity, and metabolism of chemical compounds [18-22].

Presently quantitative structure-activity relationships (QSAR) remain one of the most frequently employed techniques for the discovery of novel and effective compounds against various diseases, including malaria [23], diabetes [24], cancer [25], and many others. AQSAR model is defined as an equation that incorporates molecular descriptors that have a significant impact on a specific

biological activity. These molecular descriptors serve as quantitative representations of the chemical and structural features of compounds, allowing for the prediction of their activity or potency.

QSPR models are mathematical relationships established between molecular descriptors and target properties, aiming to predict specific properties based on the molecular structure [26]. Various modeling techniques, including multiple linear regressions (MLR) and artificial neural networks, are commonly employed to develop QSPR models [27].

The objective of the current study is to construct a robust QSPR model capable of predicting the Lipophilicity ($\log K_{ow}$) values for a collection of aromatic aldehydes. To achieve this, general molecular descriptors are computed using DRAGON software, which provides comprehensive information about the chemical and structural features of the compounds.

2. Materials and Methods

2.1. Data set

The selection of 77 aromatic aldehyde compounds and their corresponding experimental log K_{ow} values was derived from the study conducted by Schultz and Netzeva [29]. The log K_{ow} values obtained from this research ranged from 0.79 to 3.89, as mentioned in Table 1 of the Results and Discussion section. To facilitate model development and evaluation, the dataset was divided into a training set and a test set.

2. 2 Descriptors Generation

The structural representation of the compounds under study holds immense significance in describing, communicating, and elucidating essential structural information based on their specific characteristics [30]. In the investigation of quantitative structure-property relationships (QSPR), the numerical representation of chemical structures through molecular descriptors plays a crucial role.To accomplish this, the molecules were initially created using the Hyperchem package (Version7.5) [31]. Subsequently, these molecular structures underwent pre-optimization utilizing the MM+ molecular mechanics force field. The final geometries of the minimum energy conformations were attained through more precise optimization employing the semi-empirical PM3 method. For the optimization process, a gradient limit of kcal/Å was utilized as a stopping criterion to obtain optimized structures. Subsequently, these optimized structures served as input for the generation of 1664 molecular descriptors from 20 different classes. Various classes of molecular descriptors were employed in this study, encompassing Constitutional, Topological, Geometrical, Charge, GETAWAY (Geometry, Topology, and Atoms Weighted Assembly), WHIM (Weighted Holistic Invariant Molecular descriptors), and 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction). The generation of these molecular descriptors was facilitated by utilizing Dragon software (version 5.4) [28]. During a preliminary step, constant values and descriptors exhibiting pairwise correlation were eliminated. If the pairwise correlation between descriptors exceeded 98%, one of the variables was removed. Subsequently, a genetic algorithm was employed for the selection of variables, resulting in a final set of descriptors.

2.3 Training and test sets Selection

To ensure the generation of a reliable model and to evaluate its predictive capability, it is essential to define a suitable training set and an external test set. The goal is to create two sets that exhibit comparable molecular diversity and encompass a wide range of structural and physicochemical properties present in the complete dataset. Typically, the test set should comprise between 15% and 40% of the compounds in the entire dataset. This ensures a reasonable representation of the data for evaluation purposes.

In this study, the data were separated into two independent subsets using the DUPLEX algorithm: a training set consisting of 54 compounds was used to build the model, while a test set containing the remaining 23 compounds was utilized to assess the model's predictive ability.

2.4 Model Development and Validation

For the analysis, MobyDigs software [32] was used to perform multiple linear regression (MLR) and variable selection. To conduct the regression, the Ordinary Least Square (OLS) method and Genetic Algorithm-Variable Subset Selection (GA-VSS) [33] were employed. The GA-VSS generated a series of 100 regression models, ranked based on their internal predictive performance in descending order. These models were then validated using R2CV. The validated models with fewer descriptors were found to have lower R2CV values. To explore various low-dimensional combinations, models with 1-2 variables were initially developed using the all-subset method. Then, the number of descriptors was incrementally increased, and new models were created. The top models at each rank were selected, and the final model was chosen from among them. This selection process aimed to ensure sufficient correlation while avoiding over-parameterization, which could lead to reduced predictive ability for molecules not in the training set. The recommended statistical guideline of $n/m \geq 5$ (where n is the number of samples and m is the number of descriptors) was considered [34]. The genetic algorithm was terminated when the R2CV value showed no significant improvement despite the model's size increase. Collinearity among the selected molecular descriptors was assessed using the QUIK rule (Q Under Influence of K) [35], a vital condition for the model's validity.

Due to the collinearity effect observed in the initial set of molecular descriptors, a large number of models with similar predictive power were generated in MOBYDIGS using different dimensionalities. To address this issue and avoid selecting models with comparable performance, a selection process was implemented to identify high-performance models. These models were chosen based on the difference in the K index (ΔK), calculated as Kxy - Kxx, and subsequently subjected to a thorough validation process. the quick rule mentioned in the study compares the multivariate correlation index KX, which is calculated from the X-block of predictor variables, with the multivariate correlation index KXY, obtained by augmenting the X-block matrix with a column representing the response variable. According to this rule, if KXY is greater than KX, the model is considered predictive [36]. In the specific case mentioned, the values obtained for these two indexes in the problem were KX = 37.02 and KXY = 45.14. As a result, based on this quick rule, the obtained model is deemed predictive since KXY is greater than KX (KXY > KX). This implies that the augmented model, which takes into account the relationship between the

predictor variables and the response variable, exhibits a higher level of predictive power compared to the original model.

