*Chemistry Research Journal*, 2020, 5(6):79-100

*Available online www.chemrj.org*

Research Article

ISSN: 2455-8990
CODEN(USA): CRJHA5

# QSAR and Molecular Docking Analysis of Substituted Tetraketone and Benzyl-benzoate Analogs as Anti-tyrosine: A Novel Approach to anti-tyrosine kinase Drug Design and Discovery

**Emmanuel Israel Edache[1]*, Hassan Samuel[2], Yusuf Ishola Sulyman[2], Okeh Arinze[3], Oba Ismaila Ayine[4]**

[1]Department of Pure and Applied Chemistry, University of Maiduguri, Brono State, Nigeria
[2]Department of Science Laboratory Technology, Nigerian Institute of Leather and Science Technology, PMB 1034, Samaru, Zaria
[3]Owan EasT Extension Centre, Nigerian Institute of Leather and Science Technology, Edo State, Nigeria
[4]Department of Physical and Chemical Oceanography, Nigerian Institute for Oceanography and Marine Research, PMB 12729, Victoria Island, Lagos, Nigeria
*Corresponding author: edacheson2004@gmail.com, +2348066776802

**Abstract** In this paper, an attempt was made to develop a quantitative structure-activity relationship (QSAR) and molecular docking studies on a series of tetraketone and benzyl-benzoate Derivatives acting as protein tyrosine kinases (EGFR) inhibitors. QSAR was performed using the genetic algorithm-multiple linear regression (GA-MLR) method that comes out with a very promising result. According to Model-1 by GA-MLR anti-tyrosine activity of tetraketone and benzyl-benzoate derivatives were influenced by an individual (ATS0s, AATS6p, and VR1_Dze) and alignment independent descriptor (ATSC1i and SpMAD_Dzv) help in understanding the effect of ionization potential and electronegativities respectively at different position of tetraketone and benzyl-benzoate. The contribution plot of steric, geometric, and electrostatic field interactions generated by QSAR shows interesting results in terms of internal and external predictability. Finally, molecular docking analysis was carried out to better understand the interactions between EGFR targets and inhibitors in this series. Hydrophobic and hydrogen bond interactions lead to the identification of active binding sites of EGFR protein in the docked complex. The model proposed in this work can be employed to design new derivatives of tetraketone and benzyl-benzoate with specific tyrosine kinase (EGFR) inhibitory activity.

**Keywords** Tyrosinase, tetraketone and benzyl benzoate, Semi-empirical (PM6), GFA-MLR, QSAR, Applicability Domain, Molecular Docking, Drug Design, and Discovery

## Introduction

Tyrosinase is an enzyme that catalyzes the oxidation of phenols. It is also known as monophenol monooxygenase. It is a copper-containing enzyme present in animal tissues, higher plants, and fungi that catalyzes the production of melanin [1,2]. Production of melanin causes many kinds of skin diseases, such as hyperpigmented spots on the face and freckles [3]. Tyrosinase catalyzes both the hydroxylation of monophenols to o-di-phenols (monophenelase or cresolase activity) and the oxidation of o-di-phenols to o-quinones both using molecular oxygen followed by a series of non-enzymatic steps resulting in the formation of melanin which plays a crucial protective role against skin photocarcinogenesis [3-5]. Tyrosinase may involve in neuromelanin formation in the human brain and contribute to neurodegeneration associated with Parkinson's disease [5,6]. In fungi, the role of melanin is correlated with the differentiation of reproductive organ and spore formation, the virulence of pathogenic fungi, and tissue protection after injury. Besides, it causes undesired enzymatic browning such as injured cut fruits and vegetables which leads to a significant decrease in nutritional values [6]. As tyrosinase inhibitors have increasing importance due to enormous application prospects in recent periods, the various tyrosinase inhibitors are extracted from natural sources and synthesized. Among which some apply to pharmaceutical and cosmetic fields [7]. Tyrosinase inhibitors are useful for the treatment of some dermatological disorders associated with melanin hyperpigmentation, wound healing, parasite encapsulation, and also important in cosmetics for whitening and depigmentation after sunburn [3].

Lead optimization is a vital component of the drug discovery process in which a chemical showing promise is modified to greatly improve its usefulness as a drug. Computational methods like quantitative structure-activity relationships (QSAR) can facilitate this process by elucidating the chemical characteristics that are favorable and unfavorable through statistical analysis of a series of chemicals [8,9]. QSAR methods derive correlations between the properties/descriptors of molecules and their biological activities (e.g., inhibition constants or binding affinities). Since the advent of Free Wilson and Hansch's analysis, numerous methods have been published in the literature for structure-activity relationship modeling [10]. It is a meaningful correlation (model) between a set of independent variables (chemical descriptors) calculated from chemical graphs, and a dependent variable such as binding affinity, log P, or the pKa value whose value one wishes to predict for the compound of interest [11]. Docking Studies, as the structures of more potential drug targets, are elucidated the opportunity for computers to perform initial binding studies is increasing. By computationally docking a ligand to a protein, one limits concerns about assay complications such as compound solubility and the need to maintain extensive physical compound libraries. The objective of computational docking is to determine how molecules of known structure will interact. The molecule may bind to the receptor and modify their function [12].

This paper aims to find a correlation between molecular and electronic structures of 37 investigated tyrosinase inhibitors (Table 1) which were found to have tyrosinase activity through inhibiting tyrosinase reductase as their inhibition efficiency IC50 was reported [3,5]. Molecular orbital calculations were performed looking for good theoretical parameters to characterize the inhibition property of inhibitors which will be helpful to gain insight into the mechanism of inhibition.

## Materials

The materials used in this study include; DELL INSPIRON computer system (Intel Pentium), T4500 2.30GHz 2.30GHz processor Dual-core, 3GB ram size on Microsoft windows 10 operating system, Spartan' 14 version 1.1.2, ChemDraw ultra version 12.0.1, PaDEL descriptor tool kit version 2.20 and Microsoft Office Excel 2013 statistical software, Material Studio (modeling and simulation software) version 7.0, DTC_Euclidean program version 1.0, PyRx-Python prescription (version 0.8) (http://pyrx.sourceforge.net/downloads). 4R3P, retrieved from RCSB and prepared by Discovery Studio visualizer version 16.1.01 (http://www.accelyrs.com).

## Methods

The data set tetraketone and benzyl benzoate derivatives used in this study were taken from the work of [3,5] and are shown in Table 1. This set contains the values of the anti-tyrosinase inhibition potency compounds. The data set was divided into two groups, a training set consisted of 25 compounds and a test set with 12 compounds. The training and test sets were used for the construction of the models and to evaluate the predictive power of the generated models, respectively. The inhibitory activities in the logarithmic scale (pIC50 = log 1/IC50) fall in the range of -0.314 to 2.233, with a mean value of 0.0428. The various steps are presented in a flowchart in Fig. 1.

### Molecular Modeling and Generation of Molecular Descriptors

The dual-core personal computer equipped with the operating system Windows ten (10) was used for making calculations of this work. The structure of all the compounds was drawn using the ChemDraw Ultra module of the program and transferred to Spartan'14 (2013) version 1.1.2 (14) module to create the three-dimensional (3D) structure. These structures were then subjected to energy minimization using molecular mechanics (MMFF). Energy minimized molecules were subjected to optimization via the parameterization method (PM6) [13,14]. These methods have become popular in recent years because they can reach similar precision to other methods in less time and less cost from the computational point of view. The geometry optimization of the lowest energy structure was carried out without any symmetry constraints were also transferred to PaDEL-Descriptor [15] version 2.20 and were subjected to re-optimization (with the MMFF94 force field). The most stable structure for each compound was generated and used for calculating various physicochemical parameters used for the statistical analysis. The resulted geometries were used for the docking study.

### Calculation of fragment-based descriptors

For the generated descriptors, a pool of about 856 2D-3D descriptors was calculated using the PaDEL-Descriptor v2.20 software package. These descriptors include Acidic group count, ALOGP, Apol, Aromatic atoms count, BCUT, Chi cluster, constitutional, Eccentric connectivity index, electrotopological state, XLogP, Zagreb index, Moment of inertia, Zagreb index, Topological charge, Charged partial surface area, Wiener numbers, Petitjean shape index, RDF, WHIM, etc. All descriptors with constant values among the dataset were deleted, resulting in 316 different descriptors (independent variables) that were used in the QSAR analysis.

