American Journal of Chemistry 2025, 15(3): 64-75
DOI: 10.5923/j.chemistry.20251503.02

Modeling the Anti-Prostatic Activity of a Series of Turmeric Derivatives Using the Quantitative Structure-Activity Relationship

Pétah Ismaëli Aziz Do-Koné[1], Nobel Kouakou N'guessan[2], Georges Stéphane Dembele[1,*], Kafoumba Bamba[1], Abdoulaye Konaté[2], Doh Soro[1]

[1]Laboratoire de Thermodynamique et de Physico-chimie du Milieu, UFR SFA, Université Nangui Abrogoua 02 BP 801 Abidjan 02, Côte-d'Ivoire
[2]Laboratoire Chimie Organique Structurale, UFR SSMT, Université Félix Houphouët Boigny BP 582 Abidjan 22, Côte-d'Ivoire

Abstract Curcumin derivatives are promising cytotoxic agents in the treatment of prostate cancer. The primary goal of this study is to establish a quantitative correlation between the structural features and anti-prostate activity of a series of sixteen (16) curcumin derivatives. Density functional theory (DFT) calculations at the B3LYP/6-31+G(d,p) level of theory were performed to determine the relevant molecular descriptors. The RQSA model developed using the multiple linear regression method (RLM) is a function of hardness (η), bond angle α(C-C=C), surface tension (TSurface) and density. Hardness (η) proved to be the most important descriptor for predicting the cytotoxic potential of the compounds studied. The statistical indicators associated with the model (R2=0.922; S= 0.068; F= 17.027) show that this model is robust with good predictive power. Our model's quality and performance were confirmed through internal validation using both leave-one-out (LOO) cross-validation and Tropsha criteria. This model is not due to chance and could be used to determine the cytotoxic activity of turmeric derivatives belonging to the same field of applicability.

Keywords QSAR, Curcumin, Cancer, Prostate, DFT

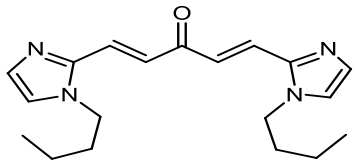
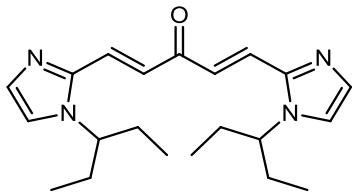
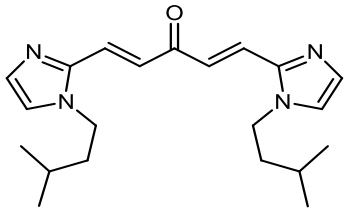
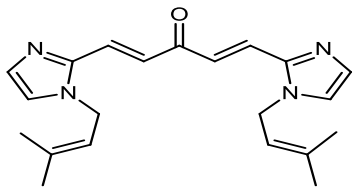
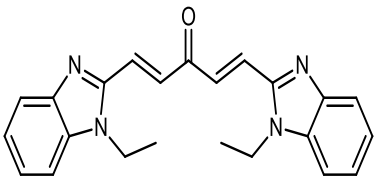
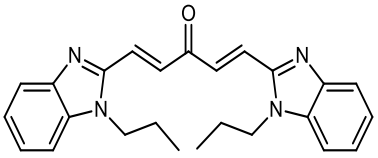
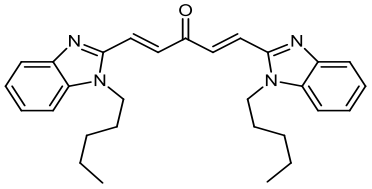
1. Introduction

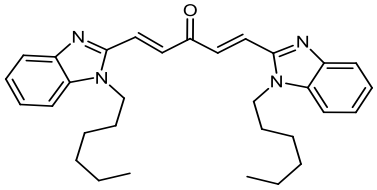
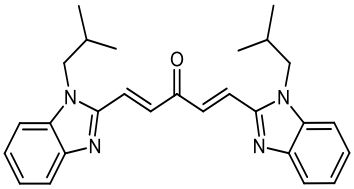
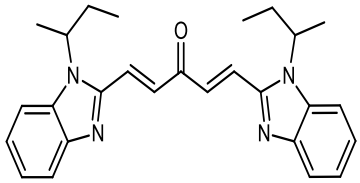
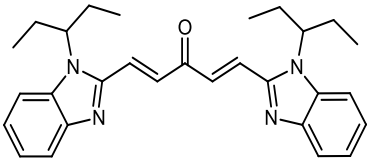
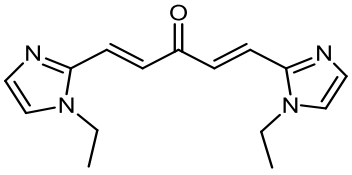
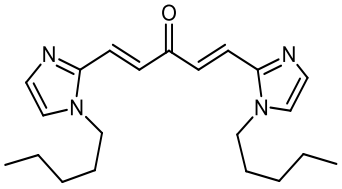
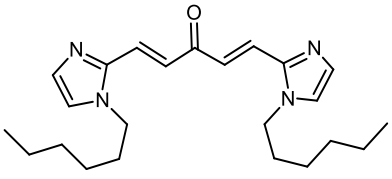
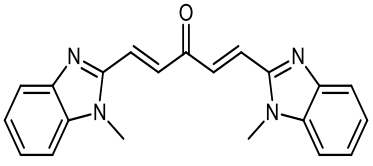
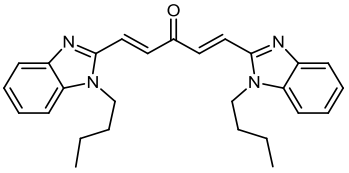
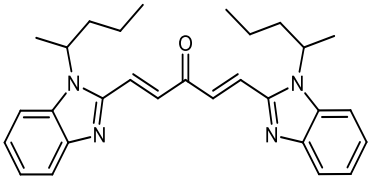
Cancer encompasses a diverse group of diseases that can develop in virtually any organ or tissue of the body. Cancer develops through a multi-step process that transforms normal cells into malignant tumor cells, often progressing from precancerous lesions. Long considered "a disease of the rich", cancer remains one of the world's most significant causes of mortality. According to the latest estimates from the International Agency for Research on Cancer (IARC), approximately 20 million new cases of cancer were detected internationally in 2022, and 9.7 million individuals succumbed to the disease [1]. The IARC also estimates that by 2050, the incidence of cancer worldwide is set to rise by 77%. The most common cancers in men are lung, prostate, colorectal, stomach, and liver cancers, while breast, colorectal, lung, cervical, and thyroid cancers are most common in women [1]. In our present study, we are interested in prostate cancer. The prostate is a gland of the male reproductive system whose volume increases with age. It lies below the bladder and in front of the rectum. It surrounds the beginning of the urethra, the channel through which urine and semen are evacuated. The prostate plays an important role in sperm production, producing a liquid called Spermatic fluid. The seminal vesicles, a pair of glands situated behind the bladder and above the prostate, are primarily responsible for producing the majority of the fluid that comprises semen. When initially normal prostate cells change and multiply in an uncontrolled way, forming a mass called a malignant tumor, this is called prostate cancer. At first, the mass is limited to the prostate. As the tumor progresses, it may grow larger and extend outside the prostate envelope. Cancer cells may break away from the cancer and travel via plasma or lymph spread to other parts of the body through blood vessels. This disease often progresses slowly, over several years. Prostate cancer develops without causing any particular symptoms. When the cancer is at an advanced stage, it may cause symptoms that raise suspicions of its presence, such as urinary tract infection, blood in the urine, urine retention, lower back pain or bone pain. For prostate cancer, family history has been identified as a risk factor. It has also been identified that men of Afro-Caribbean origin have an increased risk of developing this cancer. Occupational exposure to pesticides