The model's performance was evaluated based on parameters related to its predictive capability (R^2CV), fitting power (R^2), standard deviation error in prediction (SDEP), standard deviation error in calculation (SDEC), and standard error of estimation (s) within the domain of chemicals. To ensure the reliability and stability of the QSPR model developed using the MLR method, both internal and external validations were conducted. The quality and reliability of the fitting were initially assessed by calculating the coefficient of determination (R^2) between the experimental and calculated values of the training set particles. This assessment is represented by equation (1):

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

Where yi, ŷ,and are the observed, calculated and mean values of the lipophylicity, respectively.

The adjusted R^2 : Gives the percentage of variation explained by only the independent variables that actually affect the dependent variable. The formula is given by the equation (2):

$$R^2_{adj} = 1 - \left[\left(\frac{n-1}{n-m-1} \right) (1 - R^2) \right] \quad (2)$$

where n and m are the numbers of observations and descriptors, respectively.

Cross-validation is a widely used method for assessing the model's robustness. It involves creating modified data sets by systematically excluding one or a small group of molecules in each iteration, known as "leave-one-out" and "leave-some-out" procedures [37-39].

In this study, the internal predictive capability of the model was evaluated using the leave-one-out cross-validation technique (R2CV). The mathematical expression for this evaluation is as follows:

$$R^2_{cv} = \frac{SCT - PRESS}{SCT} \quad (3)$$

$$PRESS = \sum_i^n (y_i - \hat{y}_{(i)})^2 \quad (4)$$

$$SCT = \sum_i^n (y_i - \bar{y})^2 \quad (5)$$

The model generated using the initial selected objects is employed to predict values for the excluded sample, and subsequently, the (R^2CV) is calculated for each model. To ensure robustness, bootstrapping was repeated 8000 times. However, obtaining a robust model does not provide insights into its prediction power.

To assess the model's predictive capability, the compounds in the test set are utilized for evaluation. The R^2CV external (R^2CVext) for the test set is determined using equation (6):

$$R^2_{CV,ext} = 1 - \frac{\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{ext}} (y_i - \bar{y}_{tra})^2} \quad (6)$$

Where n_{ext} and n_{tr} are the number of objects in the external set (or left out by bootstrap) and n_{tr} the number of training set objects, respectively.

The standard deviation error in calculation (SDEC) with equation (7).

$$SDEC = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \quad (7)$$

The standard deviation error in prediction (SDEP) with equation (8)

$$SDEP = \sqrt{PRESS/n} \quad (8)$$

The external standard deviation error of prediction ($SDEP_{ext}$), defined as:

$$SDEP_{ext} = \sqrt{\frac{1}{n_{ext}} \sum_{i=1}^{n_{ext}} (yi - \bar{y})^2} \quad (9)$$

2.5 Analysis of Applicability Domain

The notion of the applicability domain (AD) [40, 41] pertains to a theoretical region in the descriptor space utilized by a model and its corresponding modeled response. This region defines the scope within which the model can offer dependable predictions. In this study, the applicability domain of the QSAR models was described using the leverage (hii) approach [42]. The leverage approach utilizes the concept of leverage, denoted as hii, to determine the applicability domain. The warning leverage, h*, is typically set at a value of 3(K + 1)/n, where n represents the total number of samples in the training set, and K is the number of descriptors involved in the correlation. By setting this threshold, the leverage values beyond h* are considered indicative of potential outliers. To identify both response outliers (Y outliers) and structurally influential compounds (X outliers), the Williams plot [43] was employed. This plot involves the graphical representation of standardized residuals versus leverage values. By analyzing the distribution of data points on the Williams plot, it becomes possible to detect samples that exhibit significant deviations from the expected behavior. Such outliers can correspond to either unusual response values or structurally unique compounds that exert a disproportionate influence on the model.

3. Results and Discussion

In this study, a total of 77 compounds representing aromatic aldehydes were selected due to their significant interactive activity. The hydrophobicity of these compounds was quantified using the

logarithmic $logK_{ow}$ values, which ranged from 0.42 to 3.96. The $logK_{ow}$ values for each compound are listed in Table 1. In this study, MLR was employed to establish a relationship between the $logK_{ow}$values and descriptors from different blocks. The resulting equation model consists of five variables and exhibits good statistical parameters for the training set. Additionally, it demonstrates high generalization and prediction ability for the prediction set, as indicated in Tables The 1 and 2. The derived MLR equation model, denoted as equation (10), describes the relationship between the $logK_{ow}$ values and the selected descriptors. Unfortunately, the specific

equation is not provided in the given text, so it cannot be regenerated.

$LogK_{OW}$ = -3,00841 -0,15428 RDF050m -0,77374Mor03p +2,75239 Mor24p +0,58083HIC-0,74133BLTF96 (10)

Where, RDF050m is the Radial Distribution Function - 5.0 / weighted by atomic

masses [44]; Mor03p is the signal 03/weighted by polarizability [45]; Mor24p is the 3D-MoRSE-signal 24 / weighted by atomic polarizabilities [46]; HIC information content on the leverage magnitude [47]; BLTF96 is the Verhaar model of algae base-line toxicity from MLOGP ($mmol^{-1}$) [48]. The characteristics and specifications related to the descriptors computed using the MLR technique and present in the model are listed in Table3.

The positive correlation of the Mor24p with the HIC shows that an increase in the values of these factors implies a increase in the value of the $logk_{ow}$, while the negative correlation of the RDF050m with the Mor03p and BLTF96 shows that an increase in the values of these factors implies a decrease in the value of the $logk_{ow}$.

