

Research Article

Quantitative Structure-Property Relationship Modeling of the Octanol/Water Partition Coefficient of a Series of Alkylpyridines Using Quantum Chemical Descriptors

Fatogoma Diarrassouba^{1, 2, *}¹Faculty of Fundamental and Applied Sciences, Nangui Abrogoua University, Abidjan, Côte d'Ivoire²Faculty of Science and Technology, Alassane Ouattara University, Bouake, Côte d'Ivoire

Abstract

Lipophilicity is therefore the characteristic of certain substances of "liking" fatty substances. From a chemical standpoint, lipophilicity is a substance's affinity for non-polar solvents such as lipids. Lipophilicity is intimately linked to the concept of the partitioning of a molecule between an aqueous phase and a lipid phase. In this study, we modeled the octanol/water partition coefficient of a family of alkylpyridines using quantum-chemical molecular descriptors. Two molecular descriptors were calculated in this context: electronic energy (E_T) and the energy gap (ΔE_{gap}). QSPR (Quantitative Structure-Property Relationship) methods enabled us to correlate the octanol/water partition coefficient with these two molecular descriptors. The statistical and validation parameters are satisfactory. Statistical indicators have demonstrated a strong correlation between the octanol/water partition coefficient and these two independent variables. Validation parameters confirm that the established model is valid and performs effectively in predicting the octanol/water partition coefficient for the family of alkylpyridines studied. Consequently, it can be used to reliably predict the octanol/water partition coefficient for alkylpyridines within this same family provided they fall within the model's applicability domain at a 95% confidence level. From now on, designing new alkylpyridines with specific octanol/water partition coefficients requires only adjusting the two parameters of this model: electronic energy (E_T) and the energy gap (ΔE_{gap}). This work will thus simplify the task for experimental chemists synthesizing this family of molecules, provided they fall within the applicability domain of the developed QSPR model.

Keywords

Alkylpyridine, Lipophilicity, Octanol/water Partition Coefficient, Quantitative Structure-Property Relationship (QSPR), Quantum Chemistry, Molecular Descriptors, Multiple Linear Regression, Statistical Data Analysis

1. Introduction

For over a century, chemists have been developing methods for the synthesis of pyridine; interest in this field is unlikely to wane, given the enduring importance of the pyridine core in

both biological and chemical contexts. Many naturally occurring pyridine-based alkaloids are derivatives of nicotinic acid. Although picoline was isolated in 1846 [1], the elucidation of the pyridine structure by Körner and Dewar in 1869 and 1871,

*Correspondence: Fatogoma Diarrassouba (fatogomadiarras89@gmail.com)

Received: 24 September 2026; **Accepted:** 24 October 2026; **Published:** 26 October 2026



Copyright: © The Author(s), 2026. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

respectively marked the beginning of significant chemical research in this area [2]. Historically, many pyridine syntheses have relied on the condensation of amines and carbonyl compounds. These methods, along with other techniques for preparing the pyridine ring, are often characterized by the number of atoms contributed by each fragment to the formation of the six-membered aza-heteroaromatic ring. Pyridine derivatives constitute an important class of aza-heterocycles found in numerous natural products, active pharmaceutical ingredients, and functional materials [3-6]. Diploclidine [7] and nakinadine A [8] are two examples of recently isolated, structurally diverse natural products containing a pyridine core. Pyridine-derived drugs include atazanavir [9] (Reyataz) and imatinib mesylate [10] (Gleevec). These drugs are prescribed for the treatment of human immunodeficiency virus (HIV) infection and chronic myeloid leukemia, respectively. Pyridine derivatives are also incorporated into polymers such as polyvinylpyridine [11]. Prior to the organic synthesis of drugs, knowledge of the octanol/water partition coefficient is crucial, as it helps elucidate how drugs are absorbed by the human body. When experimental values for these physicochemical properties are unavailable, alternative methods to experimentation are often employed. Among these, Quantitative Structure-Property/Activity Relationships (QSPR/QSAR) have attracted significant interest and are even recommended under new regulations [12, 13]. They enable the development of mathematical relationships (models) linking physicochemical properties and biological activities to molecular structure; furthermore, they allow for explaining the origins of these activities or properties and predicting them for molecules lacking experimental data. The general objective is to model the octanol/water partition coefficient using QSPR methods. More specifically, we will correlate the octanol/water with certain quantum-chemical molecular descriptors, using an experimental dataset comprising fifteen (15) alkylpyridines.

2. Materials and Methods

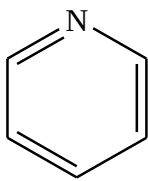
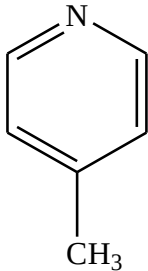
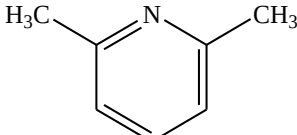
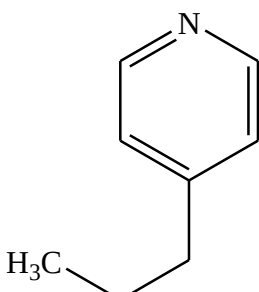
2.1. Materials

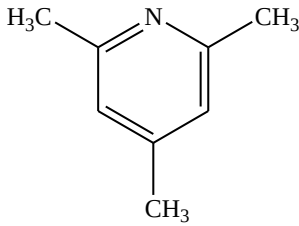
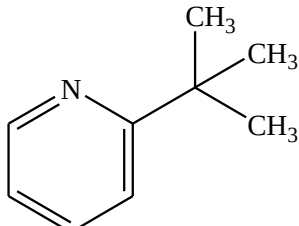
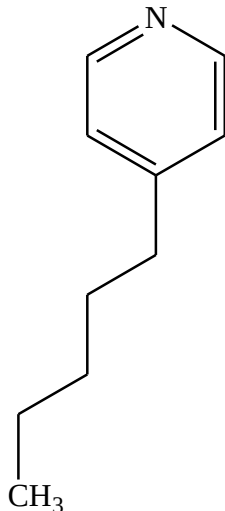
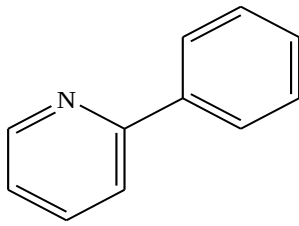
2.1.1. Experimental Database

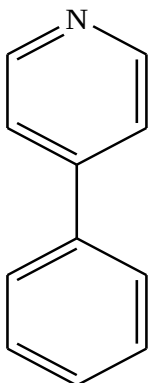
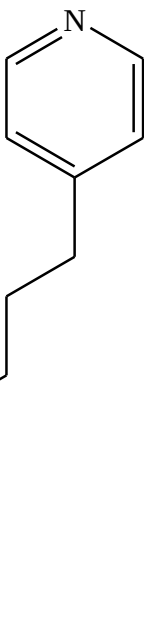
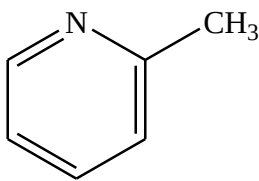
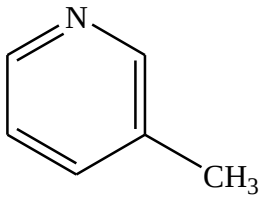
Our study focused on an experimental dataset comprising fifteen (15) alkylpyridines. These compounds were sourced from the work of James Sangster [14] and have available experimental water/octanol partition coefficients. This dataset was split into training and test sets. Several rules exist for selecting training and test sets for any given family of compounds [15]. For instance, a common guideline is to allocate 75% of the dataset to the training set and 25% to the test set [16]. The most widely used method among QSPR/QSAR (Quantitative Structure-Property Relationship/Quantitative Structure-Activity Relationship) practitioners is random splitting, which involves dividing the database through simple random selection. However, it is recommended that the test set