* Corresponding author:
1997sageme@gmail.com (Georges Stéphane Dembele)
Received: Jul. 10, 2025; Accepted: Aug. 3, 2025; Published: Aug. 7, 2025
Published online at http://journal.sapub.org/chemistry

used in agriculture has also is known to contribute to the development of prostate cancer since the end of 2021 [2]. In Côte d'Ivoire, according to the national cancer control program (PNLCa), prostate cancer is the most common cancer among men, with nearly 2,757 new cases diagnosed, representing an incidence rate of 48.0 per 100,000 men in 2020 [3]. This mortality rate will continue to rise, as the latency of its symptomatology makes it a cancer that is often discovered late, and the clinical signs unfortunately reflect an advanced stage of the disease. To treat this cancer, several therapeutic methods are applied to patients, including surgery, radiotherapy, chemotherapy, targeted therapies, hormone therapy and immunotherapy. Castration and chemotherapy are the most frequently used because of the late stage of diagnosis. Treatment is tailored to each patient and depends on the tumor's potential to progress. Localized tumors can be treated curatively by surgery (total prostatectomy) or radiotherapy (external or brachytherapy). Locally advanced tumors are treated with a combination of radiotherapy and hormone therapy. Metastatic tumors can be treated with hormonal therapy. Chemotherapy and, increasingly, second-generation anti-androgens are used to treat castration-resistant forms. Chemotherapy remains the best alternative due to the more or less effective anti-cancer drug treatments based on docetaxel, mitoxantrone and cabazitaxel [4]. However, these molecules come up against resistant strains of prostate cancer. Hence the urgent need to find other anti-cancer agents with improved properties. The plant kingdom abounds in a number of natural compounds with efficient biological and medical properties. Curcumin is a phenolic compound identified in the rhizome of turmeric. Turmeric is a rhizomatous herbaceous plant that rhizomes have been used for many years in cultural Indian remedy and Asian cooking. Turmeric is part of the Asian diet, and it has been discovered that Asians have developed resistance to prostate cancer thanks to their consumption of turmeric [5]. As a result, scientists were able to identify curcumin as the main molecule responsible for turmeric's anti-prostate cancer activity [6]. Having shown a capacity to prevent prostate cancer, in 2000 curcumin was tested on cell culture systems in vivo and in vitro as a potential treatment for prostate cancer [7]. Like many other researchers, Rubing Wang et al. in 2015 investigated the anti-prostate cancer activity of curcumin derivatives. Their study, which focused on a series of curcumin analogs derived from (1E,4E)-1,5-di(1H-imidazol-2-yl)penta-1,4-dien-3-one, showed that these compounds effectively led to apoptosis of hormone-refractory metastatic PC-3 prostate cancer cells [8]. Therefore, in order to anticipate any resistance by designing curcumin derivatives with improved anti-prostatic activity, it is important to identify the origin of the anti-cancer properties of curcumin derivatives. This is the context of our study, the general aim of which is to model the anti-prostatic activity of a series of curcumin derivatives. The aim is specifically to found a quantitative structure activity relationship (QSAR) to explain the cytotoxicity of a series of sixteen curcumin derivatives whose anti-prostatic activities have been determined experimentally (Table 1) [8]. The QSAR methodology used in this work is a valuable tool for predicting and understanding the cytotoxicity of compounds for pharmaceutical interest.

Table 1. Molecular structures and cytotoxicity of DC compounds in the training and validation datasets employed for the QSAR model

CODE	STRUCTURE	IC_{50}
	Training set	
DC2		0.68
DC5		0.69
DC6		0.71
DC7		0.60
DC9		0.52
DC10		0.34
DC12		0.37

DC13		0.64
DC14		0.22
DC15		0.34
DC17		0.26
CODE	STRUCTURE	IC_{50}
	Validation set	
DC1		1.07
DC3		0.71
DC4		0.90
DC8		0.21
DC11		**0.40**
DC16		**0.30**

2. Materials and Experimental Procedures

Data collection methods

Modeling the anti-prostatic activity of curcumin derivatives requires the determination of descriptors. An infinite number of descriptors exist, including 2D and 3D descriptors. The 2D descriptors used in our model are derived from the compounds' 2D structure obtained with ChemSketch [9]. The 3D descriptors are determined from the optimized structure of the compounds. Optimization and frequency calculation are performed using Gaussian 09 software [10]. For these various calculation operations, density functional theory (DFT) is used at the B3LYP/6-31+G(d,p) level [10]. This level conceptual is a combination of the hybrid three-parameter Lee-Yang-Parr functional and the double split valence basis with polarization effects on heteroatoms. In addition, the multilinear regression method implemented in XLSTAT software version 2014 [11] was used to develop the QSAR model.

Molecular descriptors

Various physico-chemical descriptors can be used to develop QSAR models. However, in our study, we used a descriptor derived from the global reactivity of compounds and 3D and 2D geometric descriptors. Chemical hardness is a valuable tool for understanding and predicting the behavior of molecules in various chemical environments. It measures a molecule's resistance to electronic deformation during chemical reactions, expressing in other words a system's resistance to changes in its electron number. It is therefore a measure of a molecule's stability against nucleophilic and electrophilic attack. A hard molecule is less reactive, while a soft molecule is more reactive. Its expression is as follows:

$$\eta = IP - EA \quad (1)$$

With IP being the ionization potential, conversely the HOMO energy, and EA corresponds to the electron affinity, which is the opposite of the LUMO energy. This descriptor is determined from the ground-state structure of the compounds.

The geometric descriptors employed are the bond angle α(C-C=C) (Figure 1) formed by three carbon atoms, the

surface tension ($T_{Surface}$) and the density of the molecule. The descriptor bond angle α(C-C=C) is illustrated in the figure. Surface tension is a fundamental property of liquids that influences many physical and chemical phenomena. Surface tension is fundamental to biochemical reactions such as respiration (gas exchange in pulmonary alveoli) or substance transport in blood vessels. Density is a fundamental property that helps us understand the characteristics of molecules and their behavior in various media.

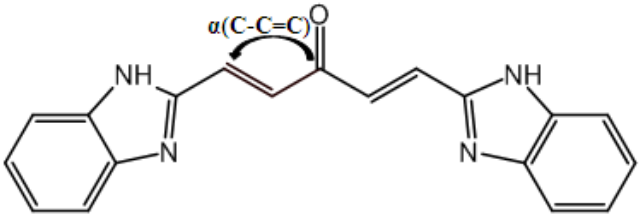


Figure 1. Geometric descriptor for curcumin analogues

Satistical analysis

The process of developing a QSAR model relies on a data analysis method that establishes a quantitative relationship between descriptors and the biological property under study. In this work, Multiple Linear Regression (MLR) is the chosen analysis method. This method is adapted to show a linear association between a dependent variable Y (in our case, the cytotoxic potential pIC_{50} and a series of n independent variables X_i (here, the descriptors) [12]. The aim is to obtain a mathematical equation of the form:

$$pIC_{50} = a_0 + a_1x_1 + \ldots\ldots + a_nx_n \quad (2)$$

Where a_i (i =1,…, n) are the coefficients of the regression parameters and a_0 is the constant of the model equation.