Table 1. Experimental and pretected$LogK_{ow}$ for the studied aromatic aldehydes

Id	Object	$logk_{ow}$(exp)	$logk_{ow}$(pred)	Err. (Pred)	Std.Err. (pred)
1	1-Naphthaldehyde	2,67	2,95	0,28	0,9
2	2,3,5-Trichlorobenzaldehyde	3,69	3,64	-0,05	-0,18
3	2,3-Dihydroxybenzaldehyde	1,03	1,13	0,1	0,31
4	2,4,5-Trimethoxybenzaldehyde	1,19	0,81	-0,38	-1,37
5	2,4-Dihydroxybenzaldehyde	1,33	1,15	-0,18	-0,59
6	2,4-Dimethoxybenzaldehyde	1,79	1,53	-0,26	-0,88
7	2,5-Dihydroxybenzaldehyde	1,33	1,01	-0,32	-1,04
8	2-Bromobenzaldehyde	2,48	2,43	-0,05	-0,17
9	2-Chloro-4-hydroxycarboxaldehyde	0,93	1,57	0,65	2,09

10	2-Chloro-5-nitrobenzaldehyde	2,25	1,82	-0,43	-1,47
11	2-Chloro-6-fluorobenzaldehyde	2,51	2,58	0,07	0,23
12	2-Fluorenecarboxaldehyde	3,43	3,54	0,11	0,37
13	2-Fluorobenzaldehyde	1,76	2,12	0,36	1,15
14	2-Hydroxy-1-naphthaldehyde	2,99	2,44	-0,55	-1,78
15	2-Hydroxy-3-nitrocarboxaldehyde	1,84	1,5	-0,34	-1,12
16	2-Hydroxybenzaldehyde	1,81	1,39	-0,42	-1,32
17	2-Tolualdehyde	2,26	2	-0,26	-0,82
18	3,4,5-Trihydroxybenzaldehyde	0,42	0,9	0,48	1,61
19	3,4-Dihydroxybenzaldehyde	1,03	1,12	0,09	0,3
20	3,5-Dibromo-4-hydroxycarboxaldehyde	3,3	3,68	0,38	1,39
21	3,5-Dibromosalicylaldehyde	3,42	3,48	0,06	0,23
22	3-Anisaldehyde	1,71	1,65	-0,06	-0,19
23	3-Bromo-4-hydroxycarboxaldehyde	3,42	2,33	-1.09 **	-3,48
24	3-Chloro-2-fluoro-5-(trifluoromethyl)benzaldehyde	2	2,94	0.94 *	3,69
25	3-Chlorobenzaldehyde	2,26	2,3	0,04	0,14
26	3-Cyanobenzaldehyde	1,18	1,43	0,25	0,78
27	3-Ethoxy-2-hydroxycarboxaldehyde	2,17	1,88	-0,29	-0,93
28	3-Ethoxy-4-hydroxybenzaldehyde	1,01	1,69	0,67	2,16
29	3-Fluorobenzaldehyde	1,76	2	0,24	0,77
30	3-Hydroxy-4-methoxybenzaldehyde	0,97	1,13	0,16	0,5
31	3-Hydroxy-4-nitrobenzaldehyde	1,42	1,25	-0,18	-0,56

32	3-Hydroxybenzaldehyde	1,38	1,4	0,02	0,07
33	3-Methoxy-4-hydroxybenzaldehyde	1,21	1,17	-0,04	-0,11
34	3-Methoxysalicylaldehyde	1,37	1,25	-0,12	-0,39
35	3-Tolualdehyde	1,99	1,93	-0,06	-0,2
36	4-(Pentyloxy)benzaldehyd	3,89	4,13	0,24	0,87
37	4,6-Dimethoxy-2-hydroxybenzaldehyde	1,26	1,64	0,38	1,25
38	4-Acetamidobenzaldehyde	1,25	1,73	0,48	1,49
39	4-Butoxybenzaldehyde	3,37	3,47	0,1	0,35
40	4-Chlorobenzaldehyde	2,13	2,21	0,08	0,27
41	4-Ethylbenzaldehyde	2,52	2,57	0,05	0,16
42	4-Hydroxy-1-naphthaldehyde	2,42	2,63	0,21	0,68
43	4-Hydroxybenzaldehyde	1,35	1,4	0,05	0,17
44	4-Isopropylbenzaldehyde	2,92	3	0,08	0,26
45	4-Methyl-1-naphthaldehyde	3,17	2,95	-0,22	-0,7
46	4-Nitrobenzaldehyde	1,56	1,66	0,1	0,31
47	4-Phenoxybenzaldehyde	3,96	4,1	0,14	0,52
48	5-Chlorosalicylaldehyde	2,65	1,99	-0,66	-2,09
49	Benzaldehyde	1,48	1,73	0,25	0,78
50	Pentafluorobenzaldehyde	3,39	2,82	-0,57	-1,93
51	Phenanthrene-9-carboxaldehyd	3,84	3,33	-0,51	-2,09
52	Phenyl-1,3-dialdehyde	1,36	1,46	0,1	0,3
53	p-Tolualdehyde	1,99	1,92	-0,07	-0,24
54	Terephthaldicarboxaldehyde	1,36	1,49	0,13	0,41
55	2,3,4-Trihydroxybenzaldehyde*	0,79	0,85	0,06	0,2
56	2,4,6-Trihydroxybenzaldehyde*	0,72	1,16	0,44	1,47