### Selection of the training and test sets

To compare the biological activities of the set of compounds that have a wide range of chemical structures (i.e., descriptors), the dataset was divided into representative training and test sets using a dissimilarity-based compound selection method called Kennard-Stone algorithms. The program is intended to split a source dataset into training and test sets for further modeling. There are many cases when splitting to training and test set is complicated because of poor endpoint variables range, etc. In this program, the authors implemented the Kennard-Stone algorithm which takes into account all available information (descriptors) to make a splitting, to get an evenly distributed set of data in both sets. The program is very quick, easy to use, with a well-documented manual that includes background information and steps to run the software. In my opinion, the program is very useful and can be applied for many kinds of datasets, which need to be split to develop and validate a predictive model.

### Optimized variable selection

Owed to the fact that it is tedious and unreasonable to investigate all possible combinations of the descriptor pool, genetic function approximation and multiple linear regression, which simplify the process and reduce the time required to execute algorithms, were implemented [16].

### Docking Studies Methods

3D structure of the enzyme tyrosinase with PDB code: 4R3P by [17]. The protein structure was obtained from the database online Protein Data Bank (http://www.rcsb.org/pdb/home).

### Preparation of protein structure

The 3D coordinates of the crystal structure of EGFR (PDB ID: 4R3P) were downloaded from the Protein Data Bank (http://www.rcsb.org/pdb/home). EGFR (chains A) were selected for the docking simulations. Before docking, all water molecules are removed from protein file 4R3P. After removing the water molecules H-atom was added to protein for correct ionization and tautomeric states of amino acid residues such as GLU, MET, ASP, THR, LEU, GLY, etc.

### Preparation of ligand structures

The ligands used for the docking study were selected from the literature [3]. The bioactive compounds that are mainly present in the plants were considered for the study. Geometry optimizations of the ligands were performed using the semi-empirical (PM6) calculation method using Spartan '14 software (www.wavefun.com). The compounds included in the study are 2,2'-((3-aminophenyl)methylene)bis(cyclohexane-1,3-dione, 2,2'-(phenylmethylene)bis(5,5-dimethylcyclohexane-1,3-dione, 2,2'-((3-aminophenyl)methylene)bis(5,5-dimethylcyclohexane-1,3-dione and 2,2'-((3-aminophenyl)methylene)bis(5,5-dimethylcyclohexane-1,3-dione. The bioactive compounds considered for the study are listed in Table 1. The ligand structures were generated using the tool ChemDraw ultra v12.0.2 (www.cambridgesoft.com) Three-dimensional optimizations of the ligand structures were done and saved as 'PDB file'.

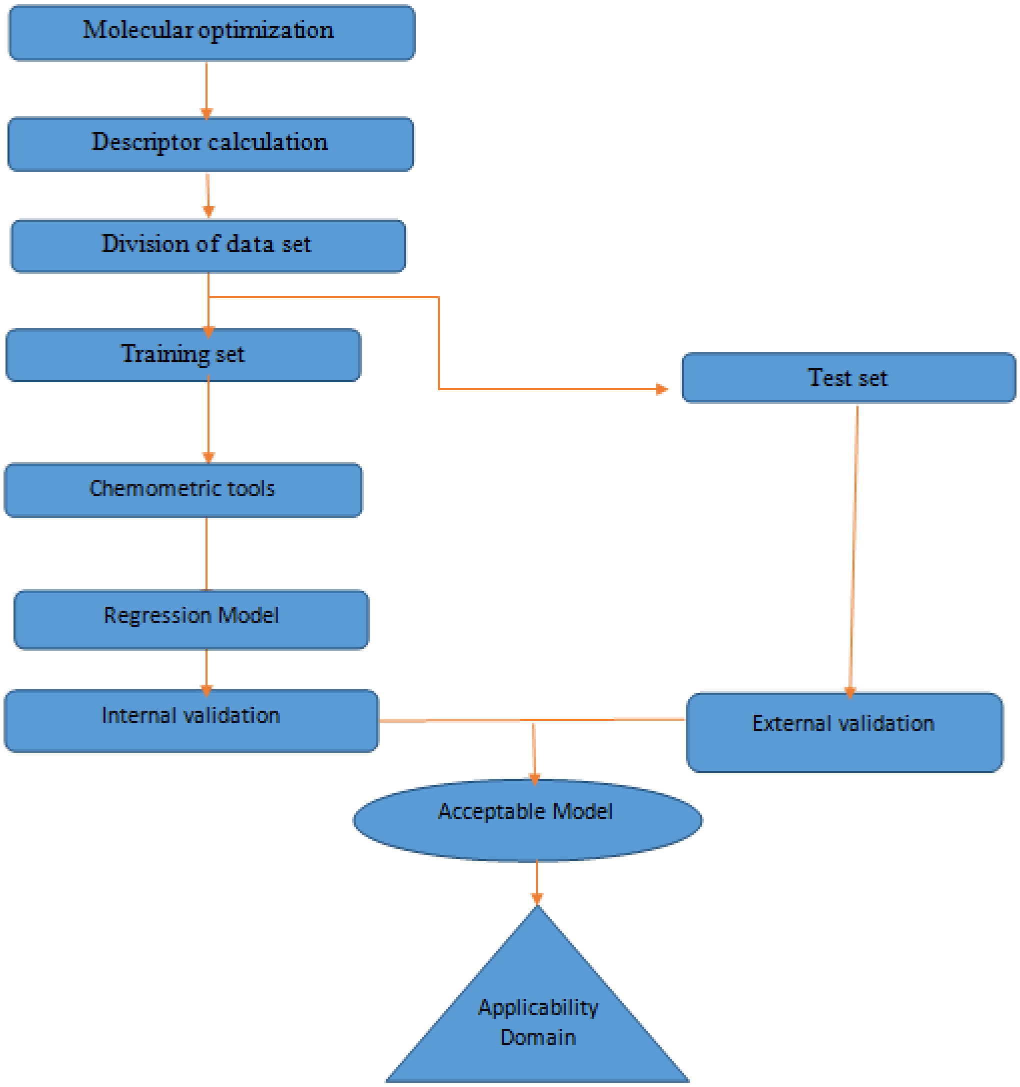

*Figure 1: Quantitative Structure-Activity Relationship Flow Chart*

### Protein-ligand interaction using PyRx (autodock vina)

The docking studies were conceded by PyRx (Autodock vina) tools (http://pyrx.sourceforge.net/downloads) version v0.8 programs. The searching grid extended above the preferred target proteins; polar hydrogen was added to the ligand moieties. Kollman charges were assigned and atomic solvation parameters were added. Polar hydrogen charges of the Gasteiger-type were assigned and the non-polar hydrogen was merged with the carbons and the internal degrees of freedom and torsions were set. Tetraketone compounds were docked to target protein complex (4R3P) with the molecule considered as a rigid body and the ligand being flexible. Evaluation of the results was done by sorting the different complexes concerning the predicted binding energy. A cluster analysis based on root mean square deviation values, regarding the starting geometry, was subsequently performed and the lowest energy conformation of the more populated cluster was considered as the most trustable solution.

**Table 1:** Structures of the dataset used for GA-MLR QSAR analysis with the corresponding observed and predicted class of tyrosinase inhibitors.

| **Compound ID** | **Structures of dataset** | **Observed $pIC_{50}$** | **Predicted $pIC_{50}$** | **Residual** |
|---|---|---|---|---|
| ID01 | | -0.816 | -1.332 | 0.515 |
| ID02 | | -1.425 | -1.008 | -0.417 |
| ID03 | | -1.090 | -1.200 | 0.110 |
| ID04 | | -1.230 | -1.308 | 0.078 |
| ID05 | | -1.071 | -0.776 | -0.295 |
| ID06 | | -0.684 | -0.950 | 0.266 |

| | | | |
|---|---|---|---|
| ID07 | -1.295 | -1.540 | 0.245 |
| ID09 | -0.681 | -0.839 | 0.158 |
| ID10 | -0.831 | -1.128 | 0.297 |
| ID11 | -0.320 | -0.013 | -0.307 |
| ID15 | -0.417 | -0.352 | -0.065 |
| ID16 | -0.616 | -0.713 | 0.097 |
| ID17 | -1.164 | -1.315 | 0.151 |
| ID18 | -0.957 | -0.577 | -0.380 |

| | | | |
|---|---|---|---|
| ID19 | -0.568 | -0.746 | 0.178 |
| ID20 | -1.108 | -0.918 | -0.190 |
| ID21 | -1.186 | -1.206 | 0.020 |
| ID22 | -0.819 | -0.558 | -0.261 |
| ID23 | -1.854 | -1.913 | 0.059 |
| ID24 | -0.603 | 0.503 | -1.106 |
| ID25 | -0.314 | 0.631 | 0.945 |
| ID26 | -1.127 | -0.939 | -0.188 |

| | | | |
|---|---|---|---|
| ID27 | -0.504 | -0.773 | 0.269 |
| ID28 | -1.103 | -0.894 | -0.209 |
| ID29 | 2.233 | 2.136 | 0.097 |
| ID30 | 1.909 | 1.807 | 0.102 |
| ID31 | 2.000 | 1.822 | 0.178 |
| ID32 | 1.580 | 2.163 | -0.583 |
| ID33 | 2.097 | 1.866 | 0.231 |
| ID34 | 2.213 | 2.298 | -0.085 |
| ID35 | 1.613 | 1.979 | -0.366 |

| | | | |
|---|---|---|---|
| ID36 | 2.205 | 1.948 | 0.257 |
| ID37 | 1.940 | 1.864 | 0.075 |
| ID38 | 1.699 | 1.967 | -0.268 |
| ID39 | 1.000 | 1.652 | -0.652 |
| ID42 | 1.699 | 1.598 | 0.101 |
| ID40 | 1.177 | 1.6502 | -0.4732 |