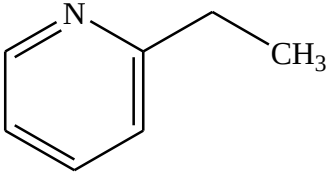
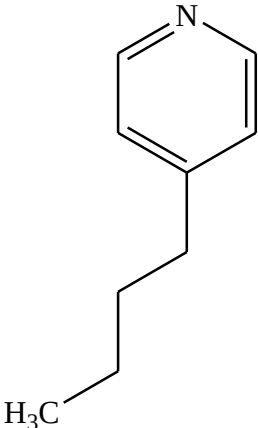
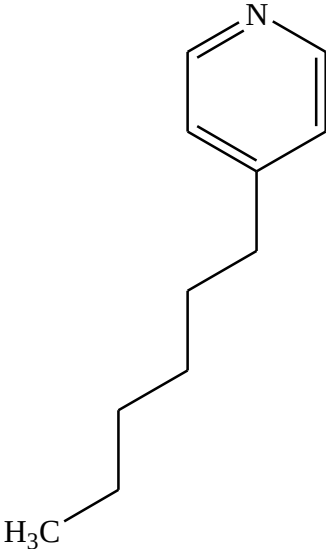
comprise at least 20% of the database [17, 18]. Accordingly, the experimental database was split randomly: ten (10) molecules were used for the training set (representing two-thirds of the experimental database), while the remaining five (5) molecules (representing one-third of the experimental dataset) were used for the test set. Table 1 presents the two-dimensional structures of these molecules along with their octanol/water partition coefficients.

Table 1. Series of studied alkylpyridines (experimental database) [14].

Compound	2D structure and name	LogP _{exp}
Training set		
PM_1	 Pyridine	0.650±0.05
PM_2	 4-Methylpyridine	1.220±0.10
PM_3	 2,6-Dimethylpyridine	1.680±0.15
PM_4	 4-n-Propylpyridine	2.100±0.20

Compound	2D structure and name	LogP _{exp}
PM_5	 2,4,6-Trimethylpyridine	1.880±0.25
PM_6	 2-Tert-butylpyridine	2.400±0.25
PM_7	 4-Pentylpyridine	3.450±0.30
PM_8	 2-Phenylpyridine	2.630±0.25

Compound	2D structure and name	LogP _{exp}
PM_9	 4-Phenylpyridine	2.590±0.10
PM_10	 4-Heptylpyridine	5.000±0.50
Test set		
PM_11	 2-Methylpyridine	1.110±0.15
PM_12	 3-Methylpyridine	1.200±0.15

Compound	2D structure and name	LogP _{exp}
PM_13	 2-Ethylpyridine	1.690±0.15
PM_14	 4-n-Butylpyridine	3.130±0.30
PM_15	 4-Hexylpyridine	4.350±0.40

2.1.2. Used Theory Level and Softwares

The GaussView 5.0 [19] software was used to represent the 3D structure and visualize the studied molecules. Then, the Gaussian 09 software [20] was used for optimization and frequency calculation (temperature 298.15 Kelvin, pressure 1 atmosphere, in vacuum). The theory level used is B3LYP/6-311G (d,p). As for 2D structures, they have been represented with chemsketch¹. The EXCEL software² was used for graphic representation. The XLSTAT software³ was used for modeling and statistical tests. For the calculation of the observation levers, the minitab 18⁴ software was used.

- 1) (2015) ACDLABS 10. Advanced Chemistry Development Inc.
- 2) Microsoft Corporation (2010) Microsoft Excel 2010
- 3) Addinsoft (2014) XLSTAT (Version 2014.5.03)
- 4) Minitab, LLC. (2017) Minitab Statistical Software (Version 18)

2.2. Methodes

2.2.1. Statistical Analysis

To develop a Quantitative Structure-Property Relationship (QSPR) model, a data analysis method is required. This method quantifies the relationship between the studied property and the molecular structure (descriptors). There are several methods for the implementation of a model and the analysis of its statistical data. But the one we used in our study is Multiple Linear Regression (MLR) (Many explanatory variables).

The generalized relation of the equation of a Multiple Linear Regression (MLR) [21] is as follows:

$$Y = a_0 + a_1X_1 + a_2X_2 + \dots + a_nX_n \quad (1)$$

In this expression, Y is the response or dependent variable, $X_1, X_2, \dots, X_n$ are descriptors (characteristics or independent variables) present in the model with corresponding regression coefficients $a_1, a_2, \dots, a_n$, respectively, and a_0 is the constant term of the model. For given values of $X_1, X_2, \dots, X_n$, the response Y must also follow a normal distribution. The descriptors in a regression model should not be highly correlated. For a statistically reliable model, the maximum number of descriptors should be on the order of one-fifth of the number of components in the training set. A regression model that fits the provided data well will produce a scatter plot, thus showing the deviation of the points from the line of best fit.

2.2.2. Molecular Descriptors

(i). Selection Criterion for Molecular Descriptors

In a QSAR study, two molecular descriptors are said to be independent when they are not correlated with each other. Statistically, this means that knowing the value of one does not allow us to predict or explain the value of the other. For a set of different descriptors, it is therefore necessary to perform a bivariate data analysis, that is, to calculate the correlation coefficient r between each pair of descriptors in order to assess the degree of intercorrelation. This type of analysis, called objective analysis, makes it possible to reduce the number of descriptors without involving the dependent variable (activity or property). Depending on the value of the correlation coefficient r , the relationships between two descriptors can be interpreted as follows [22]:

- 1) If $r > 0.97$, the two descriptors are considered strongly correlated,

- 2) If $r > 0.90$, the two descriptors are considered moderately correlated,
- 3) If $r < 0.90$, the two descriptors are considered weakly correlated,
- 4) If $r < 0.50$, the two descriptors are considered uncorrelated.

Two descriptors are considered strongly correlated ($r > 0.97$) or moderately correlated ($r > 0.90$) and cannot be used simultaneously in the same RQSA/RQSP model. However, since a correlation can also be negative, we will consider the absolute value of the correlation coefficient in our work. The molecular descriptors calculated in this work are described as follows:

(ii). Calculated Molecular Descriptors

Two molecular descriptors were calculated in this study: the electronic energy (E_T) and the energy gap (ΔE_{gap})

Electronic Energy (E_T)

This is the total energy of the molecule. It is given by the Gaussian 09 software after optimization and frequency calculations.