57	2,4-Dichlorobenzaldehyde*	3,08	2,98	-0,1	-0,32
58	2-Anisaldehyde*	1,72	1,68	-0,04	-0,11
59	2-Chloro-3-hydroxy-4-methoxybenzaldehyde*	1,17	1,06	-0,11	-0,37
60	2-Chlorobenzaldehyde*	2,33	2,38	0,05	0,15
61	2-Methyl-1-naphthaldehyde*	3,17	2,88	-0,29	-0,92
62	2-Nitrobenzaldehyde*	0,86	1,72	0,86	2,7
63	3,4-Dimethoxy-5-hydroxycarboxaldehyde*	1,25	1,58	0,33	1,17
64	3-Bromobenzaldehyde*	2,48	2,63	0,15	0,49
65	3-Nitrobenzaldehyde*	0,65	1,53	0,89	2,83
66	4-(Dimethylamino)benzaldehyde*	1,81	1,81	0	0
67	4-Anisaldehyde*	1,45	1,56	0,11	0,34
68	4-Biphenylcarboxaldehyde*	3,38	3,4	0,02	0,05
69	4-Bromobenzaldehyde*	2,48	2,39	-0,09	-0,29
70	4-Cyanobenzaldehyde*	1,21	1,42	0,21	0,66
71	4-Ethoxybenzaldehyde*	2,31	2,14	-0,17	-0,53
72	4-Fluorobenzaldehyde*	1,54	2,03	0,49	1,59
73	4-Hydroxy-3-nitrobenzaldehyde*	1,48	1,14	-0,34	-1,13
74	5-Bromosalicylaldehyde*	2,8	2,34	-0,46	-1,46
75	5-Bromovanillin*	1,92	2,39	0,47	1,59
76	5-Hydroxy-2-nitrobenzaldehyde*	1,75	1,16	-0,59	-1,9
77	6-Chloro-2-fluoro-3-methylbenzaldehyde*	2,3	2,67	0,37	1,19

* Members for the test set.

Table 2. The statistical parameters with n_{tr}=54, n_{test}=23.

Statistical parameters	n_{tr}	n_{ext}	Nv	R^2 (%)	R^2_{adj} (%)	R^2_{cv} (%)	$R^2_{CV,ext}$
	54	23	5	88,71	87,54	84,85	82,47
Statistical parameters	F	SDEC	s	SDEP	$SDEP_{ext}$	K_{xy}	K_x
	75,45	0,3055	0,324	0,3539	0,381	48,57	40,66

Table 3. Characteristics of the selected descriptors in MLR model.

Variable	Reg.coeff.	Err.coeff	Std.Coeff.	Err.std.coeff.
Constant	-4.235691	0.69553	1.39107	
RDF050m	-0.1490972	2.80E-02	-0.3072317	1.188974
Mor03p	-0.720197	0.1496242	-0.3457702	1.482311
Mor24p	2.38871353	0.8671428	0.14089108	1.055392
HIC	0.85788583	0.1820004	0.27177078	1.189704
BLTF96	-0.830512	8.71E-02	-0.674236	1.459617

The generated QSPR model indicates that the $logk_{ow}$ of 54 aromatic aldehydes to Tetrahymena pyriformis can be explained by the five selected descriptors in equation (10). It is important to note that the errors in the entire dataset are distributed on both sides of the zero line, suggesting the absence of any systematic error in the developed model.

The computer model serves two purposes: predicting $logK_{ow}$ and assessing the quality of its fit through the graph of calculated and predicted values of $logK_{ow}$ compared to experimental values. Based on this model, the relationship between the observed and computed $logK_{ow}$ is highly significant. The statistical parameter values of the model are presented in Table 2, and the corresponding values have been plotted in Figure 1.

Upon examining the statistical coefficients in Table 2, we observe that the QSPR model exhibits a higher determination coefficient (R^2 = 0.8871) and a lower standard deviation of errors (SDEC = 0.3055), indicating its reliability. The model's cross-validation is assessed using the leave-one-

out (LOO) method, resulting in a notable R^2cv value (R^2cv = 0.8485). This satisfies the conditions for predictability as suggested by R. Veerasamy [49] and A. Golbraikh [50]. The cross-validated MLR model, with an R2cv value of 0.8485, demonstrates its reliability, sensitivity, and statistical significance. Based on the given conditions, the regenerated statement can be as follows:

"The MLR model satisfies the following conditions: (1) R^2 is greater than 0.7, (2) R2cv is greater than 0.6, and (3) the difference between R^2 and Q2Loo is smaller than 0.1 [51, 52]."

To evaluate the predictive power of the developed model, external validation is employed. This involves using a set of compounds (remaining 23 compounds) that were not included in the training of the model. Comparing the values of $logK_{OW}$-test and $logK_{OW}$-obs, we observe a good prediction for the test set compounds (R^2cv ext in Table 2). Through these results, we can conclude that the model exhibits strong predictive performance, and the descriptors employed effectively describe the partition coefficient.