**Genetic Function Approximation**

Genetic Function Approximation (GFA) [18] is used to determine the best initialization of clusters as well as optimization of initial parameters. Genetic Function Approximation attempt to incorporate the ideas of natural evolution [19]. In general, they start with an initial population, and then a new population is created based on the notion of survival of the fittest. Typically, fitness is the measure of how good this population is and can be calculated depending on the nature of the application, where a distance measure is the most common [20]. Then a process called crossover is done over the new population where substrings from selected pairs are swapped [21].

Multiple Linear Regression is a method used for modeling the linear relationship between dependent variable Y ($pIC_{50}$) and independent variable X (descriptors). MLR is based on the least-squares method: the model is fitted such that the sum-of-squares of differences of the observed and predicted value is minimized. MLR estimates the values of regression coefficients ($R^2$) by applying the least-squares curve fitting method. The model creates a relationship in the form of a straight line (linear) that best approximates all the individual data points. In regression analysis, the conditional mean of the dependent variable ($pIC_{50}$) Y depends on (descriptors) X. MLR analysis extends this idea to include more than one independent variable.

Regression equation takes the form:

Y=B1*X1 + B2*X2 + B3*X3 + ………+ c

where Y is dependent variable, 'B's are regression coefficients for corresponding 'X's (independent variable), 'c' is a regression constant or intercept [22].

### QSAR Results

All molecules in each data set were successfully optimized by Spartan 14 V1.1.2 software. The following properties were obtained from the optimized structures: Molecular properties, QSAR descriptors, thermodynamic properties as well as acidity and basicity properties. The successful optimization of the molecules implies that all the molecules used have geometries close to their real or test tube geometries. Thus, properties computed from these optimized molecules are reliable. A statistically significant 2D-QSAR model was obtained using the properly selected training set of 25 ligands.

### Descriptor Calculation

The descriptors of each molecular structure were successfully computed with the aid of the PaDEL version 2.20 descriptor tool kit. Approximately 856 descriptors ranging from 1D, 2D, and 3D were obtained from these soft wares.

### GA-MLR Derived models for $pIC_{50}$ Anti-tyrosinase Compounds

Models 1 give the best Genetic Function Approximation-Multiple Linear Regression (GA-MLR) derived QSAR models for pIC50 of anti-tyrosinase molecules. Based on the model with the best statistical parameters identified using the parameters in Table 2 as standard, Model 1 was chosen as the best model for predicting the pIC50 of anti-tyrosinase molecules. The internal and external validation parameters of the models conform to the minimum standard for a robust QSAR model shown in Table 2, confirming the stability and robustness of the models.

### Genetic algorithm-multi-parameter linear regression

$$pMIC50 = 1.40103(+/-1.69175) - 0.0257(+/-0.0011)\,ATS0s - 3.73751(+/-0.98995)\,AATS6p + 0.19682(+/-0.01267)\,ATSC1i + 1.17379(+/-0.10375)\,SpMAD\_Dzv - 0(+/-0)\,VR1\_Dze. \quad \underline{\hspace{8cm}} \text{ Model 1}$$

### Comparison of observed and predicted $pIC_{50}$ of model 1

The comparison of the predicted pIC50 of the model with their experimental values are presented in Tables 1. The low residual values shown in the tables confirms the high predictive power of the models.

### The plot of Experimental Versus Predicted $pIC_{50}$ of model 1

The agreement between the experimental $pIC_{50}$ values of molecules used in the training and test set and the predicted values by the optimization models 1 presented in Fig. 2 and Fig. 3, respectively. The high Linearity of these plots indicates the high predictive power of the models.

### Residual plot of model 1

The measure of the dispersion of residual $pIC_{50}$ values from the predicted $pIC_{50}$ values is presented in Fig. 4. The propagation of the errors on both sides of zero is an indication of the robustness of the QSAR models.

**Table 2:** Statistical analysis of the QSAR models derived using GA-MLR approaches

| **Parameters** | **Values** | **Parameters** | **Values** | **Parameters** | **Values** |
|---|---|---|---|---|---|
| SEE | 0.2086 | $Q^2$ | 0.9705 | $r^2$ | 0.8756 |
| $R^2$ | 0.9823 | PRESS | 1.3774 | $r0^2$ | 0.8574 |
| $R^2$ adjusted | 0.9777 | SDEP | 0.2347 | reverse $r0^2$ | 0.8406 |
| F | 210.92 | $rm^2$ (Loo) | 0.9651 | $rm^2$ (test) | 0.7574 |
| Q | 4.7512 | $rm^{2'}$ (Loo) | 0.9671 | reverse $rm^2$ (test) | 0.7118 |
| FIT | 21.0889 | average $rm^2$ (LOO) | 0.9661 | average $rm^2$ (test) | 0.7346 |
| $\lvert r0^2-r'0^2 \rvert$ | 0.017,<0.3 | delta $rm^2$ (LOO) | 0.0020 | delta $rm^2$ (test) | 0.0456 |
| K | 0.8192, 0.85<k<1.15 | $rm^2$ (overall) | 0.8797 | RMSEP | 0.5389 |
| $[(r^2-r0^2)/r^2]$ | 0.02, < 0.1 | reverse $rm^2$ (overall) | 0.8756 | $Rpred^2$ | 0.8323 |
| k' | 1.0583, 0.85<k'<1.15 | average $rm^2$ (overall) | 0.8777 | $Q^2_{f1}$ | 0.8323 |
| $[(r^2-r'0^2)/r_2]$ | 0.04, < 0.1 | delta $rm^2$ (overall) | 0.0041 | $Q^2_{f2}$ | 0.8121 |

**Table 3:** *R*, *R*$^2$, *Q*$^2$, and *Rp*$^2$ values after several Y-Randomization tests

| **Model** | **R** | **$R^2$** | **$Q^2$** |
|---|---|---|---|
| Original | 0.9911 | 0.9823 | 0.9705 |
| Random 1 | 0.4627 | 0.2141 | -0.2846 |
| Random 2 | 0.4213 | 0.1775 | -1.0689 |
| Random 3 | 0.3681 | 0.1355 | -0.3185 |
| Random 4 | 0.5475 | 0.2997 | -0.2090 |
| Random 5 | 0.3707 | 0.1374 | -0.6789 |
| Random 6 | 0.3160 | 0.0998 | -0.4111 |
| Random 7 | 0.4059 | 0.1647 | -0.7249 |
| Random 8 | 0.4484 | 0.2011 | -0.4412 |
| Random 9 | 0.2769 | 0.0766 | -0.6252 |
| Random 10 | 0.3824 | 0.1462 | -0.5918 |
| Random Models Parameters | | | |
| Average r: | 0.3999 | | |
| Average $r^2$: | 0.1652 | | |
| Average $Q^2$: | -0.5354 | | |
| $Rp^2$: | 0.8988 | | |

**Table 4:** The linear model based on the five parameters selected by the GA-MLR method

| **Descriptors name** | **Symbol** | **VIF** | **MF** |
|---|---|---|---|
| Broto-Moreau autocorrelation - lag 0 / weighted by I-state | ATS0s | 1.5719 | 5.1762 |
| Average Broto-Moreau autocorrelation - lag 6 / weighted by polarizabilities | AATS6p | 1.3968 | 3.2661 |
| Centered Broto-Moreau autocorrelation - lag 1 / weighted by first ionization potential | ATSC1i | 1.3411 | -0.7138 |
| Spectral mean absolute deviation from Barysz matrix / weighted by van der Waals volumes | SpMAD_Dzv | 1.7905 | -6.8643 |
| Randic-like eigenvector-based index from Barysz matrix / weighted by Sanderson electronegativities | VR1_Dze | 1.9766 | 0.1357 |