Energy gap (ΔE_{gap})

The energy gap is the energy between the Highest Occupied Molecular Orbital (HOMO) and the Lowest Unoccupied Molecular Orbital (LUMO). It is an important stability indicator [23].

$$\Delta E_{\text{gap}} = E_{\text{LUMO}} - E_{\text{HOMO}} \quad (2)$$

(iii). Paramètres Statistiques Et De Validations

The relationships 3, 4, 5 and 6 were used to calculate many statistical and validation parameters:

$$\text{ESS} = \sum (Y_{i,\text{cal}} - \bar{Y}_{\text{exp}})^2 \quad (3)$$

$$\text{TSS} = \sum (Y_{i,\text{exp}} - \bar{Y}_{\text{exp}})^2 \quad (4)$$

$$\text{RSS} = \sum (Y_{i,\text{exp}} - Y_{i,\text{cal}})^2 \quad (5)$$

$$\text{TSS} = \text{ESS} + \text{RSS} \quad (6)$$

where TSS: Total Sum of Squares; ESS: Extended Sum of Squares; RSS: Residual Sum of Squares.

Determination coefficient (R^2) [24]

The determination coefficient is given by the following relationship:

$$R^2 = 1 - \frac{\sum (Y_{i,\text{exp}} - Y_{i,\text{cal}})^2}{\sum (Y_{i,\text{exp}} - \bar{Y}_{\text{exp}})^2} = 1 - \frac{\text{RSS}}{\text{TSS}} \quad (7)$$

with

$$R = \pm \sqrt{\frac{\sum (Y_{i,\text{cal}} - \bar{Y}_{\text{exp}})^2}{\sum (Y_{i,\text{exp}} - \bar{Y}_{\text{exp}})^2}} = \pm \sqrt{\frac{\text{ESS}}{\text{TSS}}} \quad (8)$$

Standard deviation (s) [25]

It is an indicator of dispersion. It provides information on how the distribution of data is performed around the average. The closer its value is to 0, the better the adjustment and the more reliable the prediction.

$$s = \sqrt{\frac{\sum (Y_{i,\text{exp}} - Y_{i,\text{cal}})^2}{n-p-1}} = \sqrt{\frac{\text{RSS}}{n-p-1}} \quad (9)$$

Adjusted determination coefficient (R^2_{adjusted}) [26]

It allows to measure the robustness of a model unlike R^2 . This coefficient is used in multiple regressions because it takes into account the number of parameters (descriptors) of the model.

$$R^2_{\text{adjusted}} = 1 - \frac{(n - \text{Intercept})}{n-p-1} \cdot \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{(n - \text{Intercept})}{n-p-1} \cdot (1 - R^2) \quad (10)$$

Fisher-Snedecor coefficient (F) [27]

It allows to test the global significance of linear regression. A globally significant regression equation contains at least a relevant explanatory variable to explain the dependent variable. The Fisher-Snedecor coefficient is related to the determination coefficient by the following relationship:

$$F = \frac{n-p-1}{p} \cdot \frac{\text{ESS}}{\text{RSS}} = \frac{n-p-1}{p} \cdot \frac{R^2}{1-R^2} \quad (11)$$

Leave-One-Out cross-validation (Q^2_{LOO}) [28]

It measures the accuracy of the prediction on the data of the training set

$$Q^2_{\text{LOO}} = 1 - \frac{\sum (y_{i,\text{exp}} - y_{i,\text{pred}})^2}{\sum (y_{i,\text{exp}} - \bar{y}_{\text{exp}})^2} = 1 - \frac{\text{PRESS}}{\text{TSS}} \quad (12)$$

According to Erickson et al. [29], a model with a $Q^2_{\text{LOO}} > 0.50$ is satisfactory for prediction. Conversely, a model with a $Q^2_{\text{LOO}} > 0.90$ is considered excellent.

Cross-validation criteria (PRESS) [28]

The sum of the quadratic prediction errors. PRESS (Prediction Sum of Squares) is defined by the relationship:

$$\text{PRESS} = \sum (y_{i,\text{exp}} - y_{i,\text{pred}})^2 \quad (13)$$

This criterion is used to select models with good predictive power. (we always look for the smallest PRESS). A standard deviation of error of prediction (SDEP) is calculated from PRESS:

$$\text{SDEP} = \sqrt{\frac{\sum (y_{i,\text{exp}} - y_{i,\text{pred}})^2}{n}} = \sqrt{\frac{\text{PRESS}}{n}} \quad (14)$$

In these expressions, n is the number of molecules in the training set, p is the number of explanatory variables. $y_{i,\text{exp}}$ and $y_{i,\text{pred}}$ are respectively the experimental and predicted

values of property for molecule i and $\bar{y}_{\text{exp}}$ is average value of the property for the training set.

Todeschini's parameter ($^cR_p^2$) [30]

$^cR_p^2$ is the corrected form of P. P. Roy's parameter noted R_p^2 [31]. It allows to know if the model is due to chance correlations or not. If this parameter is greater than 0.50, the model is not due to a chance correlations. It is defined as:

$$^cR_p^2 = R \sqrt{R^2 - R_f^2} \quad (15)$$

with R_f^2 , the average value of R_{ri}^2 of the models obtained with the randomized property.

External validation coefficient (Q_{ext}^2) [32]

It measures the accuracy of the prediction on the test set data.

$$Q_{\text{ext}}^2 = 1 - \frac{n}{n_{\text{ext}}} \frac{\text{PRESS}(\text{test})}{\text{TSS}} \quad (16)$$

here, n_{ext} refers to the number of test set compounds

Parameter (RMSEP) [32]

External predictive ability of QSPR (Quantitative Structure-Property Relationship) model may further be determined by root mean square error in prediction given by:

$$\text{RMSEP} = \sqrt{\frac{\sum (y_{\text{exp}}(\text{test}) - y_{\text{pred}}(\text{test}))^2}{n_{\text{ext}}}} \quad (17)$$

Roy K. et al. parameters ($\bar{r}_m^2$ and Δr_m^2) [33]

For the acceptable prediction, the value of Δr_m^2 should preferably be lower than 0.20 when the value of $\bar{r}_m^2$ is more than 0.50.

$$\bar{r}_m^2 = \frac{(r_m^2 + r_m'^2)}{2} \quad (18)$$

$$\Delta r_m^2 = |r_m^2 - r_m'^2| \quad (19)$$

$$\text{here } r_m^2 = r^2 (1 - \sqrt{r^2 - r_0'^2}) \quad (20)$$

and

$$r_m'^2 = r^2 (1 - \sqrt{r^2 - r_0'^2}) \quad (21)$$

The parameters r^2 and $r_0'^2$ are the determination coefficients between the observed and predicted values of the compounds (training set or test set) with and without intercept, respectively. The parameter $r_0'^2$ bears the same meaning but uses the reversed axes.

External validation criteria or "Tropsha's criteria" [28, 34]

There are five such criteria:

1) Criterion 1: $R_{\text{ext}}^2 > 0.70$

2) Criterion 2: $Q_{\text{ext}}^2 > 0.60$

3) Criterion 3: $\frac{|R_{\text{ext}}^2 - R_0'^2|}{R_{\text{ext}}^2} < 0.1$ and $0.85 < k < 1.15$

4) Criterion 4: $\frac{|R_{\text{ext}}^2 - R_0'^2|}{R_{\text{ext}}^2} < 0.1$ and $0.85 < k' < 1.15$

5) Criterion 5: $|R_{\text{ext}}^2 - R_0'^2| < 0.3$

where, R_{ext}^2 stands for the determination coefficient of molecules for the test set; $R_0'^2$ represents the determination coefficient of the regression between predicted and experimental values for the test set without intercept; $R_0'^2$ is the determination coefficient of the regression between experimental and predicted values for the test set without intercept. k stands for the slope of the correlation line (values predicted according to the experimental values with intercept = 0) and k' is the slope of the correlation line (experimental values according to the predicted values with intercept = 0).