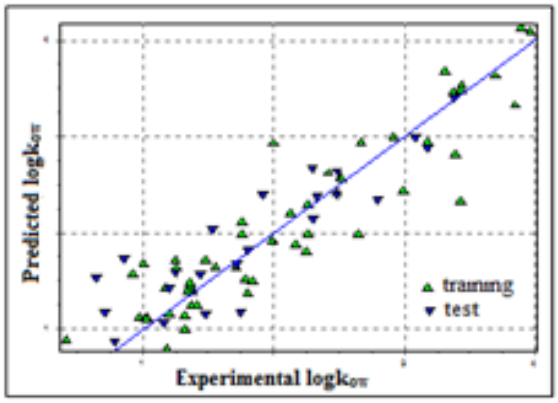

Figure 1. Plot of predicted vs. experimental $logk_{ow}$ for the entire data set

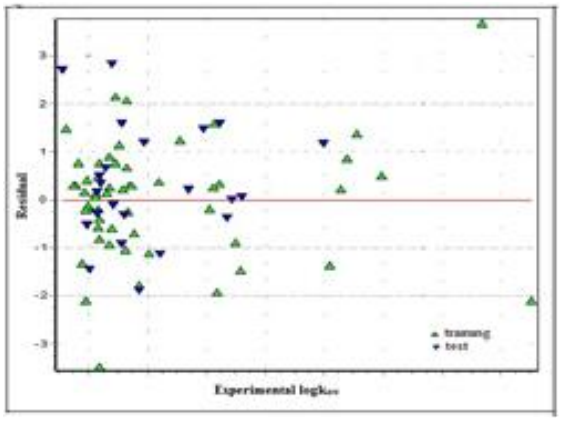

Figure 2. Plot of residual vs. experimental $logk_{ow}$ for the entire data set

A comparison between the results of the randomized models and the actual starting model can be made by plotting essential statistical coefficients, such as R^2 and Q2. Figure 3 illustrates this comparison. The statistical values for the modified logkow vectors are notably lower than those

of the genuine QSPR model, confirming the presence of a meaningful structure-property relationship.

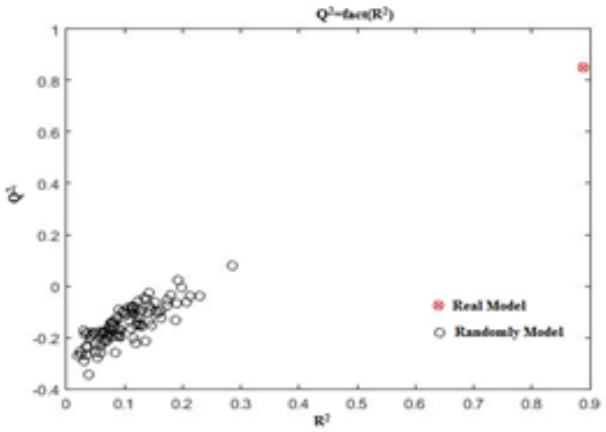

Figure 3. Randomization test associated with the previous QSPR model

3.2. Analysis of Descriptor Contribution

Based on the aforementioned procedure [40, 41], we conducted an analysis to determine the relative importance and contribution of the five descriptors in the model. The results were then plotted in Figure 4. By examining the figure and considering the percentage of contribution, we observed that the descriptors decrease in importance in the following order: BLTF96 (27.5635%) > Mor03p (19.1273%) > Mor24p (18.5050%) > RDF050m (18.4442%) > HIC (16.3600%). It is worth noting that the disparity in descriptor contribution between any two descriptors utilized in the model is not significant. This finding suggests that all descriptors play an essential role in constructing the predictive model and none of them can be deemed dispensable.

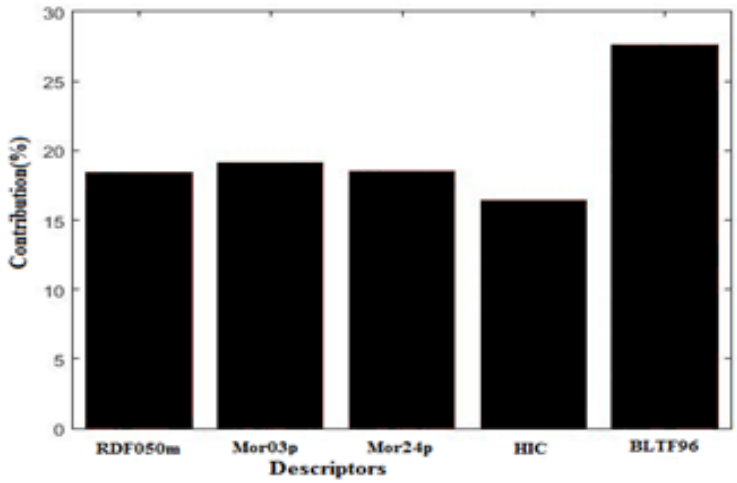

Figure 4. Contributions of Selected Descriptors in the MLR Model

3. 3. Domain of applicability

Based on Figure 5, we identified three compounds as outliers, one on the x-axis and the other two on the y-axis. These outliers are considered outside the scope of the descendant MLR model

and account for approximately 3.89% of the total compounds studied. Examining the diagram, we observed one outlier on the x-axis, specifically compound 51, which exhibits a leverage value higher than the warning limit of 0.33. This outlier is identified as Phenanthrene-9-carboxaldehyde. Its structure consists of three aromatic rings and one CHO functional group. Due to the presence of a high number of aromatic rings, this compound exhibits a high value of lipophilicity ($logk_{ow}$=3.84). This observation aligns with the findings of Timothy J. et al., who discovered a strong correlation between lipophilicity and the count of aromatic rings. Their research suggests that the addition of an aromatic ring often leads to a distinct and statistically significant increase in clogP [53].

The y-axis outliers in the model are represented by compounds 23 (3-Bromo-4-hydroxycarboxaldehyde) and 24 (3-Chloro-2-fluoro-5-(trifluoromethyl) benzaldehyde), which exhibit residuals higher than ±3σ in the training set. These compounds contain halogen atoms, specifically bromo, chloride, and fluoride. The presence of these halogen atoms in the molecular structure influences the lipophilicity of the compounds, resulting in their classification as outliers.