The multiple linear regression method proved useful for expressing the IC_{50} inhibitory activity of DC as a function of descriptors. It is essential that the descriptors are independent of each other to guarantee the efficiency of the model. To achieve this, the partial correlation coefficient a_{ij} between the values of descriptors i and j should be less than 0.70 ($a_{ij} < 0.70$) [13]. It is determined according to the following relation.

$$a_{ij} = \frac{COV(X_i,X_i)}{Var(X_i)} \quad (3)$$

The quality of a QSAR model is closely linked to a set of statistical indicators such as the coefficient of determination R^2, the standard deviation S, the cross-validation coefficient Q^2_{cv} and the Fischer test. Validation of a quantitative structure-activity relationship (QSAR) model relies on a number of statistical parameters. These parameters enable internal validation of the QSAR model. The coefficient of determination R^2 determines the percentage of data variance explained by the model. An R^2 close to 1 indicates a good fit. The cross-prediction coefficient Q^2 assesses the model's predictive ability. It is calculated using cross-validation methods. A high, positive Q^2 indicates a good predictive model. Standard deviation S measures the mean error of model predictions. A lower standard deviation indicates a better fit. The Fischer F-test evaluates the overall predictive power of the model. A high F-statistic indicates that the model explains the data variance well. Moreover, the p-value indicates the significance of the coefficients of the explanatory variables. A p-value below the 0.05 threshold suggests the variable has a statistically significant effect. The combination of these parameters enables us to judge the quality and robustness of a QSAR model, guaranteeing reliable and relevant predictions.

- The coefficient of determination R^2[14] is calculated using:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \text{ with } R = \sqrt{R^2} \quad (4)$$

TSS: Total Sum of Squares:

$$TSS = \sum(Y_{i,exp} - \overline{Y}_{exp})^2 \quad (5)$$

ESS: Extended Sum of Squares:

$$ESS = \sum(Y_{i,cal} - \overline{Y}_{exp})^2 \quad (6)$$

RSS: Residual Sum of Squares:

$$RSS = \sum(Y_{i,exp} - Y_{i,cal})^2 \quad (7)$$

❖ Adjusted coefficient of determination $R^2_{ajusté}$ [15]

Unlike R^2, this coefficient is used to assess the robustness of the model, considering the number of descriptors used in the multiple regression model

$$R^2_{ajusté} = 1 - \frac{(n-intercept)(1-R^2)}{n-p} \quad (8)$$

❖ Fisher-Snedecor coefficient F [16]

The Fisher-Snedecor coefficient is used to assess the overall significance of the linear regression and is correlated with the coefficient of determination by the following relationship:

$$F = \frac{R^2}{1-R^2}\frac{n-p-1}{p} \quad (9)$$

- Cross-validation criterion PRESS [17]

PRESS (Prediction Sum of Squares) is a criterion that evaluates the model's prediction errors and is calculated according to the following relationship:

$$PRESS = \sum(y_{i,exp} - y_{i,préd})^2 \quad (10)$$

To select a model with good predictive power, we use PRESS as a criterion, aiming for the smallest possible PRESS value.

- Cross-validation coefficient Q^2_{LOO} [17]

It quantifies the predictive power of the model on training data.

$$Q^2_{LOO} = 1 - \frac{\sum(y_{i,exp} - y_{i,préd})^2}{\sum(y_{i,exp} - \overline{y}_{exp})^2} = 1 - \frac{PRESS}{TSS} \quad (11)$$

- External validation coefficient Q^2_{ext} [17]

This criterion evaluates the model's ability to make accurate predictions on the test set.

$$Q^2_{ext} = 1 - \frac{n}{n_{ext}}\frac{PRESS\ (test)}{TSS} \quad (12)$$

➢ Standard deviation [18]

This is a statistical indicator that reveals how the data is distributed around the mean. When this value approaches 0, it reflects better model fit and greater prediction reliability.

$$s = \sqrt{\frac{RSS}{n-p-1}} \quad (13)$$

$$s_{PRESSS} = \sqrt{\frac{PRESS}{n-p-1}} \quad (14)$$

In this configuration, p corresponds to the number of descriptors (independent variables), n to the number of molecular samples in the training set, with n-p-1 representing the degree of freedom. In addition to these statistical indicators, criteria such as Kubinyi's are needed to validate a model.

➢ Kubinyi criterion FIT [19]

The ITF measures the strength of model; a smaller ITF value signifies that model is stronge and includes a greater number of variables.

$$FIT = \frac{(n-p-1)}{(n+p)^2}\frac{R^2}{(1-R^2)} \quad (15)$$

It is also necessary to show that the QSAR model is not due to chance. This is done using the parameters of Roy and al.

❖ Parameter from Roy and al R_p^2, r_m^2 et Δr_m^2 [20]

By evaluating this parameter, we can tell whether the pattern is the product of chance or not. A value is larger than 0.5indicates that the model is significant and not by chance.

$$R_p^2 = R\sqrt{R^2 - R_r^2} \quad (16)$$

With R_r^2, the mean value of the R_{ri}^2 of the models generated using random parameters. For the prediction to be acceptable, the value of the metric Δr_m^2 must be less than 0.20 when that of $\overline{r_m^2}$ is greater than 0.50.

$$\overline{r_m^2} = \frac{(r_m^2 + r'^2_m)}{2} \quad (17)$$

$$\Delta r_m^2 = |r_m^2 - r'^2_m| \quad (18)$$

$$\text{Here, } r_m^2 = r^2(1 - \sqrt{r^2 - r_0^2})$$

$$\text{et } r_m'^2 = r^2(1 - \sqrt{r^2 - r_0'^2}) \quad (19)$$

The parameters r^2 and r_0^2 represent the coefficients of determination Between observed and predicted values, with and without interception respectively. In addition to internal validation, external validation is also required. It is carried out by using the ratio (theoretical activity) / (experimental activity) and the Tropsha criteria.

➢ Tropsha criteria [20] [21]

There are five Tropsha criteria:

❖ Criteria 1: $R_{ext}^2 > 0{,}70$

❖ Criteria 2: $Q_{ext}^2 > 0{,}60$

❖ Criteria 3: $\frac{|R_{ext}^2 - R_0^2|}{R_{ext}^2} < 0{,}1$

and k=0,9475 avec 0,85<k<1,15 (20)

❖ Criteria 4: $\frac{|R_{ext}^2 - R_0'^2|}{R_{ext}^2} < 0{,}1$

and k'= 1,0521 avec 0,85<k'<1,15 (21)

❖ Criteria 5: $|R_{ext}^2 - R_0^2| < 0{,}3$ (22)

R_{ext}^2: Coefficient of determination for the molecules in the test series;

R_0^2: Coefficient of determination of the regression model for the test set;

$R_0'^2$: Coefficient of determination of the regression among experimental and predicted values for the test series;

k: Slope of the correlation line (predicted values vs experimental values);

k': Slope of correlation line (experimental values vs. predicted values).

QSAR model applicability domain:

The applicability domain of a QSAR model is established using the threshold lever [22]. The concept of leverage is essential for assessing the validity and robustness of QSAR models, helping to identify data points that could disproportionately influence model results. Leverage" refers to a parameter that measures how much a specific data point influences model predictions. A point with high leverage has a disproportionate impact on model results. Leverage is often calculated from the model's matrix of explanatory variables. It is related to the interval of a point from the mean of the points in the variable space. For observation i, it should be calculated as follows:

$$h_i = x_i(X^T X)^{-1} x^T \quad (i=1,..,n) \quad (23)$$

x_i is the row vector of compound i descriptors

X is the model matrix deduced using the information contained in the training data.

The superscript T denotes the transpose operation, which interchanges the rows and columns of a matrix or vector.

The critical leverage value h* is generally set at $\frac{3(k+1)}{N}$ [23], where **N** signifies the sample size of the training set, and k indicates the number of independent variables used in the model. A compound is considered outside the model's applicability domain if its residual and leverage exceed the critical h* value.