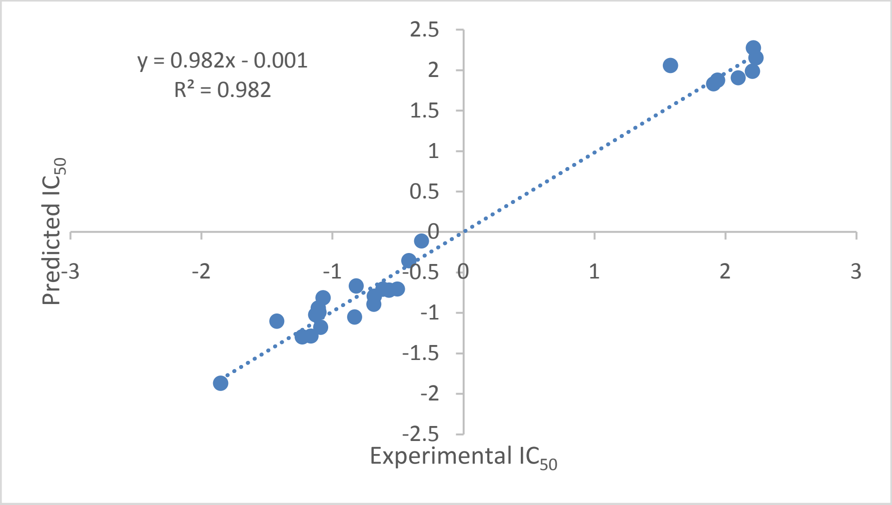

*Figure 2***:** *Scatter plot of the experimental activities versus predicted activities for the QSAR model, LOO cross-validated predictions on the full training set*

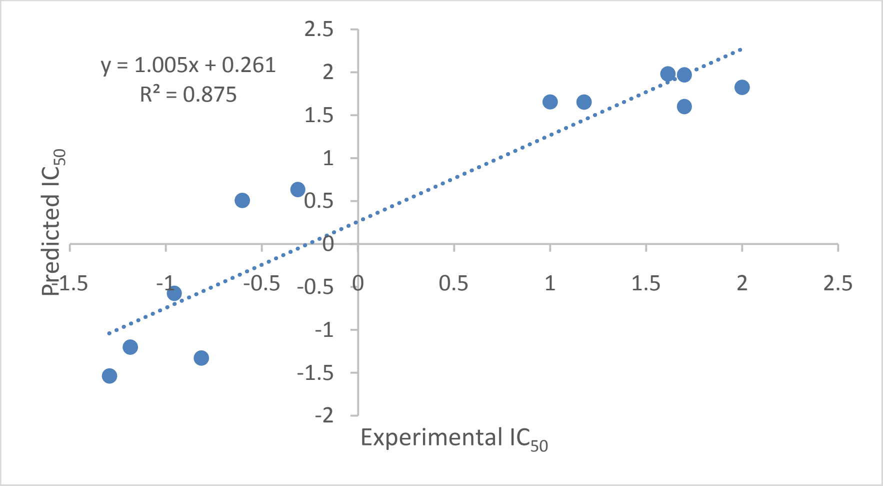

*Figure 3: Scatter plot of the experimental activities versus predicted activities for test-set predictions*