Gerebtzoff, G. et al., emphasize that halogenation of sp2 carbons is commonly employed to increase lipophilicity, which can enhance membrane permeability and oral absorption [54].

By removing these outliers, it is possible to improve the cross-validated R^2 (R^2cv) between the experimental $logk_{ow}$ values and the selected descriptors.

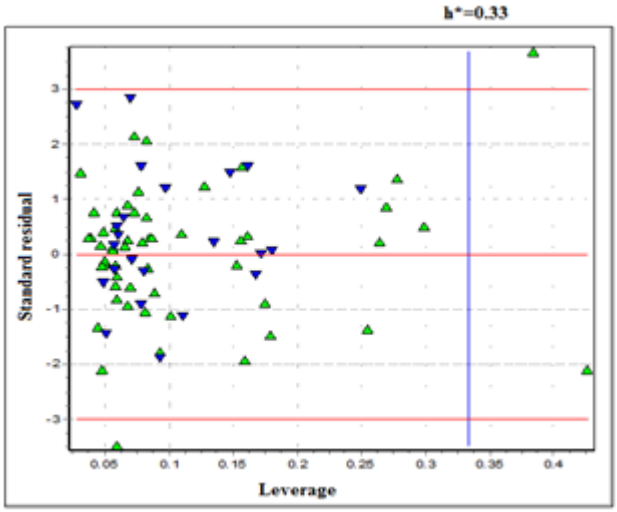

Figure 3. Williams plot for the descendant MLR model (with h* = 0. 33)

Conclusion

In this study, we employed multiple linear regressions (MLR) to predict the lipophilicity ($logK_{ow}$) of aromatic aldehydes to Tetrahymena pyriformis. The analysis involved the development of a linear model using a series of 77 compounds. The model utilized five parameters. The overall performance of the prediction was determined to be 88.71%, and the Standard Deviation of Error of Prediction (SDEC) for the training set was found to be 0.3055. The difference between R^2 (coefficient of determination) and R2cv (cross-validated coefficient of determination) is

significantly less than 0.1. This indicates that the MLR model exhibits significant predictive power and reliability.

The selected optimal model demonstrates adequate fitting precision and strong predictability, as supported by the statistical parameters and validation technique employed. As a result, this model can be effectively utilized for estimating the n-octanol/water partition behavior of aromatic aldehydes. The descriptors incorporated in this quantitative structure-property relationship (QSPR) model offer valuable insights into various molecular properties that contribute to intermolecular interactions affecting the n-octanol-water partition coefficient. These properties play a crucial role in determining the biological activity of the compounds. this model, can gain valuable information about the partition behavior of aromatic aldehydes, enabling us to better understand and predict their biological activity.

Reference

[1] E. P. K. Olsen, T. Singh, P. Harris, P. G. Andersson and R. Madsen, Experimental and Theoretical Mechanistic Investigation of the Iridium-Catalyzed Dehydrogenative Decarbonylation of Primary Alcohols J. Am. Chem. Soc. 137 (2) (2015) 834–842

[2] J. T. Spletstoser, J. M. White, A. R. Tunoori and G. I. Georg, J. Am. Chem. Soc., 2007, 129, 3408

[3] M. Y. S. Ibrahim and S. E. Denmark, Palladium/Rhodium Cooperative Catalysis for the Production of Aryl Aldehydes and Their Deuterated Analogues Using the Water–Gas Shift Reaction Angew. Chem., Int. Ed. 57, 2018, 10362 (Angew. Chem. 130, 2018 , 10519)

[4] M. Zhang, X.-A. Yuan, C.-J. Zhu and J. Xie, Angew. deoxygenative deuteration of carboxylic acids with D2O CHEM., INT. ED. 58 (2019) 312-316

[5] S .J. Ahmadi , M .Hosseinpour& S. Sadjadi , Non-catalytic condensation of aromatic aldehydes with aniline in high temperature water , Green Chemistry Letters and Reviews. 5 (3) (2012) 403-407.

[6] Y. Riadi, R.Mamouni, R.Azzalou, M.E.Haddadc, S.R.G.Guillaumetd, S. Lazara, An efficient and reusable heterogeneous catalyst animal bone meal for facile synthesis of benzimidazoles, benzoxazoles, and benzothiazoles,Tetrahedron. Lett. 52 (2011) 3492–3495.

[7] M. A. Theodoropoulou , N. F. Nikitas and G. K. Christoforos , Aldehydes as powerful initiators for photochemical transformations . Beilstein J. Org. Chem. 16 (2020) 833–857. https://doi.org/10.3762/bjoc.16.76

[8] E. Rutkowska, K. Pajak and K. Jozwiak, Lipophilicity methods of determination and its role in medicinal chemistry, Acta Pol. Pharm -Drug Research, 70(1) (2013) 3-18.

[9] V. S. Talismanov, S. V. Popkov, O. G. Karmanova, S.S. Zykova, A. P., Chernobrovkina, Convenient way to experimentally determine the lipophilicity(logP) of new synthetic biologically active compounds using high-performance liquid chromatography in the series

of 2,2-disubstituted 4-(1,2,4-triazol-1-ylmethyl)- 1,3-dioxolanes J. Pharm. Sci. & Res. 9(12) (2017) 2372-2375.

[10] F. Tsopelas, C. Giaginis, A. Tsantili, Lipophilicity and biomimetic properties to support drug discovery. Expert Opin. Drug Discov. 12 (2017) 885–896.