Normality test

Normality tests are statistical techniques used to evaluate the hypothesis that a given sample originates from a normally distributed population. A wide range of parametric statistical tests, such as t-tests and ANOVA, are based on the normality assumption, which stipulates that the data are normally distributed. Testing this assumption is essential to guarantee the validity of the results. Normality testing helps to understand the probability distribution of the data, and assists to select the appropriate statistical methods for analysis. Normality tests are essential for validating statistical analysis hypotheses. If the data fails to meet the normality assumption, non-parametric statistical tests, which do not rely on the assumption of a normal distribution, can be employed. In practice, for example, verification of the normality of residuals in linear regression is rarely carried out, although it

is essential to guarantee the reliability of confidence intervals for parameters and predictions. This normality of residuals can be verified by analyzing certain graphs, or by using a normality test that relies on the independence of residuals using certain graphs [24].

Shapiro-wilk test

The Shapiro-Wilk test [25] is a statistical procedure used to test the hypothesis that a given sample originates from a normally distributed population. The test is based on two competing hypotheses: the null hypothesis, which assumes that the data follows a normal distribution, and the alternative hypothesis, which suggests that the data does not follow a normal distribution.

- The null hypothesis(H_0) posits that the data are drawn from a population that is normally distributed.
- The alternative hypothesis (H_1) posits that the data are not drawn from a population that is normally distributed.

Generally recommended for small sample sizes (n < 50), but can be used up to n = 2000. The test uses the correlation coefficients between the observed data points and the theoretical values of a normal distribution. The result is a W statistic, which varies between 0 and 1. The p-value obtained is compared with a threshold, often set at 0.05:

- p-value < 0.05: Rejection of the null hypothesis, data are not normally distributed.
- p-value ≥ 0.05: No rejection of the null hypothesis, data can be considered as normally distributed.

Application of this test to the predicted cytotoxic activity values of the established model produced the following results in XLSTAT software [11].

Probability-probability graph

A probability-probability plot (Q-Q plot, or quantile-quantile plot) is a graphical method used to compare the empirical distribution of a dataset to a theoretical distribution, such as the normal distribution. It is often used in conjunction with normality tests such as the Shapiro-Wilk test. The Q-Q plot is utilized to test for normality. The Q-Q plot visually compares the empirical quantiles of the dataset to the theoretical quantiles of a specified distribution. with those of a theoretical distribution. The Q-Q Plot is constructed in two stages: calculating quantiles and plotting points: On a graph, the quantiles of the observed data are placed on the y-axis and the quantiles of the theoretical distribution on the x-axis. If the points are collinear, they all fall along the same straight line (usually the diagonal), this indicates that he data follow a normal distribution pattern, characterized by a symmetric, bell-shaped curve centered around the mean.

Contribution of an explanatory variable to the prediction of an activity

The contribution of the explanatory variable X_i, denoted by C_{X_i}, to the prediction of activity Y [24] [25] is evaluated by Student's t test, allowing us to determine its relevance in the model. It measures the importance or contribution of variable X_i in the QSPR/QSAR model according to the following relationship:

$$C_{X_i}=\frac{\lfloor t(X_i)\rfloor}{\sum\lfloor t(X_i)\rfloor}\times 100 \quad (24)$$

The contribution is expressed as a percentage (%) with:

$\lfloor t(X_i)\rfloor$: Absolute test value of the variable.

$\sum\lfloor t(X_i)\rfloor$: Sum of the absolute values of the t-tests of all the Xi variables in the model.

3. Resultats Et Discussion

A series of sixteen curcumin derivatives are being investigated for their anti-prostatic activity. In order to carry out our QSAR study, this experimental database was split into two groups. The first group, comprising 11 molecules (i.e. 2/3 of the database), constitutes the training set. The second group, comprising 5 molecules (1/3 of the database), is called the validation set. The values of the descriptors and the cytotoxic activity of the molecules are given in Table 2.

Table 2. Values of molecular descriptors obtained after optimization

Code	$\eta(ev)$	$\alpha_{(C-C=C)}$ (Å)	$T_{Surface}$ (dyne/cm)	Densité (g/cm^3)	$pIC_{50(exp)}$
Training set					
DC2	1.774	120.304	40.400	1.060	6.167
DC5	1.784	120.018	37.700	1.040	6.161
DC6	1.773	120.248	37.700	1.030	6.149
DC7	1.773	120.118	37.700	1.030	6.222
DC9	1.648	120.373	46.100	1.640	6.284
DC10	1.651	120.138	44.800	1.140	6.469
DC12	1.660	119.804	42.900	1.600	6.432
DC13	1.663	131.359	42.200	1.080	6.194
DC14	1.656	120.070	41.900	1.120	6.658
DC15	1.657	120.102	41.900	1.120	6.469
DC17	1.655	120.646	41.200	1.100	6.585
Validation set					
DC1	1.777	120.265	42.500	1.100	5.971
DC3	1.774	120.247	39.700	1.040	6.149
DC4	1.774	120.228	39.100	1.020	6.046
DC8	1.555	120.904	47.800	1.200	6.678
DC11	1.650	120.178	43.800	1.120	6.398
DC16	1.657	120.046	41.200	1.100	6.523

Interdependence of molecular descriptors

The descriptors used in a QSAR model must be independent of each other. This interdependence is measured using the partial correlation coefficient a_{ij}. An extract from the correlation matrix is shown in Table 3. The various values of the partial correlation coefficient between two descriptors are less than 0.7 ($a_{ij} < 0.7$). The different descriptors are therefore independent of each other.

Table 3. Correlation matrix between descriptors

Variables	η	αc-c=c	TSurface	Densité
η	1.000			
αc-c=c	-0.203	1.000		
TSurface	-0.851	0.109	1.000	
Densité	-0.508	-0.158	0.686	1.000

Principal Component Analysis (PCA) and Hierarchical Clustering (HCA)

- Principal component analysis (PCA)

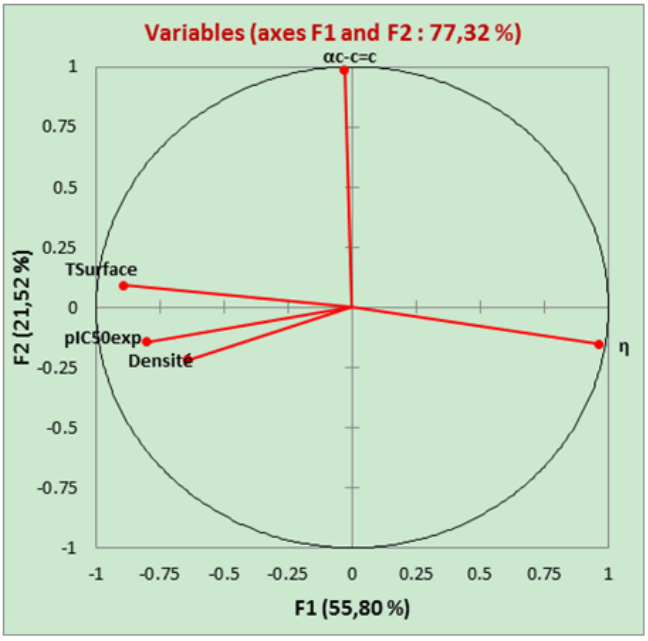


Figure 2. Correlation circle

Principal component analysis (PCA) is applied to identify the relationship between the different descriptors [26]. In our study, principal component analysis was identify to data from the 17 DC compounds, based on the four descriptors. The primary two axes in the factor space, F1 and F2, explain a significant proportion of the total variability (77.32%), suggesting that these two axes are sufficient to represent the information contained in the data matrix.