2.2.3. Domain of Applicability

The usefulness of a QSPR model lies in its predictive capacity for new chemical compounds. Thus, once the QSPR model is built, its domain of applicability (DA) must be defined. A model is considered valid only within its domain of applicability, and only the prediction of new compounds falling within this domain can be considered reliable, not an extrapolation of the model. The most common method for defining the DA is based on determining the leverage value of each compound [35]. The Williams plot (plotting of standardized or studentized residuals R as a function of leverage values h_{ii}) is used in this study to visualize the DA of the developed QSPR models. A lever is a kind of distance from the centroid of points in the space of explanatory variables. It identifies observations that are anomalously far from the others. For an observation i , the lever is expressed as follows:

$$h_{ii} = x_i (X^T X)^{-1} x_i^T \quad (22)$$

where x_i is the descriptor vector of the considered composite, X is the descriptor matrix derived from the descriptor values of the training set, and the threshold is defined as follows:

$$h^* = \frac{3(p+1)}{n} \quad (23)$$

where n is the number of compounds in the training set, and p is the number of descriptors in the proposed model. It should be noted that [36]:

1) If the h_{ii} value of a compound in the training set is greater than the threshold value h^* , the structure of that compound strengthens the developed model.

2) If all data points are located in the region of $0 \leq h_{ii} \leq h^*$ and $-3\sigma \leq R \leq 3\sigma$, the developed model can be considered statistically acceptable and valid.

3) Any compound with a standardized residual greater than three standard deviations ($R > 3\sigma$) and a leverage value less than the threshold ($h_{ii} < h^*$) can be considered an outlier in the response (Y outlier).

4) Conversely, if the leverage value of a compound is greater than the threshold value ($h_{ii} > h^*$) while the R -squared value is less than 3 standard deviations, this compound can be

considered structurally influential (X outlier). This stabilizes the developed model and makes it more accurate. It is called an influential good point.

5) Compounds with $h_{ii} > h^*$ strengthen the model when they are part of the training set, but otherwise will have questionable predicted values without necessarily being outliers, as the residuals may be low.

2.2.4. Shapiro-Wilk Test

The Shapiro-Wilk test [37] is very popular and is based on the parameter W. Compared to other tests, it is particularly powerful for small sample sizes ($n \leq 50$). The test statistic is written as:

$$W = \frac{\left[\sum_{i=1}^{k=\lfloor \frac{n}{2} \rfloor} a_i (y_{n-i+1} - y_i) \right]^2}{\sum_i (y_i - \bar{y})^2} \quad \text{with } -a_i = a_{n-i+1} \quad (24)$$

y_i corresponds to the series of sorted data.

$\lfloor \frac{n}{2} \rfloor$ is the integer part of the ratio $\frac{n}{2}$

a_i are constants generated from the mean and the variance-covariance matrix of the quantiles of a sample of size n following a normal distribution. The constants a_{n-i+1} are provided in specific tables. The W statistic can therefore be interpreted as the coefficient of determination (the square of the correlation coefficient) between the series of quantiles generated from the normal distribution and the empirical quantiles obtained from the data.

The null and alternative hypotheses of the Shapiro-Wilk test are as follows:

H_0 : The sample of size n follows a normal distribution.

H_1 : The sample of size n does not follow a normal distribution.

The value of the W statistic is higher the more credible the agreement with the normal distribution.

$$\text{If } W < W_{\text{critical}}, \text{ reject } H_0 \quad (25)$$

$$\text{If } W > W_{\text{critical}}, \text{ do not reject } H_0 \quad (26)$$

3. Results and Discussion

3.1. Values of Calculated Molecular Descriptors

In this study, two molecular descriptors were calculated: the energy gap (ΔE_{gap}) and the electronic energy (E_T). Additionally, the correlation between these two descriptors was examined using a correlation matrix. Table 2 presents the values of these molecular descriptors for both the training set and the test set.

Table 2. Values of calculated molecular descriptors.

Training set			
Compound	LogP _{exp}	ΔE_{gap} (eV)	E_T (a. u.)
PM_1	0.650	4.8054	-248.3469
PM_2	1.220	4.8517	-287.6760
PM_3	1.680	4.7042	-327.0091
PM_4	2.100	4.8379	-366.3245
PM_5	1.880	4.7348	-366.3380
PM_6	2.400	4.7392	-405.6490
PM_7	3.450	4.8396	-444.9729
PM_8	2.630	3.8626	-479.4599
PM_9	2.590	3.9333	-479.4544
PM_10	5.000	4.8400	-523.6214
test set			
Compound	LogP _{exp}	ΔE_{gap} (eV)	E_T (a. u.)
PM_11	1.110	4.6760	-284.6780
PM_12	1.200	4.7579	-287.6751
PM_13	1.690	4.7755	-327.0024
PM_14	3.130	4.8398	-405.6487
PM_15	4.350	4.8398	-484.2971

3.2. Normality Test for Experimental Octanol/water Partition Coefficients

Before any QSPR modeling, it is essential to test the normality of the experimental data to be modeled in this case, the octanol/water partition coefficient. The Shapiro-Wilk test is used to verify the normality of the data. A summary of the test parameters is provided in Table 3.

Table 3. Shapiro-Wilk test parameter values.

W	W_{critical}	p-value	$1 - \alpha$
0.9403	0.881	0.3865	0.05

Analysis of the data in Table 3 shows that the calculated p-value is greater than $1 - \alpha = 0.05$ (5% significance level). Regarding W_{critical} , the value is lower than the calculated W. Consequently, the experimental octanol/water partition coefficient values follow a normal distribution. This normality is confirmed by the distribution of the scatter plot along the first bisector (the line defined by $y = x$) (Figure 1).

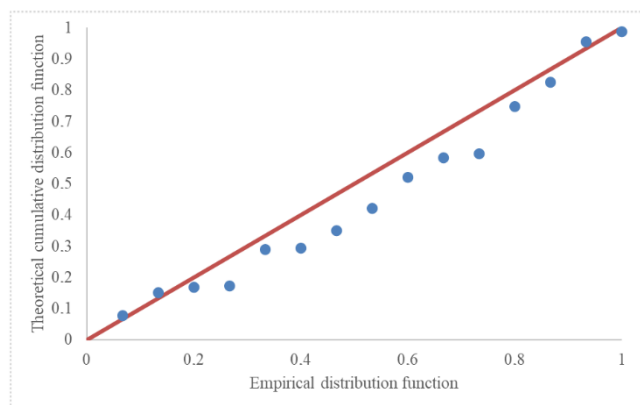


Figure 1. Model P-P plot (LogPexp).

3.3. Study of the Intercorrelation Between Molecular Descriptors

The correlation between the two descriptors must be zero or low to avoid any redundancy of information. The table below shows the correlation coefficient value between the two calculated molecular descriptors.