[11] S. Lobo, Is there enough focus on lipophilicity in drug discovery? Expert Opin. Drug Discov. 15 (2020) 261–263

[12] M. Dabrowska, Ł. Komsta, J. Krzek, K.Kokoszka, Lipophilicity study of eight cephalosporins by reversed-phase thin-layer chromatographic method. Biomed. Chromatogr. 29 (2015) 1759–1768.

[13] K. Kulig, B.Malawska, Estimation of the lipophilicity of antiarrhythmic and antihypertensive active 1-substituted pyrrolidin-2-one and pyrrolidine derivatives. Biomed. Chromatogr. [CrossRef] 17 (2003) 318–324.

[14] S. G. Machatha, S. H. Yalkowsky, Comparison of the octanol/water partition coefficients calculated by ClogP®, ACDlogP and KowWin® to experimentally determined values .Int. J. Pharm, 294 (2005) 185–192. DOI:10.1016/j.ijpharm.2005.01.023

[15] H. Geof, E. Charles, B. Bart, B. Alain, E. Marie Helene, G. Marc, K.Matthias, M. Eleanor, M. Dennis, M. Josef, O. Gunter, R. Jayne, S. Diederik, S. Ping and V.Joachim. A comparison of log Kow (n-octanol– water partition coefcient) values for non-ionic, anionic, cationic and amphoteric surfactants determined using predictions and experimental methods. Environ Sci Eur (2019) 31:1

[16] F. Spafiu, A. Mischie, P. Ionita , A. Beteringhe, T. Constantinescu et A. T. Balaban , New alternatives for estimating the octanol/water partition coefficient and water solubility for volatile organic compounds using GLC data (Kovàts retention indices) (x) (2009) 174-194

[17] G. Ghasemi, S. Arshadi, A. N. Rashtehroodi, M. Nirouei, S. Shariati and Z. Rastgoo, QSAR investigation on quinolizidinyl derivatives in Alzheimer's disease, Journal of Computational Medicine Volume 2013, Article ID 312728, 8 pages http://dx.doi.org/10.1155/2013/312728

[18] C., Hansch, and T. Fujita, p-σ-π analysis. A method for the correlation of biological activity and chemical structure. J. Am. Chem. Soc. 86(1964) 1616–1626. doi: 10.1021/ja01062a035

[19] C.Hansch, , and A.Leo, Substituent Constants for Correlation Analysis in Chemistry and Biology. New York, NY: John Wiley & Sons (1979). .

[20] H., Zhu, A.Sedykh, , S. K.Chakravarti, , and G. Klopman, , A new group contribution approach to the calculation of LogP. Curr. Comput. Aided Drug Des. 1 (2005) 3–9. doi: 10.2174/1573409052952323

[21] A.Cherkasov, , E. N.Muratov, D.Fourches, , A.Varnek, , I. I.Baskin, ,M.Cronin, , et al. QSAR modeling: where have you been? Where are you going to? J. Med. Chem., 57 (2014) 4977–5010. doi: 10.1021/jm4004285

[22] B. J. Neves, R. C.Braga, , C. C.Melo-Filho, J. T. Moreira-Filho, E. N.Muratov, , and C. H. Andrade,. QSAR-based virtual screening: advances and applications in drug discovery. Front. Pharmacol. 9(2018) 1275. doi: 10.3389/fphar.2018.01275

[23] M.Flores-Sumoza, J.J.Alcázar, E.Márquez, J.R.Mora, J.Lezama, E. Puello, Classical QSAR and docking simulation of 4-pyridone derivatives for their antimalarial activity. Molecules. 23 (2018) 3166. doi: 10.3390/molecules23123166.[PMC free article] [PubMed] [CrossRef] [Google Scholar]

[24] J.R.Mora, E.A.Márquez, L. Calle, Computational molecular modelling of N-cinnamoyl and hydroxycinnamoyl amides as potential α-glucosidase inhibitors. Med. Chem. Res. 27 (2018) 2214–2223. doi: 10.1007/s00044-018-2229-2.[CrossRef] [Google Scholar]

[25] N.Cabrera, J.R.Mora, E.A. Marquez, Computational Molecular Modeling of Pin1 Inhibition Activity of Quinazoline, Benzophenone, and Pyrimidine Derivatives. J. Chem. 2019; 11(2019). doi: 10.1155/2019/2954250

[26] P. Iswanto, E.V.Y. Delsy, E. Setiawan, F.A. Pu4tra, Quantitative Structure-Property Relationship Analysis Against Critical Micelle Concentration of Sulfonate-Based Surfactant Based on Semiempirical Zindo/1 Calculation .Molekul, 14 (2019) 78–83DOI: 10.20884/1.jm.2019.14.2.467

[27] D. Wang, Y. Yuan, S. Duan, R. Liu, S. Gu, S. Zhao, L. Liu, J. Xu, QSPR study on melting point of carbocyclic nitroaromatic compounds by multiple linear regression and artificial neural Chemometrics and Intelligent Laboratory Systems, 143 (2015) 7–15.DOI: 10.1016/j.chemolab.2015.02.009

[28] R .Todeschini, V.Consonni, A.Mauri, M.Pavan, 2005. DRAGON Software – version5.4-TALETE srl

[29] T.I.Netzeva, T.W.Schultz, QSARs for the aquatic toxicity of aromatic aldehydes from Tetrahymena data Chemosphere, 61 (2005) 1632-1643.