- Hierarchical ascending classification (HAC)

The hierarchical ascending classification (HAC) in fig 3 illustrates a distribution of DCs into four (04) classes according to their affinity. The four (04) classes are constituted as follows: C1 (2; 3; 4; 5; 6; 7; 13), C2 (8; 9), C3 (10; 11) et C4 (1; 12; 14; 15; 16; 17).

Study of the model developed

The QSAR model we developed shows a linear relationship between the descriptors and the opposite of the logarithm to base 10 of the inhibitory concentration IC50 of the compounds studied (pIC_{50}). The regression equation for our model is shown below.

$$pIC_{50_{théo}} = 18.78652 - 4.25231 * \eta - 0.03003 * \alpha(C\text{-}C\text{=}C) - 0.03074 * TSurface - 0.25950 * Densité \quad (25)$$

Table 4. Statistical indicators associated with the RQSA model developed

Number of compounds in test set N	11
Correlation coefficient R	0.960
Determination coefficient R^2	0.922
Standard deviation S	0.068
Fischer test F	17.027
FIT	0.031
p-value	< 0.0001
TSS	0.339
ESS	0.332
Confidence level α	95%

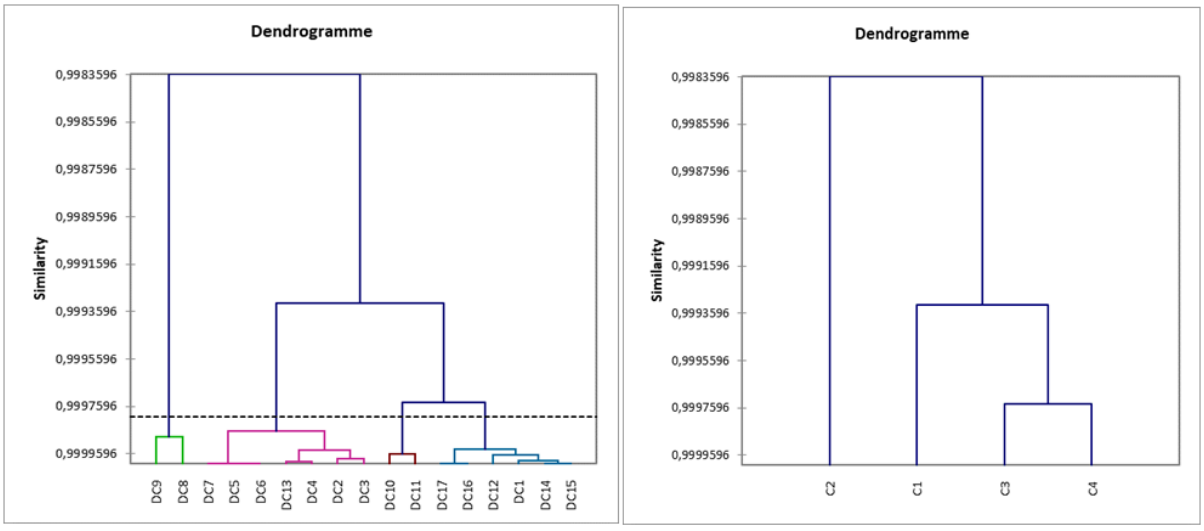


Figure 3. DC dendrograms

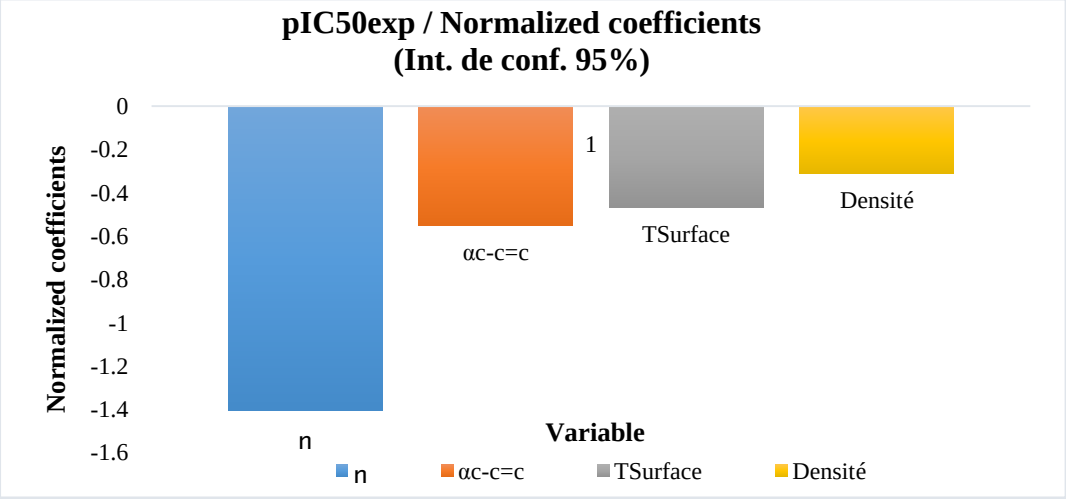


Figure 4. Standardized descriptor coefficients

In the model developed, the cytotoxic potential pIC_{50} depends on hardness, angle, surface tension and density. The model's regression equation shows negative coefficients for all variables namely η, α(C-C=C), TSurface and density. These negative coefficients indicate that cytotoxic potential increases as the values of these descriptors decrease. In other words, inhibitory concentration is enhanced for a smaller value of these descriptors. Analysis of the statistical parameters associated with the QSAR model indicates a correlation coefficient R of 0.960. This high value highlights a strong linear correlation between inhibitory activity and the descriptors used in the model. Moreover, 92.22% of the experimental variable pIC_{50} this is elucidated by the the model incorporates various descriptors (R^2 = 0.922). Furthermore, a standard deviation (S= 0.068) tending towards 0 indicates a good fit and high predictive reliability. The high Fischer test value (F = 17.027) is well above the critical value derived from the Fischer table, confirming the significance of the model. Also, parameter values such as the experimental variance (TSS = 0.339) and the variance due to the model ESS = 0.332 indicate that there is less variability in the data.

Table 5. Contribution of descriptors

Descriptors	**η**	$\alpha_{(C-C=C)}$	$\mathbf{T_{Surface}}$	Densité
Contribution in %	43.47	31.17	12.38	12.99

The contribution of the various descriptors to the development of the model is illustrated in the figure. Examination of the fig 4 and contribution values (table) shows that hardness and bond angle between carbon atoms are the main descriptors influencing the cytotoxic activity of curcumin derivatives.

Internal model validation

A QSAR model must satisfy a certain number of criteria: this is known as validation. This stage of the QSAR methodology is divided into two operations: internal validation and external validation. In our study, the model's internal validation was performed using (Leave-One-Out (LOO). Randomization was also used to show that the model is not due to chance.

• Leave-One-Out procedure

Table 6. The statistical metrics used for the leave-one-out (LOO) internal validation of the model

N	$\mathbf{Q^2_{LOO}}$	$\bar{r}^2_m(LOO)$	$\Delta\mathbf{r^2_m(LOO)}$	PRESS	S_{PRESS}
11	0.919	0.886	0.063	0.027	0.068

The value of Q^2_{LOO} = 0.919, which is greater than 0.50, indicates that 91.90% the molecules in the test set exhibit their cytotoxic potential perfectly predicted by the model. The model displays a high predictive ability in relation to the molecules included in the training set for model training. Consequently, this result indicates that the model exhibits low sensitivity to the operation of isolating a molecule before reintegrating it into the learning set (Leave-One-Out). The value of $\bar{r}^2_m(LOO)$ is greater than 0.50 while that of $\Delta r^2_m(LOO)$ is less than 0.2. Thus, the model is considered sufficient for use in predicting cytotoxic potential. The low value of the S_{PRESS} error indicates a good model.