Table 4. Descriptor correlation matrix.

Variables	ΔE_{gap}	E_T
ΔE_{gap}	1.0000	
E_T	0.4822	1.0000

Based on the data in Table 4, the absolute value of the linear correlation coefficient between the energy gap (ΔE_{gap}) and the electronic energy (E_T) is less than 0.50 ($|r| < 0.50$). This value, being below 0.50, indicates the absence of correlation between these molecular descriptors. Consequently, these two descriptors can coexist within the same QSPR model. A QSPR model will therefore be developed using these two molecular descriptors.

3.4. Development of the QSPR Model: $\text{LogP} = f(\Delta E_{\text{gap}} \text{ and } E_T)$

A QSPR model based on the energy gap (ΔE_{gap}) and electronic energy (E_T) was developed using the XLSTAT modeling software.

QSPR model regression equation

The XLSAT software enabled us to determine the coefficients of the regression equation for the developed QSPR model. The regression coefficients are summarized in Table 5.

Table 5. Regression coefficient values for the model.

Source	Value	Standard error	t	Pr > t
Constant	-10.5088	1.3658	-7.6941	0.0001 (***)
ΔE_{gap}	1.4993	0.2429	6.1719	0.0005 (***)
E_T	-0.0151	0.0010	-14.6733	< 0.0001 (***)

$$|t_{\text{critical}}| = 2.36$$

The regression coefficients in Table 5 assigned to the various explanatory variables (the two molecular descriptors) lead to the following regression equation:

$$\text{LogP}_{\text{theo}} = -10.5088 + 1.4993\Delta E_{\text{gap}} - 0.0151E_T$$

The regression equation clearly shows that the regression coefficient for the energy gap (ΔE_{gap}) is positive, whereas the coefficient for the electronic energy (E_T) has the opposite sign. Consequently, the octanol/water partition coefficient increases as the energy gap increases and decreases as the electronic energy increases. Indeed, an increase in the energy gap leads to an increase in the octanol/water partition coefficient. Conversely, an increase in the (negative) electronic energy leads to a decrease in the octanol/water partition coefficient. It is also observed that the p-values for the intercept and the two molecular descriptors in the QSPR model fall within the [0; 0.001] range, indicating a highly significant influence (***) on the octanol/water partition coefficient. Thus, determining these two molecular descriptors is essential for evaluating the octanol/water partition coefficient of the family of molecules studied. The absolute values of the t-statistic and the contribution values reveal that electronic energy (E_T) makes the largest contribution (70.39%), indicating that it is the primary predictive descriptor for the octanol/water partition coefficient of the pyridine family studied. The energy gap, meanwhile, contributes 29.61%.

Model Analysis of Variance (ANOVA) table

This table allows for the easy calculation of various statistical parameters of the model.

Table 6. Model ANOVA table.

Source	D F	Sum of squares	Mean square	F	Pr > F
Model	2	12.8534	6.4267	108.1835	< 0.0001
Error	7	0.4158	0.0594		

Source	D F	Sum of squares	Mean square	F	Pr > F
Corrected total	9	13.2692			

(***) signifies "Highly significant"

In Table 6 (ANOVA table), it can be seen that the p-value falls within the range [0; 0.001], indicating that the model's regression equation is highly significant (***) for predicting the octanol/water partition coefficient of the series of pyridines studied. This high level of significance is confirmed by the very high value of the Fisher coefficient ($F = 108.1835$), which is well above the significance threshold ($F_{\text{limit}} = 4.74$). More precisely, at least one of the explanatory variables is relevant for explaining the dependent variable (octanol/water partition coefficient). A summary of the parameters determined using the ANOVA table is presented in Table 7.

Model statistical parameters

Various parameters related to the developed QSPR model were calculated to assess its quality. These parameters are summarized in the table below:

Table 7. Model statistical parameters.

n	R	R ²	R ² _{adjusted}	s	F
10	0.9842	0.9687	0.9597	0.2437	108.1835

Table 7 shows a very high correlation coefficient ($R = 0.9842$), indicating that the octanol/water partition coefficient is strongly correlated with the three calculated molecular descriptors. Regarding the adjusted coefficient of determination (R^2_{adjusted}), it provides greater precision than the standard coefficient of determination (R^2). The R^2_{adjusted} value is 0.9597, revealing that 95.97% of the experimental variance in the octanol/water partition coefficient is explained by the model's descriptors. Furthermore, the low standard deviation ($s = 0.2437$) indicates a good fit and high prediction reliability. These statistical parameters yielded satisfactory results, demonstrating the high quality of the developed QSPR model. However, validation tests are necessary to avoid overestimating the model's predictive capability regarding truly external molecules.

3.5. Internal Model Validation

Internal validation of a QSPR model applies only to the training set. The validation methods used are Leave-One-Out (LOO) cross-validation and the property randomization test (Y-randomization).

Leave-One-Out cross-validation of the model

The Leave-One-Out cross-validation parameters are listed in Table 8.

Table 8. Statistical parameters for the LOO cross-validation of the model.

n	Q ² _{LOO}	$\overline{r^2_m(LOO)}$	$\Delta r^2_m(LOO)$	PRESS	SDEP
10	0.9185	0.6890	0.0000	1.0819	0.3289

It is noteworthy that the model's LOO cross-validation coefficient is 0.9185, exceeding 0.90 ($Q^2_{\text{LOO}} > 0.90$). This demonstrates the model's excellence in predicting the octanol/water partition coefficient for the studied family of molecules. Furthermore, the model successfully predicts the partition coefficients for approximately 92 of the 100 molecules in the training set. The developed model thus exhibits high predictive capability regarding the training set molecules. This result indicates that the developed QSPR model is highly stable showing minimal sensitivity to the Leave-One-Out procedure as evidenced by the very high Q^2_{LOO} and R^2 (coefficient of determination) values, thereby confirming its robustness. Regarding the average coefficient $\overline{r^2_m(LOO)}$, the value is 0.6890 (above the 0.50 threshold), while $\Delta r^2_m(LOO)$ is below 0.2. Consequently, the developed model is considered acceptable for predicting the octanol/water partition coefficients of the training set molecules. To determine whether the established QSPR model is the result of chance, a Y-randomization test was performed.

Model Y-randomization test

For the randomization, a circular permutation was performed (i.e., 9 iterations). A summary of the mean parameter values from the randomization test is provided in Table 9.

Table 9. Mean parameter values for model randomization.

Randomized parameter	R ² _r	s _r	F _r	'R^2_{p}
Mean value	0.1667	1.2515	0.8624	0.8814

The mean value of the randomized coefficient of determination (R^2_r) is very low ($R^2_r = 0.1667$), indicating that the regression line equation accounts for only 16.67% of the data point distribution (octanol/water partition coefficient). Furthermore, there is significant dispersion of the data points around the regression line, confirmed by a high randomized standard deviation ($s_r = 1.2515$). The very low value of the randomized model statistic ($F_r = 0.8624$) indicates that the randomized model equation is not significant. Regarding Todeschini's corrected parameter ('R^2_{p}), its value is 0.8814,

which is well above 0.50 ($R_p^2 > 0.50$). Since this parameter exceeds 0.50, it is evident that the developed QSPR model is not the result of chance correlations.