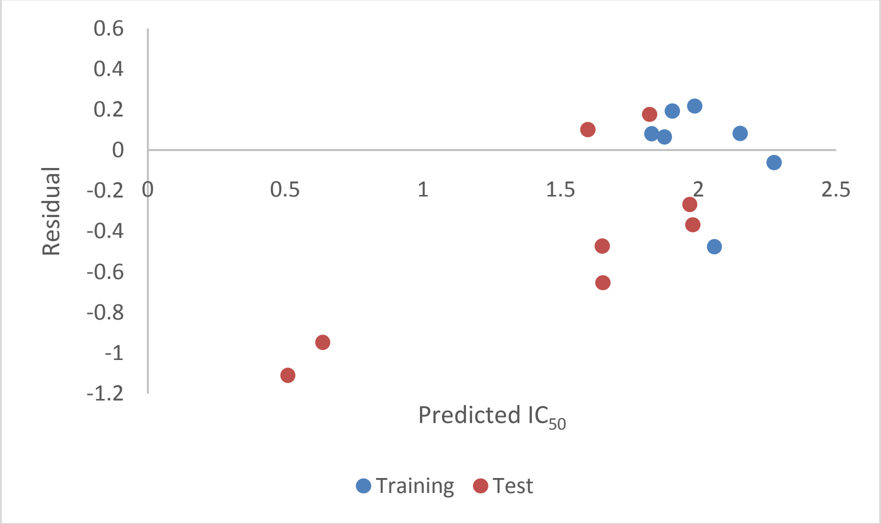

*Figure 4: The residual versus the experimental pIC$_{50}$ by measured GA-MLR*

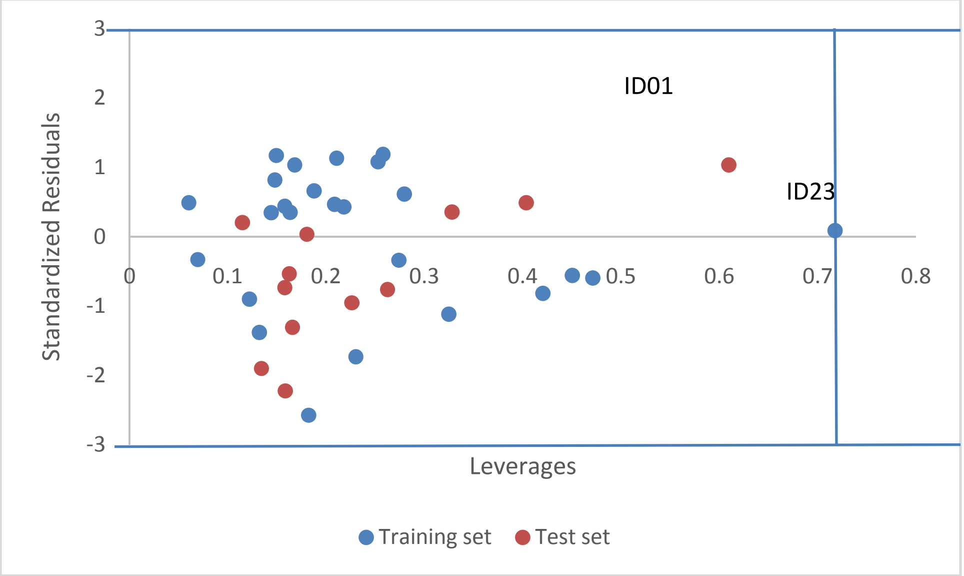

*Figure 5: William's plot of the generated GA-MLR model*

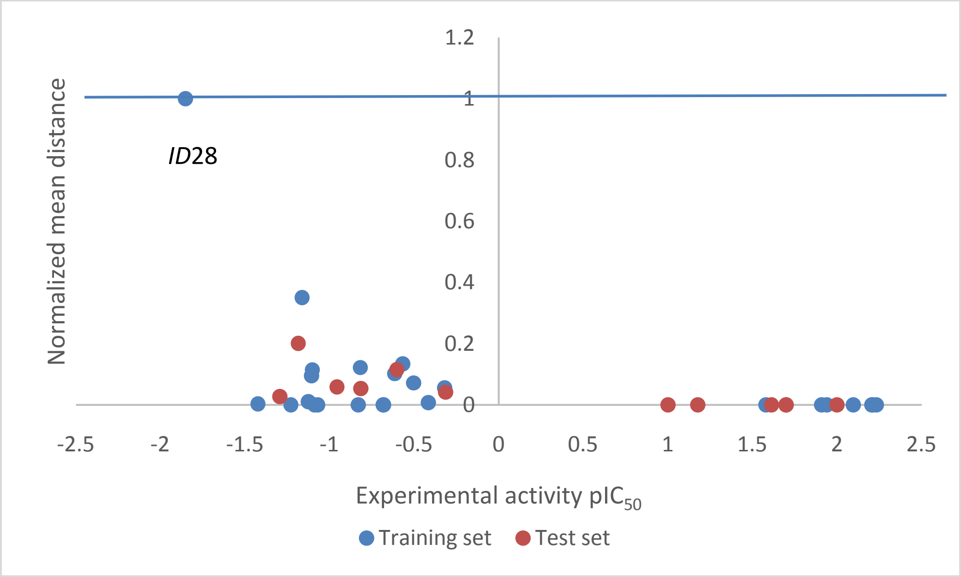

*Figure 6: Euclidean based applicability domain generated GA-MLR model*

## QSAR Discussion

After analyzing, we split the data set into the training set and query set, the next step was to select the main factors which were the most important for the anti-tyrosinase inhibition. As we do not know yet which descriptors or which particular combinations are related to the studied response and can be used in the predictive models, we applied genetic algorithms as the variable selection procedure to select only the best combinations (most relevant) for obtaining the models with the highest predictive power by using the training set. The five most significant descriptors according to the GA-MLR algorithm are Broto-Moreau autocorrelation - lag 0 / weighted by I-state (ATSOs), Average Broto-Moreau autocorrelation - lag 6 / weighted by polarizabilities, (AATS6p), Centered Broto-Moreau autocorrelation - lag 1 / weighted by first ionization potential (ATSC1i), Spectral mean absolute deviation from Barysz matrix / weighted by van der Waals volumes (SpMAD_Dzv), and Randic-like eigenvector-based index from Barysz matrix / weighted by Sanderson electronegativities (VR1_Dze). The predicted values for pIC50 for the compounds in the training and test sets using model 1 were plotted against the experimental pIC50 values in Fig. 2 and 3. A plot of the residual for the predicted values of pIC50 for both training and test sets against the experimental pIC50 values are shown in Fig. 4. Clearly, the model did not show any proportional and systematic error as suggested by Jalali-Heravi and Kyani [23], because the propagation of the residuals on both sides of zero is random. The real usefulness of QSAR models is not just their ability to reproduce known data verified by their fitting power (R2), but mainly it is their predictive application potential. The F -value is statistically significant at the 95% level since the calculated F value is higher as compared to the tabulated value. The positive value of quality factor (Q) for this QSAR's model suggests its high predictive power and lack of overfitting [24,25].

FIT Kubinyi function define the statistical quality of activity prediction, the number of variables that enter in a QSAR model is compared by using the FIT Kubinyi function as shown in the equation below, a criterion closely related to F value was proven to be useful.

$$FIT = R^2 (n - k - 1) / (n + k^2) (1 - R^2)$$

where n is the number of compounds in the training set and k is the number of variables in the QSAR equation. The main feature of the F value is its sensitivity to changes in k if k is small sensitivity is high and vice versa if k is large. The FIT criterion has a low sensitivity towards changes in k values, as long as they are small numbers and a substantial increase in sensitivity for large k values [26,27]. The best model will be the one that possesses a high value of this function. According to the statistical values of the models reported in Table 2, with five variables since this showed high FIT. The observed, calculated, and predicted values of the statistically significant five parameters QSAR model are presented in model 1.

The results of the statistical analysis are presented in Table 2. In the QSAR model, the initial GA analysis of the aligned training set was done using material studio version 7.0. This yielded a highly significant $Q^2$ value of 0.9705 (with SDEP = 0.2347 Table 2), which indicates that it is a model with high statistical significance; a $Q^2$ value of 0.6 is considered statistically significant in QSAR studies [28]. The conventional $R^2$ value of 0.9823 and low standard error of estimate (SEE) value of 0.2086 Table 1, indicate the accuracy of the predictions of the model. High values of Q2 from the leave-one-out (LOO) analysis (Table 2) can be regarded as a necessary, but not a sufficient, condition for a model to possess significant predictive power [29]. In addition to LOO, the internal predictive ability of the model was further assessed by a Y-randomization performed with 25 analogs for 10 times. The average of 10 readings was given as average $Q^2$ as shown in Table 3; the Y-randomization test ensures the robustness of a QSAR model [30] and to assess the multiple linear regression models obtained by descriptor selection [31]. In the Y-randomization test, the dependent variable or biological activity is randomly shuffled and a new QSAR model is developed keeping molecular descriptors intact. The new models are expected to have low $R^2$ and $Q^2$ values, which determine the statistical significance of the original model. Moreover, if the model development includes F-stepping, then it is necessary to shuffle both dependent and independent variables to indicate that the original model is not because of chance correlation. The low $R^2$ and $Q^2$LOO values of the random models shown in Table 3 and the value of $R^2P$ = 0.8988 ($RP^2$ ≥0.5) indicates that there is no chance of correlation or structural dependency in the proposed model. Consequently, model 1 can be considered as a perfect model with both high statistical significance and excellent predictive ability.

To satisfy with the robustness of the QSAR model developed using the training set, we have applied the QSAR model to an external data set of tetraketone and benzyl benzoate derivatives constituting the test set. As the experimental values of $IC_{50}$ for these inhibitors are already available, this set of molecules provides an excellent data set for testing the prediction power of the QSAR model for new ligands. Fig. 3 represents the predicted pIC50 values of the test set based on the model. The overall root means square error of prediction (RMSEP) between the experimental and predicted $pIC_{50}$ values was 0.5389 as showed in Table 2, which reveals good predictability. The estimated correlation coefficients between experimental and predicted $pIC_{50}$ values with intercept ($r0^2$) and without intercept ($r^2$) were 0.8406 and 0.8574, respectively. The value of $[(r^2 - r0^2)/r^2]$ = 0.002, which is less than 0.1 stipulated value [32] and thus validates the usefulness of the QSAR model for predicting the biological activity of the external data set. Also, the values of k and k′ were 0.8192 and 1.0583, which are within the specified ranges of

0.85 and 1.15 [30]. The values of $R^2$pred = 0.8323 and $rm^2$(test) = 0.7574 were found to be in the acceptable range (Table 2) [33], thereby indicating the good external predictability of the QSAR model.

Selecting the best model, values of $rm^2$(overall) for the model was determined. As shown in Table 2, this parameter penalized a model for large differences in experimental and predicted activity values. The parameter $rm^2$(overall) determines whether the predicted activities are close to the observed values or not since high values of $Q^2$ and R2pred does not necessarily mean that the predicted values are very close to the experimental ones. A model is considered satisfactory when $rm^2$(overall) is greater than 0.5 [34]. Besides rm^2(overall), we have calculated $rm^2$(test) and $rm^2$(LOO) values. These two parameters signify the differences between the experimental and predicted activities of the test and training set compounds. For an ideal predictive model, the difference between $R^2$pred and $rm^2$(test) and the difference between $Q^2$ and $rm^2$(LOO) in Table 2 should below. A large difference between the values will ultimately lead to poor values of the $rm^2$(overall) parameter. For this data set, the difference between $Q^2$ and $rm^2$(LOO) is quite less (0.0054), and that between $R^2$pred and $rm^2$(test) is also very less (0.0749). This indicates that the model obtained for this data set using those descriptors are quite robust and predictive. The $rm^2$(LOO) parameter for a given model indicates the extent of deviation of the LOO predicted activity values from the experimental ones for the training set compound while parameter $rm^2$(test) determines the extent of deviation of the predicted activity from the experimental activity values of test set compounds where the predicted activity is calculated based on the model developed using the corresponding training set. Model 1 shows acceptable values of $rm^2$(LOO) and $rm^2$(test) since they are greater than 0.5 [35].

The multi-collinearity between the above five descriptors were detected by calculating their variation inflation factors (VIF), which can be calculated as $1/1\text{-}R^2$ [36].

Where R is the correlation coefficient of the multiple regression between the variables in the model. If VIF equals 1, no inter-correlation exists for each variable; if VIF falls into the range of 1–5, the related model is acceptable; and if VIF is larger than 10, the related model is unstable and a recheck is necessary [28]. The corresponding VIF values of the seven descriptors are shown in Table 5. Based on this table, most of the variables had VIF values of less than 5, indicating that the obtained model has statistical significance. To examine the relative importance, as well as the contribution of each descriptor in the model, the value of the mean effect (MF) was calculated for each descriptor. This calculation was performed using the following equation.

$$MF_j = \frac{B_j \sum_{i=1}^{i=n} d_{ij}}{\sum_j^m B_j \sum_i^n d_{ij}}$$

## Applicability Domain of the Model

A quantitative structure-activity relationship (QSAR) model is exploited to monitor new compounds when its domain of application has been defined [30]. The prediction may be assumed reliable for only those compounds which fall into this domain [37]. Standardized residuals of the activity were computed and were plotted versus leverage values (h*). The value of leverage was calculated for every compound. Values are always between 0 and 1. A value of 0 is indicative of perfect prediction and usually is not accessible, and a value of 1 indicates very poor prediction. The lower the value, the higher confidence in the prediction. Warning leverage (h*) is another standard for an explanation of the results and is, generally, fixed at 3 (k +1)/ n, where k is the number of model parameters and n is the number of training and test sets [37]. Calculated leverage for training and test sets is useful for determining the compounds which affect the model and, in terms of the validation set, useful for assigning the applicability domain of the model. William's plot for the developed models in GA-MLR is shown in Figure 3. Response outliers are compounds that have standard residual points higher than ± 3.0 standard deviation units and a leverage value higher than the warning leverage, which is 0.72 for GA-MLR. As can be seen in Figure 3, all studied molecules in training and test sets lie with a high degree of confidence in the application domain of the developed models.