• Y-randomization test

To ensure that the model is not random, we carried out a randomization test, which consisted in performing a circular permutation in the training set. The table 7 displays the values of the randomization parameters following 10 iterations.

Table 7. The first 10 iterations of the Y-randomization test

ITERATION	1	2	3	4	5	6	7	8	9	10
$\mathbf{R^2_{ri}}$	0.129	0.131	0.153	0.197	0.145	0.180	0.190	0.037	0.127	0.129

Table 8. Average values of Y-randomization parameters

Randomized parameter	R_r^2	s_r	F_r	R_P^2
Average value	0.142	0.196	0.248	0.848

An examination of the data presented in Table 8 reveals that the mean value of R_r^2 is low (R_r^2=0.142), Showing that the equation for the regression line explains only 14.20% of the distribution of points (cytotoxic potential). The high standard deviation (s_r=0.196) also points to a dispersion of points around the regression line. F_r The value of Roy's parameter R_p^2, 0.848 (exceeding 0.50), confirms that the model's performance is statistically significant and not attributable to random variation.

External model validation

In addition to internal validation, external validation is also required. This is carried out by using the molecules in the test series. Statistical parameters for external validation are listed in Table 9.

Table 9. Statistical parameters for external model validation

n	R_{ext}^2	Q_{ext}^2	$\bar{r}_m^2(test)$	Δr_m^2(test)	PRESS (test)
6	0.973	0.636	0.941	0.027	0.0436

Examination the information provided in Table 9 indicates the presence of certain qualities in the model, such as above-average the statistical best capability to predict. Indeed the value of Q_{ext}^2 which is 0.636 indicates a prediction rate of cytotoxic potential of 63.60%. In order to $\bar{r}_m^2(test)$, It holds a value higher than 0.50 while that of the other is Δr_m^2 (test) is less than 0.2 we can therefore conclude that the model is acceptable to estimate the future value of the cytotoxic the capability of the molecules in the validation set. To further assess the robustness of our model, the five (05) Tropsha criteria are calculated.

- Checking Tropsha criteria

Criteria 1: $R_{ext}^2 = 0.973 > 0.70$

Criteria 2: $Q_{ext}^2 = 0.636 > 0.60$

Criteria 3: $\frac{|R_{ext}^2 - R_0^2|}{R_{ext}^2} = 0.00041 < 0.1$ and $k = 1.0116$ avec $0.85 < k < 1.15$

Criteria 4: $\frac{|R_{ext}^2 - R_0'^2|}{R_{ext}^2} = 0.0024 < 0.1$ and $k' = 0.9885$ avec $0.85 < k' < 1.15$

Criteria 5: $|R_{ext}^2 - R_0^2| = 0.0004 < 0.3$

It is clear that all five (05) Tropsha criteria are met. Consequently, the model developed is extremely effective in predicting the potential for the initial reduction.

Correlation between model-predicted values and experimental values

Theoretically calculated pIC_{50} values and experimentally determined pIC50 values are compared using the correlation curve (figure 5).

Table 10. pIC_{50} (*exp*) and pIC_{50} (theo) model values and their ratios

CODES	$pIC50_{exp}$	$pIC50_{théo}$	$pIC50_{théo}/pIC50_{exp}$
DC2	6.167	6.112	0.999
DC5	6.161	6.168	1.001
DC6	6.149	6.210	1.010
DC7	6.222	6.214	0.999
DC9	6.284	6.320	1.006
DC10	6.469	6.485	1.003
DC12	6.432	6.396	0.994
DC13	6.194	6.192	1.000
DC14	6.658	6.559	0.985
DC15	6.469	6.557	1.014
DC17	6.585	6.576	0.999
DC1	5.971	6.026	1.009
DC3	6.149	6.140	0.999
DC4	6.046	6.165	1.020
DC8	6.678	6.762	1.013
DC11	6.398	6.522	1.019
DC16	6.523	6.584	1.009

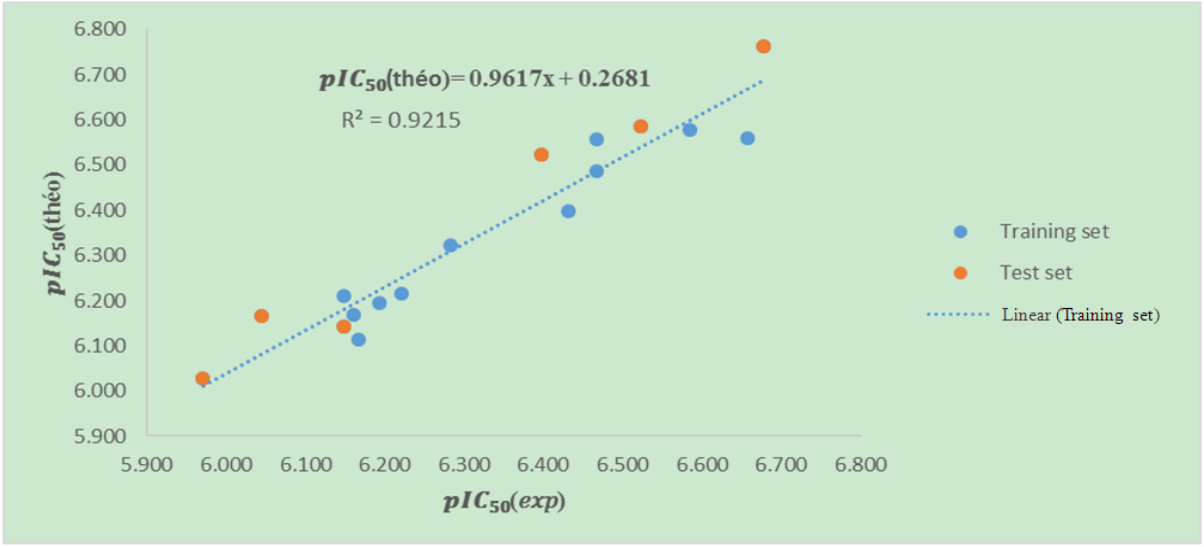


Figure 5. Correlation curve between $pIC5(exp)$-$pIC50$(theo) model

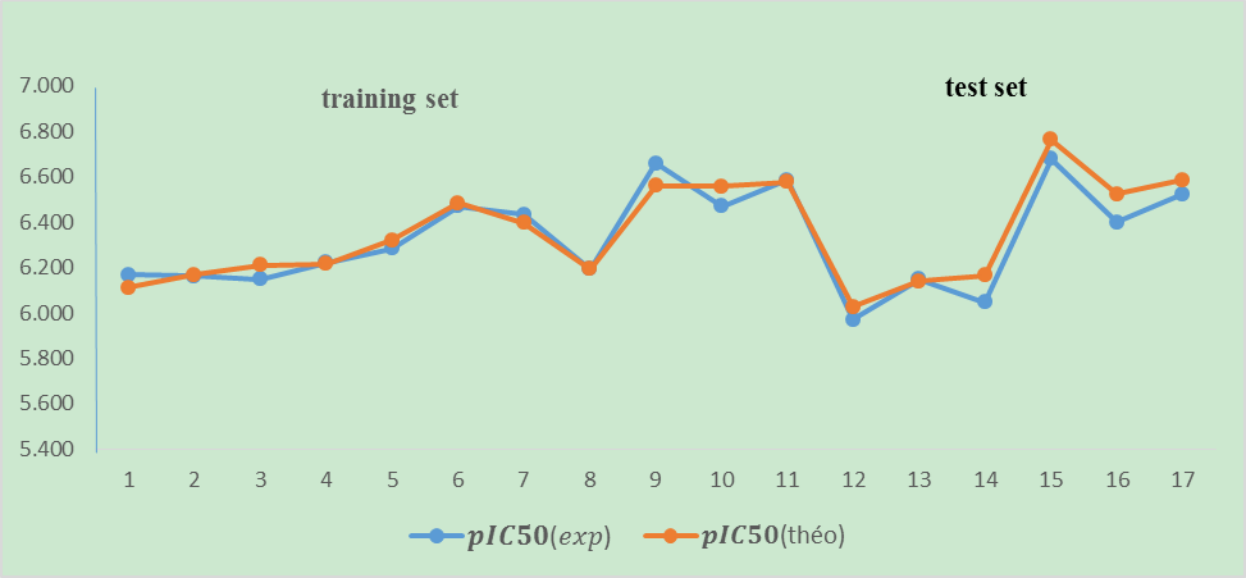


Figure 6. The match between the model-predicted values and the experimental results

Fig 5 shows that the data points generally tend towards approximate the regression line. The figure thus reveals a good linear correlation between experimental values and the cytotoxic potential values predicted by the model.