3.6. External Model Validation

External validation applies only to the test set. Parameters were determined for this validation, and Tropsha's criteria were verified.

Statistical parameters for external model validation

Table 10. Statistical parameters for external model validation.

n _{ext} t	R _{ext} ²	Q _{ext} ²	$\overline{r_m^2}(\text{test})$	$\Delta r_m^2(\text{test})$	PRES S (test)	RMSE P
5	0.9968	0.9588	0.9669	0.0000	0.2730	0.2337

An analysis of the data in Table 10 reveals that the model possesses very high predictive power, evidenced by the high external validation coefficient value ($Q_{\text{ext}}^2 = 0.9588$). This indicates that, for every 100 molecules in the test set, the model successfully predicts the octanol/water partition coefficients for approximately 96 of them. Furthermore, 99.68% of the experimental variance in the acidity constant is explained by the model's descriptors. Regarding ($\overline{r_m^2}(\text{test})$) avg, the value exceeds 0.50, while the value for $\Delta r_m^2(\text{test})$ is below 0.2. Thus, the developed model is acceptable for predicting the properties of the test set molecules. Additionally, the five Tropsha criteria have been satisfied.

Verification of the model's Tropsha criteria

- 1) Criterion 1: $R_{\text{ext}}^2 = 0.9968 > 0.70$
- 2) Criterion 2: $Q_{\text{ext}}^2 = 0.9588 > 0.60$
- 3) Criterion 3: $\frac{|R_{\text{ext}}^2 - R_0^2|}{R_{\text{ext}}^2} = 0.0009 < 0.1$ and $k = 0.9224$ with $0.85 < k < 1.15$
- 4) Criterion 4: $\frac{|R_{\text{ext}}^2 - R_0'^2|}{R_{\text{ext}}^2} = 0.0009 < 0.1$ and $k' = 1.0816$ with $0.85 < k' < 1.15$
- 5) Criterion 5: $|R_{\text{ext}}^2 - R_0^2| = 0.0009 < 0.3$

It is observed that all five of Tropsha's criteria are met. Consequently, the model performs very well in predicting the octanol/water partition coefficient of the molecules in the test set from the experimental database.

Ratio $\tau = \frac{\text{Log}P_{\text{exp}}}{\text{Log}P_{\text{theo}}}$ of the test set

The model's performance was also evaluated by analyzing the ratio $\tau = \frac{\text{Log}P_{\text{exp}}}{\text{Log}P_{\text{theo}}}$ for the test series. The values obtained are summarized in Table 11.

Table 11. Values of the ratio τ .

LogP _{exp}	LogP _{theo}	τ
1.1100	0.8130	1.3652
1.2000	0.9812	1.2230
1.6900	1.6032	1.0542
3.1300	2.8905	1.0829
4.3500	4.0815	1.0658

Analysis of Table 11 shows that the values of the ratio τ are essentially equal to unity. This demonstrates that the model performs very well in predicting the octanol/water partition coefficient for the series of molecules studied and can be used to predict values for other molecules in the same family.

Comparison between model-predicted values and experimental values of the octanol/water partition coefficient for the test set.

The theoretical values of the lipophilicity parameter were compared with the experimental values using the regression constants from the $\text{Log}P_{\text{theo}} = f(\text{Log}P_{\text{exp}})$ plots in Figure 2 and the $\text{Log}P_{\text{exp}} = f(\text{Log}P_{\text{theo}})$ plots in Figure 3.

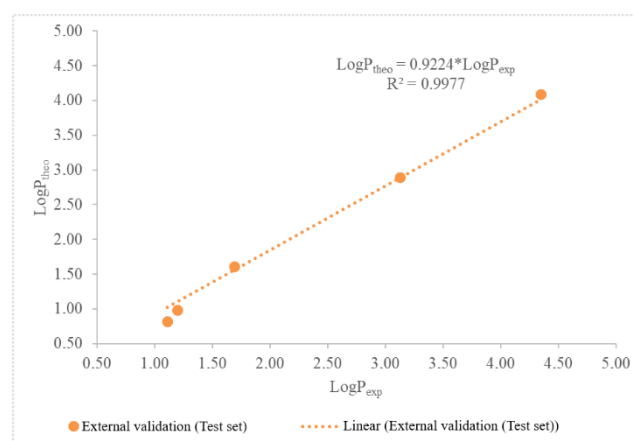


Figure 2. Plot of $\text{Log}P_{\text{theo}} = f(\text{Log}P_{\text{exp}})$ for the model test series (intercept = 0).

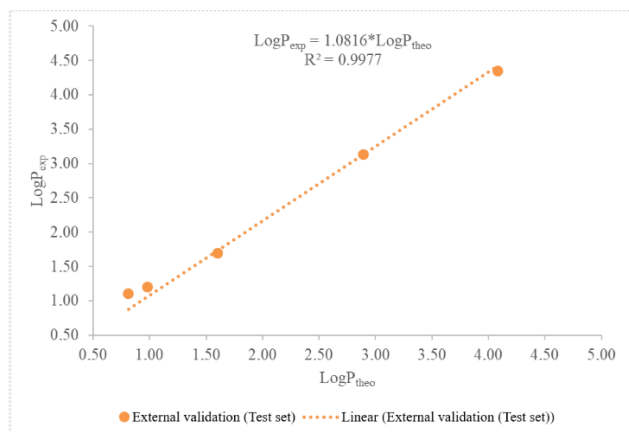


Figure 3. Plot of $\text{LogP}_{\text{exp}} = f(\text{LogP}_{\text{theo}})$ for the model test series (intercept = 0).

Based on Figures 2 and 3, the slope values $k = 0.9224$ and $k' = 1.0816$ are very close to unity. This indicates that the calculated LogP value is very close to the experimental value ($\text{LogP}_{\text{theo}} = f(\text{LogP}_{\text{exp}})$), resulting in a ratio close to unity. The external validation results further demonstrate that the model performs very well in predicting the LogP of the studied series of molecules. It can be effectively used to predict the LogP of new pyridines falling within its applicability domain.

3.7. Correlation Between Model-predicted Values and Experimental Values of the Octanol/water Partition Coefficient

The distribution of data points around the regression line in Figure 4 reveals a strong linear correlation between the model-predicted LogP values and the experimental values. Figure 5 shows a similarity between the curves representing the model-predicted values and the experimental values, particularly for the test set. Consequently, these graphs confirm that the model is validated and acceptable for LogP prediction. This also reflects the suitability of the level of theory employed to develop this QSPR model.

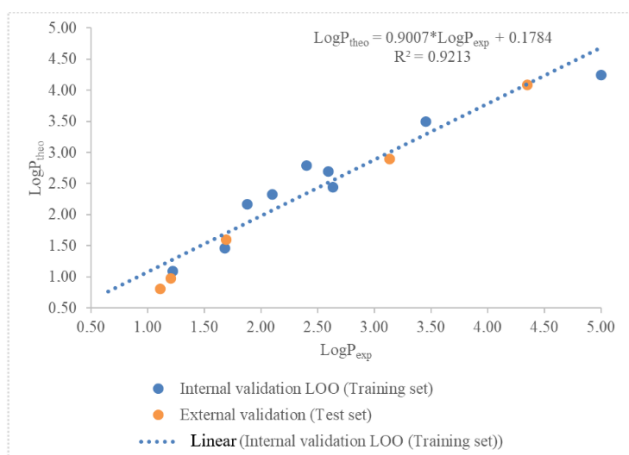


Figure 4. $\text{LogP}_{\text{theo}}\text{-LogP}_{\text{exp}}$ scatterplot of the model.