## Interpretation of Descriptors

The 2D-QSAR developed indicated that Broto-Moreau autocorrelation - lag 0 / weighted by I-state (ATSOs), Average Broto-Moreau autocorrelation - lag 6 / weighted by polarizabilities (AATS6p) and N Randic-like eigenvector-based index from Barysz matrix / weighted by Sanderson electronegativities (VR1_Dze) has positive values in the mean effect (Table 8) indicate that the indicated descriptor contributes positively to the value of $pIC_{50}$, whereas negative values of Centered Broto-Moreau autocorrelation - lag 1 / weighted by first ionization potential (ATSC1i), and Spectral mean absolute deviation from Barysz matrix / weighted by van der Waals volumes (SpMAD_Dzv) indicate that the greater the value of the descriptor the lower the value of $pIC_{50}$. In other words, increasing the ATSC1i and SpMAD_Dzv (Table 8) will decrease $pIC_{50}$, and increasing the ATS0s, AATS6p and VR1_Dze increase the extent of pIC50 of the tetraketone and benzyl benzoate derivatives. The mean effect reveals the significance of an individual descriptor presented in the regression model.

## Docking Results and Discussion

In this present study, to understand the formation of Hydrophobic and hydrogen bond interactions between the tetraketone compounds and active sites of the crystal structure of EGFR (PDB code: 4R3P) was used to explore their binding mode and docking study was performed by using PyRx (autodock vina) (http://pyrx.sourceforge.net/downloads). Four (4) naturally occurring tetraketone compounds were retrieved from Table 1. The 3D structure and energy minimization was done by Spartan '14 software (www.wavefun.com). To date, several crystal structures of EGFR in complex with different inhibitors have been reported in the literature [38], are used as inhibitors for a tyrosine kinase (EGFR). In the present study, we have used the X-ray crystallography structure of tyrosinase (PDB code: 4R3P) in ternary complex against tetraketone compounds that are used for the docking study (Table 5).The detection of ligand-binding sites is often the starting point for protein function identification and drug discovery [39].The goal of ligand-protein docking is to predict the predominant binding model(s) of a ligand with a protein of known three-dimensional structures [40]. PyRx (autodock vina) predicted the active site of the receptor EGFR (4R3P) with higher average precision. The active site of EGFR (4R3P) comprises of amino acid residues such as GLU762, MET766, ASP855, THR854, THR790, LYS745, ALA743, LEU844, LEU792, MET793, GLY796, VAL726, GLY719, and LEU718. As most of the amino acid residues in the active site are hydrophobic so they are the main contributors to the receptor and ligand-binding interaction (Table 5).

## Interaction between the Tetraketone Compounds and 4R3P

To study the binding mode of tetraketone compounds in the binding site of the crystal structure of EGFR (4R3P), docking simulations were performed employing PyRx (autodock vina) program and docking scores were calculated from the docked conformations of the crystal structure of EGFR (4R3P)-inhibitor complexes. Four tetraketone compounds such as 2,2'-((3-aminophenyl)methylene)bis(cyclohexane-1,3-dione, 2,2'-(phenylmethylene)bis(5,5-dimethylcyclohexane-1,3-dione, 2,2'-((3-aminophenyl)methylene)bis(5,5-dimethylcyclohexane-1,3-dione and 2,2'-((3-aminophenyl)methylene)bis(5,5-dimethylcyclohexane-1,3-dione were docked into the active site of crystal structure of EGFR (4R3P) by using the same protocol. Docking studies yield crucial information concerning the orientation of the inhibitors in the binding pocket of the target protein. Several potential inhibitors have been identified through the docking simulation [41]. The majority of the ligand had a greater binding affinity with the target receptor crystal structure of EGFR (4R3P). Inhibition was measured by the binding energy of chemical compounds posses (kcal/mol). It was depicted that aligned binding conformations of the tetraketone compounds in the binding pocket of the crystal structure of EGFR (4R3P), were derived from the docking simulations (PyRx software). The four tetraketone compounds such as 2,2'-((3-aminophenyl)methylene)bis(cyclohexane-1,3-dione, 2,2'-(phenylmethylene)bis(5,5-dimethylcyclohexane-1,3-dione, 2,2'-((3-aminophenyl)methylene)bis(5,5-dimethylcyclohexane-1,3-dione and 2,2'-((3-aminophenyl)methylene)bis(5,5-dimethylcyclohexane-1,3-dione were bind into the EGFR active sites. From the results it has been clearly observed 2,2'-((3-aminophenyl)methylene)bis(5,5-dimethylcyclohexane-1,3-dione (ID25) formed one hydrogen bond interaction with

EGFR. The corresponding docking energy value of ID25 (-6.7kcal/mol) with one H-bonding was shown in Fig. 9. The hydrogen bond was formed between GLU906 by a distance of 3.20Å. The docking energy of ID11 (-7.0Kal/mol) was shown in Fig. 7, ID15 (-7.5Kcal/mol) interaction with EGFR was presented in Fig. 8 and the ID27 (-7.6Kcal/mol) binding with EGFR was shown in Fig. 10. The molecular docking studies of tetraketone compounds into the EGFR binding site revealed a very clear preference for the binding pocket. Residues GLU762, MET766, ASP855, THR854, THR790, LYS745, ALA743, LEU844, LEU792, MET793, GLY796, VAL726, GLY719, and LEU718 are important for the catalytic mechanism of EGFR. Any ligand which can bind to GLU762, MET766, THR854, THR790, LYS745, ALA743, GLY796, VAL726, GLY719, and/or LEU718 and prevent the substrate from binding to the active site can behave as an inhibitor of tyrosinase. These two key residues are positioned at the end of the active site cleft. Usually binding of the substrate to EGFR occurs through a well-formed hydrophobic channel. So blocking the hydrophobic channel is an effective way to inhibit EGFR Pereanez et al., (2011) have reported that active site residues of ASP and a combination of ASP with GLY form the calcium-binding loop, which is responsible for coordinating the Ca2+ required during catalysis [42].

**Table 5:** Compounds of tetraketone and their molecular docking score

| **Comp.** | **Structure** | **IC$_{50}$ (μM)** | **Docking Score** |
|---|---|---|---|
| ID11 | | 2.09 | -7.0 |
| ID15 | | 2.61 | -7.5 |
| ID25 | | 2.06 | -6.7 |
| ID27 | | 3.19 | -7.6 |

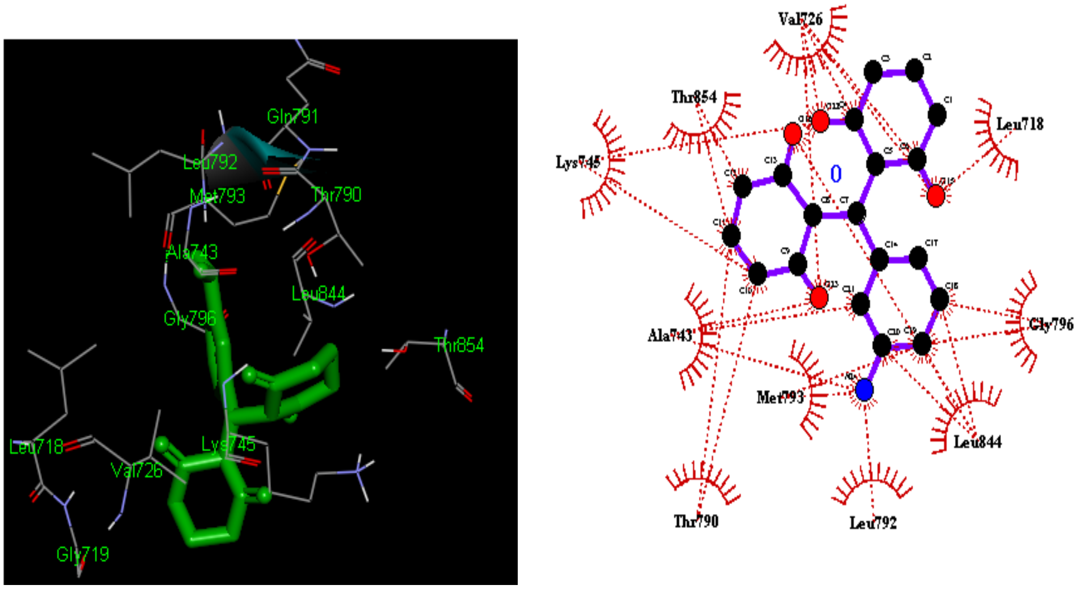

*Figure 7: Overlay of docked potent 2,2'-((3-aminophenyl)methylene)bis(cyclohexane-1,3-dione compound (ID11) at the active site of 4R3P produced using the PyRx, Discovery Studio, and LigPlot+ program*