[30] B. Souyei, A. Hadj Seyd, F. Zaiz, and A.k.Rebiai, Application of Inverse QSAR/QSPR Analysis for Pesticides Structures Generation, ActaChim. Slov. 66 (2019) 315–325.

[31] Hyperchem TM, Rel. 7.5 for Windows Molecular Modeling System; Organisation: Evaluation Copy Dealer: Copyright © 2002 Hypercube .Inc.

[32] Todeschni R., Ballabio D., Consonni V., Mauri A., Pavan M., 2009. MOBYDIGS – version 1.1 – Copyright TALETE srl (2004).

[33] R. Leardi, R.Boggia, M.Tarrile, Genetic Algorithm as a Strategy for Feature Selection, J. Chemom, 6(1992267 – 281

[34] J.Xu, H.Zhang, Lei.Wang, G.Liang, L.Wang, X.Shen, W. Xu, QSPR study of absorption maxima of organic dye- sensitized solar cells based on 3D descriptors. Spectrochimica Acta Part A, 76(2010) 239-247.

[35] R.Todeschini, A.Maiocchi, V.Consonni, The K Correlation Index: Theory Development and its Application in Chemometrics. Chemom, Int. Lab. Syst, 46 (1999) 13 – 29

[36] Todeschini R.; Consonni, V.; in Manhold, R.; Kubinyi, H.; Temmerman, H. (Series editors), Handbook of molecular descriptors, Weinheim, New York, Wiley-VCH, (2000).

[37] D. W. Osten, Selection of optimal regression models via cross-validation. J Chemomitrics. 2(1998) 39-48.

[38] K. Roy, I. Mitra, S .Kar, et al. Comparative studies on some metrics for external validation of QSPR models. J Chem Inf Model. 52 (2012) 396-408.

[39] P.K Ojha, I .Mitra, R .Das, et al. Further exploring rm2 metrics for validation of QSPR models. ChemomIntell Lab Syst. 107(2011) 194-205

[40] A .Tropsha., P. Gramatica and V. K. Gombar, The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. QSAR & Combinatorial Science, 22(1) (2003) 69–77

[41] M. Shen, C.Béguin, A.Golbraikh, J. P Stables, H. Kohn and A.Tropsha, Application of predictive QSAR models to database mining: identification and experimental validation of novel anticonvulsant compounds. Journal of Medicinal Chemistry, 47(9) (2004) 2356 – 2364

[42] P.Gramatica, E. Giani, E. Papa Statistical external validation and consensus modeling: a QSPR case study for Koc prediction. J Mol Graph Model, 25 (6) (2007) 755-766

[43] SCAN- Software for Chemometric Analysisversion 1.1- for Windows, Minitab USA (1995).

[44] N. Chanin, T. Tanawut, N. Thanakorn, C. Isarankura-Na-Ayudhya1, P.Virapong Prediction of selectivity index of pentachlorophenol-imprinted polymers EXCLI Journal 5 (2006) 150-163 – ISSN 1611-2156

[45] H.Sakagami, T. Watanabe, T. Hoshino, N. Suda, K. Mori, T. Yasui, N.Yamauchi, H. Kashiwagi, T. Gomi,8 T. Oizumi, J. Nagai,10 Y. Uesawa, K. Takao, and Y. Sugita, Recent Progress of Basic Studies of Natural Products and Their Dental Application. Journal List Medicines (Basel) 6(1) (2019) .

[46] M. Nekoei, M. Mohammad Hosseini, A. Alavi Gharahbagh Quantitative Structure-Electrochemistry Relationship (QSER) Study for Prediction of Half-Wave Reduction Potentials of Some Organic CompoundsAnal. Bioanal. Electrochem. 1 (3) (2009) 159 - 168

[47] R.Todeschini, G P.ramatica, E.Marengo, R.Provenzani,. Weighted holistic invariant molecular descriptors. Part. 2. Theory development and applications on modeling physico-

chemical properties of polyaromatic hydrocarbons (PAH), Chemom. Intell. Lab. Syst, 27(1995) 221 – 229.

[48] Yuting Li, a Zhijun Dai,a Dan Cao,a Feng Luo,b Yuan Chen*a and Zheming Yuan*acChi-MIC-share: a new feature selection algorithm for quantitative structure–activity relationship models RSC Adv. 2020, 10, 19852 DOI: 10.1039/d0ra00061b

[49] Veerasamy, R., Rajak, H., Jain, A., Sivadasan, S., Varghese, C.P. and Agrawal, R.K. Validation of QSAR Models-Strategies and Importance. International Journal of Drug Design and Discovery. 2(2011) 511-519.

[50] A. Golbraikh, A. Tropsha. Quantitative Structure Activity Relationship Analysis of Selected Chalcone Derivatives as Mycobacterium tuberculosis Inhibitors J. Mol. Graphics Mod. 20(2002) 269–276.

[51] Chirico, N.; Gramatica, P. Real external predictivity of QSAR models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. J. Chem. Inf. Model. 51(2011) 2320–2335.

[52] Hawkins, D.M.; Basak, S.C.; Mills, D. Assessing model fit by cross-validation, J. Chem. Inf. Comput. 43 (2) (2003) 579–586.

[53] Timothy J. Ritchie and Simon J.F. Macdonald. The impact of aromatic rings count on compound developability-are too many aromatic rings a liability in drug design ? Drug discovery today volume 14, Number 21/22 November 2009

[54] Gerebtzoff, G.; Li-Blatter, X.; Fischer, H.; Frentzel, A.; Seelig, A. Halogenation of Drugs Enhances Membrane Binding and Permeation. ChemBioChem 2004, 5, 676–684.