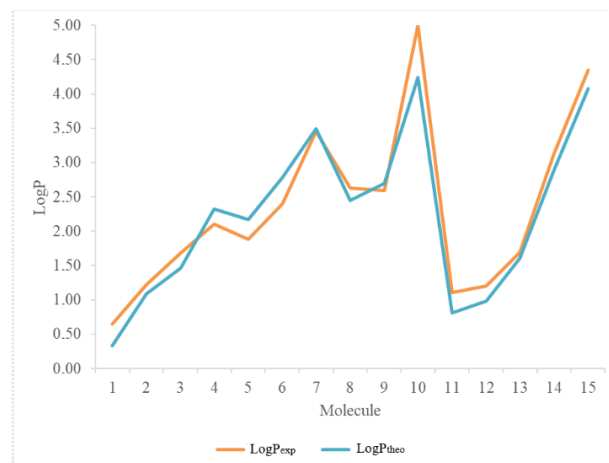


Figure 5. Similarity between model-predicted LogP values and experimental LogP values.

3.8. Normality Test for Model-derived Data

The Shapiro-Wilk test was used to verify the normality of the data derived from the model. The test parameters are summarized in Table 12.

Table 12. Shapiro-Wilk test parameters for the model.

W	W _{critical}	p-value	1 - α
0.9675	0.881	0.8196	0.05

Analysis of the values in Table 12 shows that the calculated p-value is greater than $1 - \alpha = 0.05$ (5% significance level). Regarding W_{critical} , the value is lower than the calculated W. Consequently, the normality hypothesis is consistent with our data. This normal distribution is confirmed by the scatter plot's alignment along the first bisector (the line defined by $y = x$) (Figure 6).

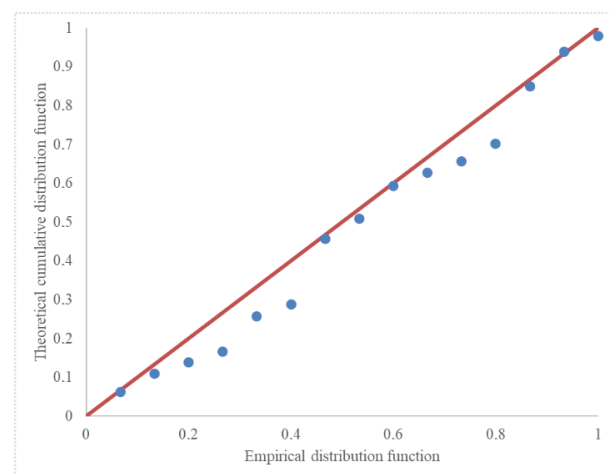


Figure 6. Model P-P plot ($\text{LogP}_{\text{theo}}$).

3.9. Applicability Domain of the Model

The applicability domain of the established QSPR model was defined using the Williams diagram, which is a graphical representation of studentized residuals versus leverage values. The figure below shows this Diagram.

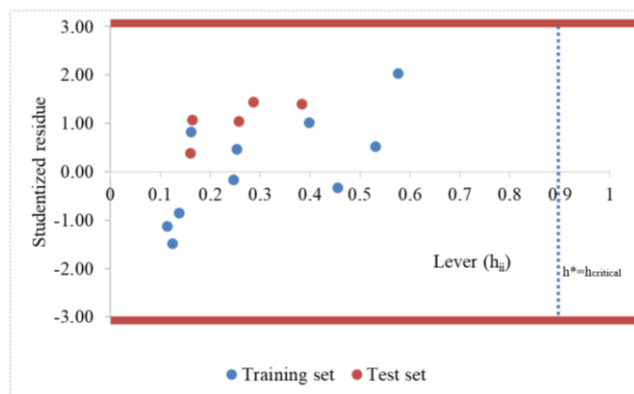


Figure 7. William's diagram of the model.

On the William's diagram, it can be seen that the studentized residuals for all observations in the training set fall within ± 3 standard deviation units. The same applies to the test set. All leverage values are below the threshold $h^* = 0.9000$. This indicates that no observation in either set is an outlier. The external validation results demonstrate that the model can be used to predict the octanol/water partition coefficient for future pyridines belonging to the same family and falling within the model's applicability domain.

4. Conclusion

This study focused on modeling the octanol/water partition coefficient of fifteen (15) pyridines using quantum chemical descriptors. A multiparametric QSPR model was developed, based on the energy gap (ΔE_{gap}) and electronic energy (E_T). This model exhibits highly satisfactory statistical and validation parameters, $R^2=0.9687$; $R^2_{\text{adjusted}}=0.9597$; $s=0.2437$; $F=108.1835$; $Q^2_{\text{LOO}}=0.9185$; $\overline{r}_m^2(\text{LOO})=0.6890$; $\Delta r_m^2(\text{LOO})=0.0000$; $R^2_{\text{ext}}=0.9968$; $Q^2_{\text{ext}}=0.9588$; $\overline{r}_m^2(\text{test})=0.9669$; $\Delta r_m^2(\text{test})=0.0000$. These various parameters demonstrate that the developed QSPR model is validated and performs well in predicting the octanol/water partition coefficient; it is considered an acceptable predictive model. Consequently, it can now be used to predict the octanol/water partition coefficient of future pyridines belonging to the same family and falling within its applicability domain. Therefore, future work will involve using this model to design new pyridines with desired octanol/water partition coefficients.

Abbreviations

QSPR	Quantitative Structure-Property Relationship
QSAR	Quantitative Structure-Activity Relationships
MLR	Multiple Linear Regression
TSS	Total Sum of Squares
ESS	Extended Sum of Squares
RSS	Residual Sum of Squares
PRESS	Prediction Sum of Squares
SDEP	Standard Deviation of Error of Prediction
HOMO	Highest Occupied Molecular Orbital
LUMO	Lowest Unoccupied Molecular Orbital
HIV	Human Immunodeficiency Virus

Author Contributions

Fatogoma Diarrassouba is the sole author. The author read and approved the final manuscript.

Conflicts of Interest

The author declares no conflicts of interest.