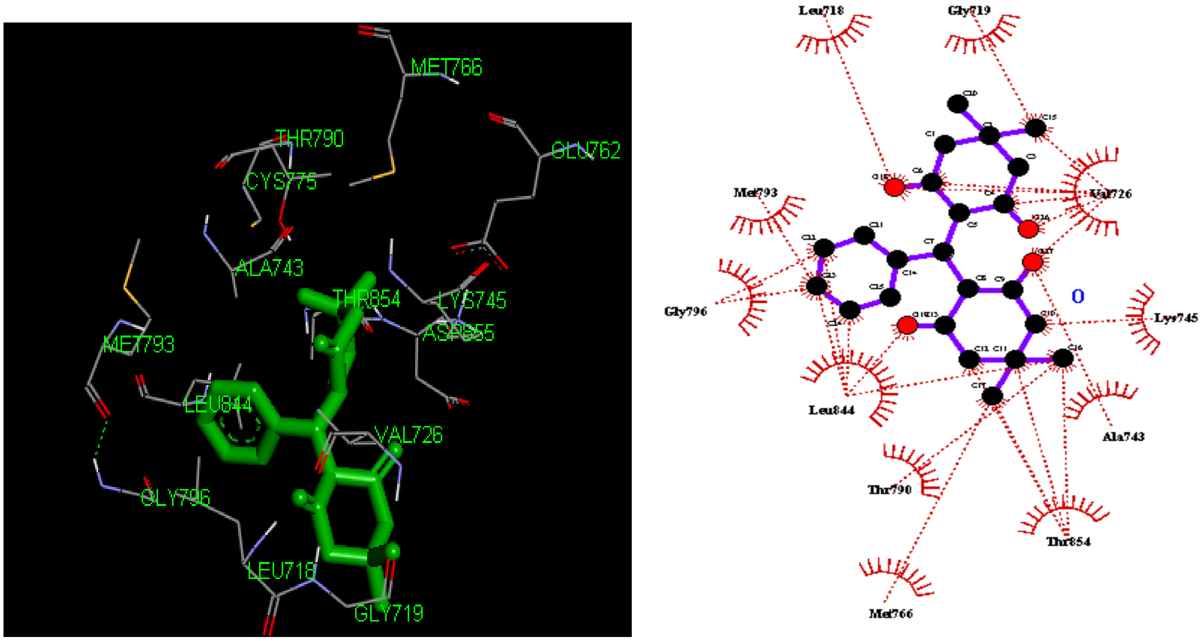

*Figure 8: Overlay of docked potent 2,2'-(phenylmethylene)bis(5,5-dimethylcyclohexane-1,3-dione compound (ID15) at the active site of 4R3P produced using the PyRx, Discovery Studio, and LigPlot+ program.*

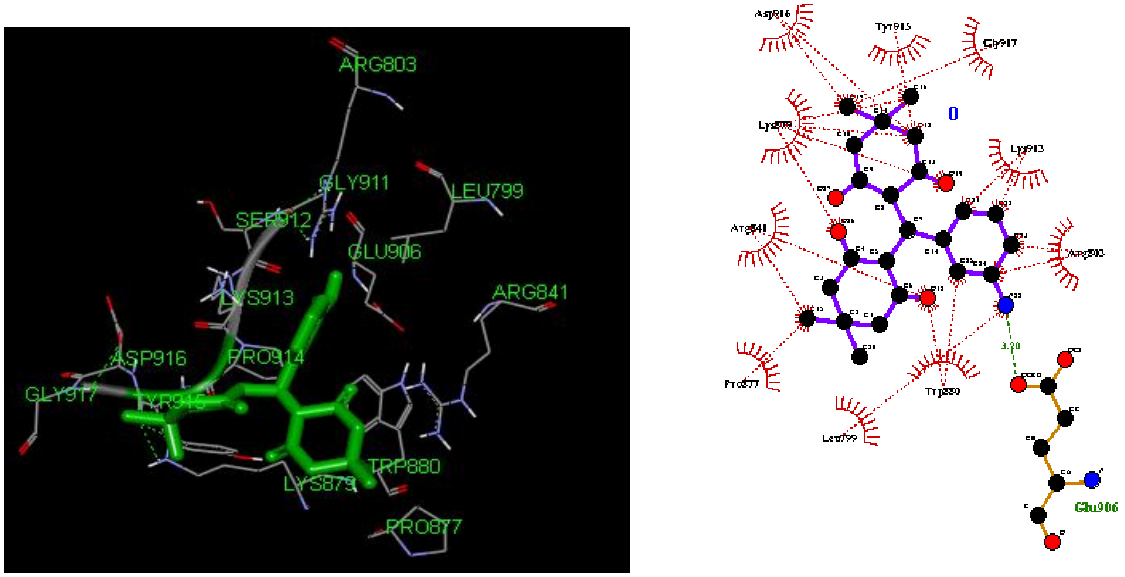

*Figure 9: Overlay of docked least potent 2,2'-((3-aminophenyl)methylene)bis(5,5-dimethylcyclohexane-1,3-dione compound (ID25) at the active site of 4R3P produced using the PyRx, Discovery Studio, and LigPlot+ program*

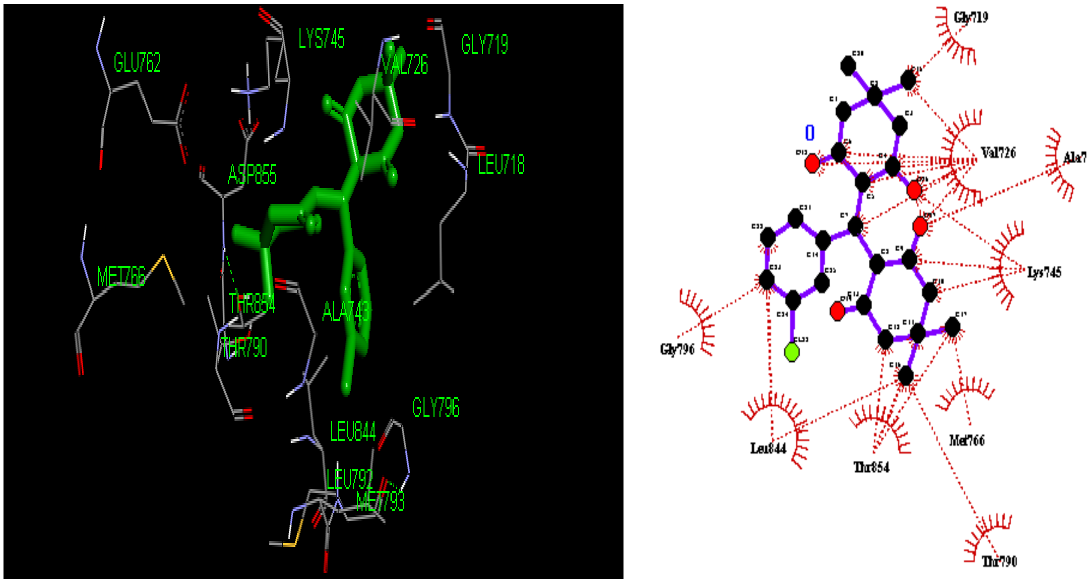

*Figure 10: Overlay of docked highest 2,2'-((3-aminophenyl)methylene)bis(5,5-dimethylcyclohexane-1,3-dione compound (ID27) at the active site of 4R3P produced using the PyRx, Discovery Studio, and LigPlot+ program*

## Conclusion

In the present QSAR investigation, the proposed QSAR models were statistically significant. However, Model-1 by genetic algorithm-multiple linear regression analysis could be considered as one in terms of excellent internal and external predictive abilities. According to Model-1 (GA-MLR), the anti-tyrosine activity of tetraketone and benzyl-benzoate derivatives was influenced by an individual (ATS0s, AATS6p, and VR1_Dze) and alignment independent descriptor (ATSC1i and SpMAD_Dzv) help in understanding the effect of ionization potential and electronegativities respectively at different position of tetraketone and benzyl-benzoate. The result obtained from the

QSAR study suggests that the electron-withdrawing group on tetraketone and benzyl-benzoate ring enhances the lipophilicity of compounds and favors the EGFR inhibition. It also suggests that a long chain group of tetraketone and benzyl-benzoate ring favors the activity. It also suggests that bulky electron-donating groups are favorable. This finding supports the experimental observations, where the presence of bulky electronegative groups signifies an increase in activities of compounds. From the molecular docking studies, it is evident that hydrophobic groups substituted of the tetraketone ring possessing strong hydrophobic interactions with nonpolar active residues are likely to enhance EGFR kinase inhibition. The tetraketone ring plays a crucial role in producing biological activity by interacting with GLU 316, an important active residue for the binding affinity of the inhibitor, which correlates with the results obtained from the crystallographic study of EGFR. These interactions underscore the importance of nitrogen atoms for binding and subsequent inhibitory capacity. The model proposed in this work can be employed to design new derivatives of tetraketone and benzyl-benzoate derivatives with specific tyrosine kinase (EGFR) inhibitory activity.