In addition, the alignment of the predicted values with the experimental measurements of anti-prostate activity is illustrated in fig 6. Examination of the curve indicates that the theoretical values predicted by the progression of the model values mirrors that of the experimental values. This demonstrates the consistency between the theory employed and the database used for model design.

Model normality tests

- **Shapiro-Wilk test $pIC50$ (theo)**

Table 11. Shapiro-Wilk test parameter values

W	P-value	alpha
0.929	0.238	0.05

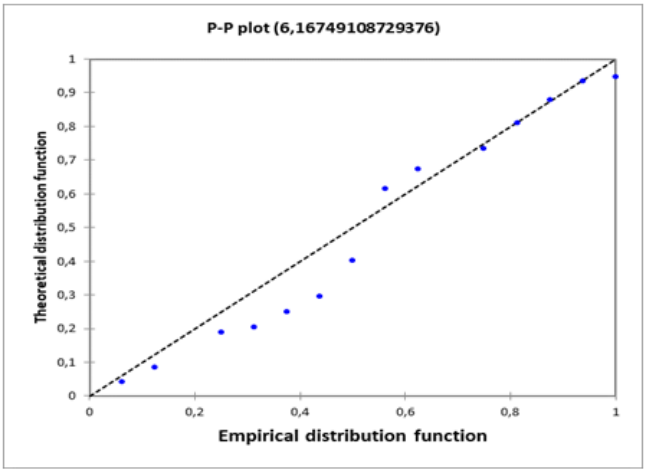


Figure 7. P-P plot ($pIC5$(theo)) of the mode

The computed p-value (Table11) exceeds the alpha = 0.05 significance level. This result shows that the cytotoxic potential values predicted by our model adhere to a normal distribution. This normal distribution is corroborated by the observed distribution of the scatter plot along the first bisector in fig 7.

- **Durbin-Watson test pIC_{50} (theo)**

Table 12. Durbin-Watson test parameter values $pIC5$(theo)

W	P-value	alpha
1.460	0.197	0.05

The values of the parameters used to perform the Durbin Watson Test are given in Table 12. The computed p-value exceeds the significance threshold of 0.05. Thus, these residuals do not contain any information capable of influencing the cytotoxic potential prediction model. The random distribution of the scatter plot in figure 8 corroborates this interpretation.

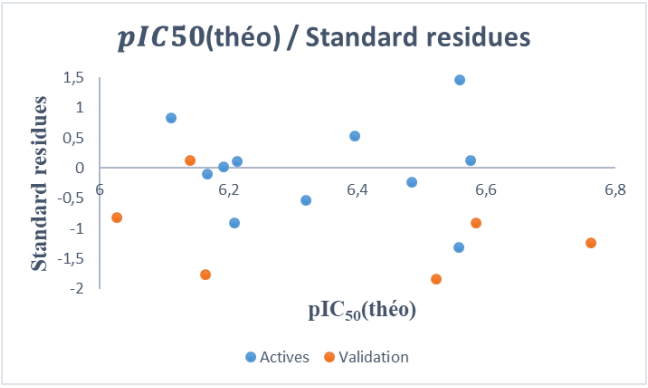


Figure 8. Normalized residuals graph=f(pIC_{50} (theo)) of the model

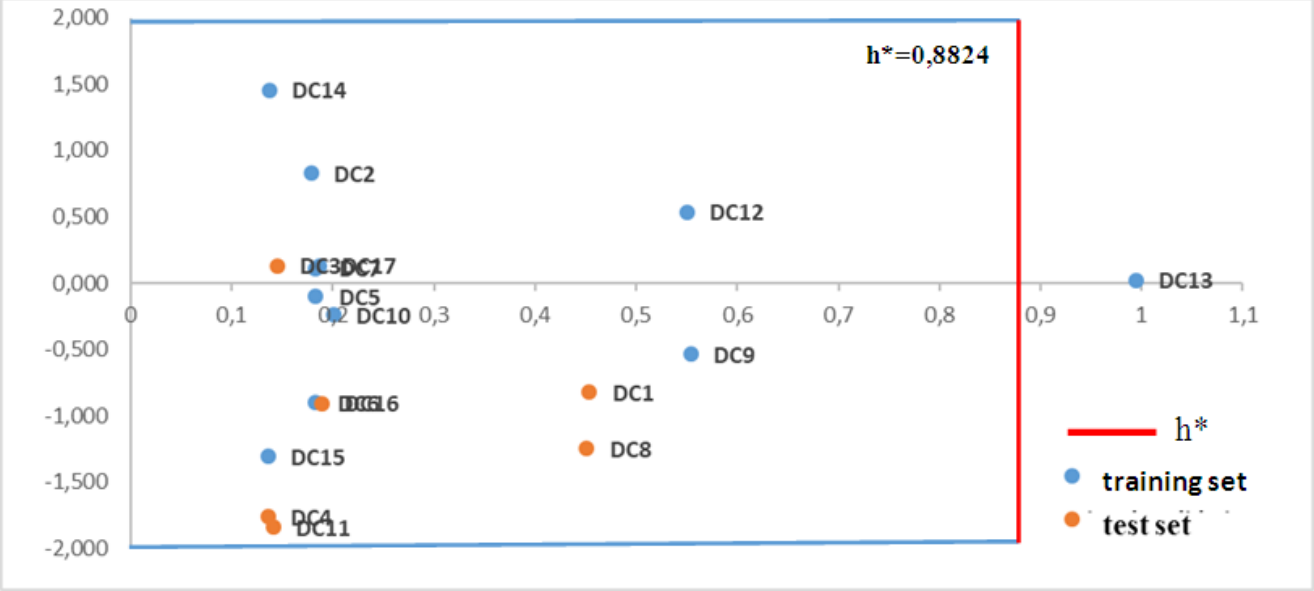


Figure 9. Williams diagram of the model

Model applicability

The activity of a curcumin derivative can be predicted by our model if and only if this molecule belongs to the same domain of applicability. The Domain of Applicability (DA) was determined by analysis of the Williams diagram in figure 9.

Upon reviewing the Williams diagram, it is evident that the data points follow a specific pattern, which suggests that the standardized residuals for both the training and validation sets do not exceed 3 standard deviations (3σ) in absolute value. The model's ability to match the observed data is further enhanced by the absence of outliers [27]. Given that the data exhibit a normal distribution the leverage effect with an interval of ±3 standard deviations is suitable. This corresponds to a confidence level of around 99.7%, in order to capture almost all observations [27]. All levers, with the exception of observation DC13 are below the defined threshold $h^* = 0.8824$. Although this last observation presents an atypical value, it does not necessarily constitute an anomaly for the model. Influential observations, such as those identified by high leverage, can improve model accuracy provided their impact on residuals is limited. The results obtained with the RQSA model indicate a good fit to the data, with no evidence of outliers likely to bias the analyses. Cross-validation and applicability domain analysis demonstrate that the model is robust and can be used to make reliable predictions on new DC molecules.