References

- [1] T. Anderson, On the Constitution and Properties of Picoline, a New Organic Base from Coal-Tar Edinburgh New Philos. J. 1846, 41, 146 - 156.
- [2] A. Pictet, The Vegetable Alkaloids With Particular Reference to Their Chemical Constitution, 2nd ed., Wiley, New York, 1904, pp. 10 - 31.
- [3] G. Jones, Comprehensive Heterocyclic Chemistry II, Vol. 5 (Eds.: A. R. Katritzky, C. W. Rees, E. F. V. Scriven, A. McKillop), Pergamon, Oxford, 1996, pp. 167 - 243;
- [4] Henry, G. D. De Novo Synthesis of Substituted Pyridines. Tetrahedron, 2004, 60, 6043-6061.
<https://doi.org/10.1016/j.tet.2004.04.043>
- [5] J. A. Joule, K. Mills, Heterocyclic Chemistry, 4th ed., Blackwell Science, Cambridge, 2000; p. 63 - 120;
- [6] J. P. Michael, Quinoline, quinazoline and acridone alkaloids, Nat. Prod. Rep. 2005, 22, 627 - 646.
- [7] L. Jayasinghe, C. P. Jayasooriya, N. Hara, Y. Fujimoto, A pyridine ring-containing ecdysteroid from *Diplocisia glaucescens*, Tetrahedron Lett. 2003, 44, 8769 - 8771.
- [8] Takaaki Kubota; Takami Nishi; Eri Fukushima; Jun Kawabata; Jane Fromont; Jun'ichi Kobayashi. Nakinadine A, a novel bis-pyridine alkaloid with a β -amino acid moiety from sponge *Amphimedon* sp., 2007, 48(29), 4983-4985.
<https://doi.org/10.1016/j.tetlet.2007.05.12>
- [9] Swainston Harrison, T. & Scott, L. J. Atazanavir: a review of its use in the management of HIV infection. Drugs, 2005, 65(16), 2309-2336.

- [10] Deininger, M. W. N. Specific Targeted Therapy of Chronic Myelogenous Leukemia with Imatinib. *Pharmacological Reviews*, 2003, 55(3), 401-423. <https://doi.org/10.1124/pr.55.3.4>
- [11] Raje, V. P., Bhat, R. P. and Samant, S. D., "Polyvinyl Pyridine as a Novel Solid Heterogeneous, Recyclable Catalyst for azo-Michael Reaction", *Synlett* 2006, 2006(16). 2676-2678.
- [12] European Parliament and Council of the European Union (2006) Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the registration, evaluation, authorisation and restriction of chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC and Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC and 2000/21/EC. Official Journal of the European Union, L 396, 1-849. <https://eurlex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32006R190>
- [13] N. Margossian, Le règlement REACH - La réglementation européenne sur les produits chimiques, Dunod / L'Usine Nouvelle, Paris, 2008.
- [14] Sangster, J. (1989) Octanol-water Partition Coefficients of Simple Organic Compounds. *Journal of Physical and Chemical Reference Data*, 1989, 18(3), 1111-1229. <https://doi.org/10.1063/1.555833>
- [15] K. Roy, S. Kar, R. N. Das, Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment, Academic Press, 2015. <https://doi.org/10.1016/C2014-0-00286-9>
- [16] Pierre-André Cornillon, Eric Matzner-Løber, *Régression avec R*, Springer-Verlag France, 2011.
- [17] Gramatica P, Giani E, Papa E., Statistical external validation and consensus modeling: A QSPR case study for Koc prediction. *J Mol Graph Model*, 2007, 25(6), 755-766. <https://doi.org/10.1016/j.jmgm.2006.06.005>
- [18] Kahn I, Fara D, Karelson M and al., *QSPR treatment of the soil sorption coefficients of organic pollutants*, *J Chem Inf Model* 45(1), 94-105, 2005. <https://doi.org/10.1021/ci0498766>
- [19] R Dennington, T Keith and J Millam., GaussView Version 5, Semichem Inc., Shawnee Mission, KS, 2009.
- [20] Gaussian 09, Revision A. 02, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2009.
- [21] Snedecor G. W., Cochran W. G., *Statistical methods*, 8, illustrée, réimprimée, Wiley-Blackwell, 276, 1991, 503 p. ISBN: 0813815614, 9780813815619.
- [22] N. Trinajstić, S. Nikolić, S. C. Basak, and I. Lukovits, —Distances indices and their hypercounterparts: Intercorrelation and use in the structure-property modeling, SAR and QSAR in Environmental Research, 12(1-2), 2001, 31-54. <https://doi.org/10.1080/10629360108035370>
- [23] Z. Zhou and R. G. Parr, Activation hardness: new index for describing the orientation of electrophilic aromatic substitution, *Journal of the American Chemical Society*, 1990, 112(15), 5720-5724. <https://doi.org/10.1021/ja00171a007>
- [24] Chatterje S. and Hadi A. S., *Regression Analysis by Example*. 4th Edition, John Wiley & Sons, Inc., Hoboken, 366 p, 2006.
- [25] A. F. Siegel, *Practical Business Statistics*, IRWIN, 3rd edition, 1997.
- [26] P. Besse., *Pratique de la modélisation statistique*, Publications du laboratoire de statistique et Probabilité, 2003.
- [27] R. D. Cook, S. Weisberg, An introduction to Regression Graphics, *Wiley Series in probability and mathematical statistics*, 1994, 265 p.
- [28] Golbraikh A., Tropsha A., Beware of q^2 !, *Journal of Molecular Graphics and Modelling*, 2002, 20(4), 269-276. [https://doi.org/10.1016/S1093-3263\(01\)00123-1](https://doi.org/10.1016/S1093-3263(01)00123-1)
- [29] L. Erikson, J. Jaworska, A. Worth, M. Cromin, R. M. McDowell, P. Gramatica., *Methods for reliability, uncertainty assessment, and applicability evaluations of regression based and classification QSPRs*, *Environmental Health Perspective*, 2003, 111(10), 1361-1375. <https://doi.org/10.1289/ehp.5758>
- [30] Mitra, Indrani; Saha, Achintya; Roy, Kunal. Exploring quantitative structure-activity relationship studies of antioxidant phenolic compounds obtained from traditional Chinese medicinal plants., 2010, 36(13), 1067-1079. <https://doi.org/10.1080/08927022.2010.503326>
- [31] P. P. Roy, S. Paul, I. Mitra, K. Roy, On Two Novel Parameters for Validation of Predictive QSAR Models, *Molecules*, 14(5), 2009, 1660-1701. <https://doi.org/10.3390/molecules14051660>
- [32] Consonni V., Ballabio D., Todeschini R., Evaluation of model predictive ability by external validation techniques, *J. Chemometrics*, 2010, 24(3-4), 194-201. <https://doi.org/10.1002/cem.1290>
- [33] Roy K., Mitra I., Kar S., Ojha P. K., Das R. N., Kabir H., Comparative studies on some metrics for external validation of QSPR models, *J. Chem Inf Model*, 2012, 52(2). <https://doi.org/10.1021/ci200520g>
- [34] Tropsha A., Gramatica P., Gombar V. K., The Importance of Being Earnest: Validation Is the Absolute Essential for Successful Application and Interpretation of QSPR, Models, *QSAR & Comb. Sc.*, 2003, 22(1), 69-77. <https://doi.org/10.1002/QSAR.200390007>

- [35] P. Gramatica, Principles of QSAR models validation: internal and external, *QSAR Comb. Sci.*, 2007, 26(5), 694-701. <https://doi.org/10.1002/qsar.200610151>
- [36] Haixia Lu, Wanqiang Liu, Fan Yang, Hu Zhou, Fengping Liu, Hua Yuan, Guanfan Chen, and Yinchun Jiao, Thermal Conductivity Estimation of Diverse Liquid Aliphatic Oxygen-Containing Organic Compounds Using the Quantitative Structure–Property Relationship Method, *ACS Omega*, 2020, 5(15), 8534–8542. <https://doi.org/10.1021/acsomega.9b04190>
- [37] Shapiro, S. S. and Wilk, M. B. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 1965, 52(3-4) 591-611. <https://doi.org/10.1093/biomet/52.3-4.591>