4. Conclusions

Curcumin is a compound of interest in the development of cytotoxic agents effective against prostate cancer. The aim of this work was to rationalize the anti-prostate activity of curcumin derivatives with a view to improving it. A quantitative structure-activity relationship was established for a series of sixteen curcumin derivatives. The model developed using the multilinear regression method is a function of hardness η, angle α(C-C=C), surface tension TSurface and density, which help to explain the property. In addition, anti-prostate activity is intimately related to hardness and angle α(C-C=C). The model is accredited with good static indicators (R = 0.960; R^2 = 0.922; F = 17.027; FIT = 0.031) highlighting excellent predictive power. Also, internal and external validation of the model elucidated its robustness and predictive power. This model is not due to chance, and follows a normal distribution law. Thus, as part of the process of designing cytotoxic agents with improved activity, this elaborate model is suitable forecasting the activity of curcumin derivatives not yet synthesized or whose activity has not yet been determined. As a follow-up to this work, we plan to carry out molecular docking to explain the activity of these compounds in interaction with the protein responsible for prostate cancer.

REFERENCES

[1] Le cancer augmente au niveau mondial, au milieu des besoins croissants en services. Accessed: Mar. 08, 2025. [Online]. Available: https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services.

[2] Burns C. J., Juberg D. R., Cancer and occupational exposure to pesticides: an umbrella review, *Int. Arch. Occup. Environ. Health*, 94(5): 945–957 (2021).

[3] "Epidémiologie | pnlca." Accessed: Mar. 08, 2025. [Online]. Available: https://www.pnlca.org/copy-of-cancer-en-cote-d-voire-2.

[4] Sun B., Lovell J. F., Zhang, Y.,Current development of cabazitaxel drug delivery systems, *Wiley Interdiscip. Rev. Nanomedicine Nanobiotechnology*, 15(2): 1–26 (2023).

[5] Tabanelli R., Brogi S., Calderone V., Improving curcumin bioavailability: Current strategies and future perspectives, *Pharmaceutics*, 13(10) (2021).

[6] Gupta B., Sharma P. K., Malviya R., Mishra P. S., Curcumin and Curcumin Derivatives for Therapeutic Applications: In vitro and In vivo Studies, *Curr. Nutr. Food Sci.*, 20(10): 1189–1204 (2024).

[7] Yousefnezhad M., Babazadeh M., Davaran S., Akbarzadeh A., Pazoki-Toroudi H., Preparation and in-vitro evaluation of PCL–PEG–PCL nanoparticles for doxorubicin-ezetimibe co-delivery against PC3 prostate cancer cell line, *Chem. Rev. Lett.*, 7(2): 159–172(2024).

[8] Wang R., Structure-Activity Relationship and Pharmacokinetic Studies of 1,5-Diheteroarylpenta-1,4-dien-3-ones: A Class of Promising Curcumin-Based Anticancer Agents, *J. Med. Chem.*, 58(11): 4713–4726 (2015).

[9] Acdllabs Advanced Chemistry Development, "ACD Chemsketch," 2010, *1994*: 10.0.

[10] H. B. S. et G. E. S. M. J. Frisch, G. W. Trucks, Gaussian 09, Revision A.02. [Online]. Available: Gaussian, Inc., Wallingford CT, 2009.

[11] Addinsoft, XLSTAT, 2014, *1995*: Version 2014.5.03.

[12] Chukwuemeka P. O., Predictive hybrid paradigm for cytotoxic activity of 1,3,4-thiadiazole derivatives as CDK6 inhibitors against human (MCF-7) breast cancer cell line and its structural modifications: rational for novel cancer therapeutics, *J. Biomol. Struct. Dyn.*, 40 (18): 8518–8537 (022).

[13] Sékou D., Bamba F., Affoué Lucie B., Gbèdodé Wilfried., Assongba Gaston K., El-Hadji Sawaliho B., Study by Quantum Chemical of Relationship between Electronic Structure and SecA Inhibitory Activity of a Series 5-cyano Thiouracil Derivatives, *J. Mater. Phys. Chem.*, 10(2): 43–48 (2022).

[14] Konate F., Diarrrassouba F., Dembele G. S., Guy-Richard Koné M., Konaté B., Ziao N., Elaboration of a Predictive Qsar Model of the Anti-Paludial Activity of a Series of Dihydrothiophenone Molecules at Theory Level B3LYP/ 6-31G (d, p), *Chem. Sci. Int. J.*, 30(8): 1–12 (2021).

[15] Diarrassouba F., Bamba K., Koné M., Kouamé K. K. R., Determination of molecular descriptors influencing the first reduction potential of a family of Tetracyanoquinodimethane molecules at HF / 6-31G (d, p) theory level 6 186–211 (2022).

[16] Songuigama C., QSAR, Docking Studies and in Silico Admet Prediction of 1, 10- Phenanthrolinone Derivatives with Antitubercular Activities, 15 17–25 (024).

[17] Koné M. G.-R., Modeling of a Series of Dihydropyrazole Derivatives with Antiproliferative Activity by Quantum Chemical Methods, *Chem. Sci. Int. J.*, 32(4): 24–38, (2023).

[18] N'dri J S., Quantitative Structure-Activity Study against Plasmodium falciparum of a Series of Derivatives of Azetidine -2-Carbonitriles by the Method of Density Functional Theory, *Mediterr. J. Chem.*, 11(2): 162 (2021).

[19] De P., Kar S., Ambure P., Roy K., Prediction reliability of QSAR models: an overview of various validation tools, *Arch. Toxicol.*, 96(5): 1279–1295(2022).

[20] Pal R., Patra S G., Chattaraj P K., Quantitative Structure–Toxicity Relationship in Bioactive Molecules from a Conceptual DFT Perspective, *Pharmaceuticals*, 15(11) (2022).

[21] Soufi H., Multi-combined QSAR, molecular docking, molecular dynamics simulation, and ADMET of Flavonoid derivatives as potent cholinesterase inhibitors *J. Biomol. Struct. Dyn.*, 42(12): 6027–6041(2024).

[22] Dutschmann T M., Schlenker V., Baumann K., Chemoinformatic regression methods and their applicability domain, *Mol. Inform.*, 43(7): 1–24(024).

[23] Moussaoui M., Design and Optimization of Quinazoline Derivatives as Potent EGFR 2 Inhibitors for Lung Cancer Treatment: A Comprehensive QSAR, 3 ADMET, and Molecular Modeling Investigation, (2024).

[24] Karadžić Banjac M., Kovačević S., Podunavac-Kuzmanović S., Jevrić L., Chemometric Modeling of Bioconcentration Factor of 6-Chloro-1,3,5-Triazine Derivatives Based on Mlr-Qspr Approach, *Acta Period. Technol.*, 55 203–213 (2024).

[25] Monter-Pozos A.; González-Estrada E., On testing the skew normal distribution by using Shapiro–Wilk test, *J. Comput. Appl. Math.*, 440 1–26 (2024).

[26] Hammoudan I., Chtita S., Bakhouch M., Temsamani D R., QSAR study of a series of peptidomimetic derivatives towards MERS-CoV inhibitors, *Moroccan J. Chem.*, 10(3): 405–416 (2022).

[27] Király P., Kiss R., Kovács D., Ballaj A., Tóth G., The Relevance of Goodness-of-fit, Robustness and Prediction Validation Categories of OECD-QSAR Principles with Respect to Sample Size and Model Type, *Mol. Inform.*, 41(11): 1–14, (2022).