## Recommendation

1. These drugs like molecules may be synthesized and formulated appropriately.
2. Their pharmacological and toxicological activities could be performed on animal models before clinical trials.

## Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper. Also, they declare that this paper or part of it as not been published elsewhere.

## References

[1]. Seo SY, Sharma VK, and Sharma N (2003) Mushroom tyrosinase: Recent prospects, J. Agric. Food Chem.51:2837–53.

[2]. Abechi SE and Edache EI (2016) Application of Genetic Algorithm-Multiple Linear Regression (GA-MLR) For Prediction of Anti-Fungal Activity. Inter J of Pharma Sci and Res. 7: 204- 20.

[3]. Khan KM, Maharvi GM, Khan MTH, Shaikh AJ, Perveen S, Begum S, Choudhary MI (2006) Tetraketones: A new class of tyrosinase inhibitors. Bioorganic & Med Chem.14: 344–51.

[4]. Shiino M, Watanabe Y, Umezawa K (2001) Synthesis of N-substituted N-nitrosohydroxylamines as inhibitors of mushroom tyrosinase. Bioorg. Med. Chem. 9: 1233-40.

[5]. Kanchanapally RC, Macha R, Vunguturi S. and Tigulla P (2011) A Theoritical Study of Benzyl Benzoates with Agaricus Bisporus Tyrosinase Inhibitory Properties. Inter J. Life Sci & Pharma Res.1: 28 -40.

[6]. Lee TH, Seo JO, Baek S,and Kim SJ (2014) Inhibitory Effects of Resveratrol on Melanin Synthesis in Ultraviolet B-Induced Pigmentation in Guinea Pig Skin. Biomol Ther (Seoul). 22(1): 35–40.

[7]. Kim YJ, Uyama H (2005) Tyrosinase inhibitors from natural and synthetic sources: structure, inhibition mechanism and perspective for the future. Cell Mol Life Sci. 62(15): 1707-23.

[8]. Cramer, RD (2003) Topomer CoMFA: A Design Methodology for Rapid Lead Optimization. J. Med. Chem. 46: 374–388.

[9]. Giersiefen H, Hilgenfeld R, Hillisch A (2003) Modern Methods of Drug Discovery: An Introduction. In Modern Methods of Drug DiscoVery; Hilgenfeldl, A. H. R., Ed.; Birkha¨user Verlag: Basel, pp 1-18.

[10]. Ruchi RM, Ross AM, and Michael JS (2009) Comparison Data Sets for Benchmarking QSAR Methodologies in Lead Optimization. J. Chem. Inf. Model. 49: 1810–1820.

[11]. Sprous DG, Palmer RK, Swanson JT, Lawless M (2010) QSAR in the pharmaceutical research setting: QSAR models for broad, large problems. Curr Top Med Chem. 10(6):619–637.

[12]. Sharma RB, Chetia D (2013) Docking studies on quinine based antimalarial drugs for plasmepsin-II using bioinformatics tools. Intl J Pharmacy and Pharmaceutical Sci. 5(3): 681-85.

[13]. Dewar MJS, Zoebisch EG, Healy EF & Stewart JJP (1985) The development and use of quantum mechanical molecular models. 76. AMI: a new general purpose quantum mechanical molecular model. J. Amer. Chem. Soc., 107; 3902-9.

[14]. Stewart JJP (2004) Optimization of parameters for semiempirical methods IV: extension of MNDO, AM1, and PM3 to more main group elements. J Mol Model. 10: 155–164.

[15]. Yap CW (2011) PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J of Computational Chem. 32(7): 1466-1474.

[16]. Kirkpatrick S, Gelatt CD, Jr. Vecchi MP (1983) Optimization by Simulated Annealing. Science, New Series 220; 671-80.

[17]. Park ., KimN, Yi Z, Cho A, Kim K, Ficarro SB, Park A, Park WY, Murray B, Meyerson M, Beroukim R, Marto JA, Cho J, Eck MJ (2015) Crystal structures of EGFR in complex with Mig6.

[18]. Rogers D, Hopfinger AJ (1994) Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. J Chem Inf & Comput Sci., 34; 854-66.

[19]. Ching-Wen H, Lin K, Wu M, Hung K, Liu G, and Jen C (2014) "Intuitionistic fuzzy c-means clustering algorithm with neighborhood attraction in segmenting medical image." Soft Computing 1-12.

[20]. Arici T, Celebi S, Aydin AS, Temiz TT (2013) "Robust gesture recognition using feature pre-processing and weighted dynamic time warping." Multimedia Tools and Appli., 1-18.

[21]. Sumathi P and Kathiresan V (2016) "A Hybrid Model for Medical Data Using Machine Learning Approaches ", Int J of Modern Trends in Eng & Res. 2349-9745.

[22]. Thakur M, Thakur A, Ojha L (2014) Surface Area Grid in Modeling of Anti HIV Activity of TIBO Derivatives. Inte J Res and Development in Pharm & Life Sci., 3(3): 983-992.

[23]. Jalali-Heravi M. and Kyani A (2004) Use of Computer-Assisted Methods for the Modeling of the Retention Time of a Variety of Volatile Organic Compounds: A PCA-MLR-ANN Approach. *J. Chem. Inf. Comput. Sci. 44* (4); 1328–35.

[24]. Pogliani L (1996) “Modeling with special descriptors derived from a medium-sized set of connectivity indices,” J Phy Chem. 100: 18065–18077.

[25]. Pogliani L (1994) “Structure property relationships of amino acids and some dipeptides,” Amino Acids 6: 141–153.

[26]. Kubinyi H (1994) Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. Quant Struct Act Relat. 13:285-294.

[27]. Kubinyi H (1994) Variable Selection in QSAR Studies. II. A Highly Efficient Combination of Systematic Search and Evolution. Quant Struct Act Relat. 13:393-401.

[28]. Böhm M, Strzebeche J, Klebe G (1999) Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa. J. Med. Chem.42: 458–477.

[29]. Hattotuwagama CK, Doytchinova IA, Flower DR (2005) In silico prediction of peptide binding affinity to class i mouse major histocompatibility complexes: A comparative molecular similarity index analysis (CoMSIA) Study. J. Chem. Inf. Model.45: 1415–1423.

[30]. Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb Sci. 22: 69–77.

[31]. Livingstone DJ, Salt DW (2005) Judging the significance of multiple linear regression models. J Med Chem. 48 (3):661-663.

[32]. Edache EI, Hambali UH, Arthur DE, Oluwaseye A and Chinweuba OC (2016) In-silico Discovery and Simulated Selection of Multi-target Anti-HIV-1 Inhibitors. Int Res. J Pure & Appl Chem. 11(1): 1-15.

[33]. Roy PP, Roy K (2008) On some aspects of variable selection for partial least squares regression models. QSAR Comb Sci. 27: 302-313.

[34]. Golbraikh A, Tropsha A (2002) Beware of $q^2$! J Mol Graphic and Modelling, 20: 269-276.

[35]. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. Molecular Informatics, *29*(6-7): 476-488.

[36]. Edache EI, Uzairu A and Abechi SE (2015) Quantitative structure and activity relationship modeling study of anti-HIV-1RT inhibitors: Genetic function approximation and density function theory Methods. J Comput Methods in Mol Design. 5 (4): 61-76.

[37]. Mansourian M, Fassihi A, Saghaie L, Madadkar-Sobhani A, Mahnam K, Abbasi M (2015) QSAR and docking analysis of A2B adenosine receptor antagonists based on non-xanthine scaffold. Med Chem Res. 24:394–407.

[38]. Malleshappa NN, Harun MP (2013) A comparative QSAR analysis and molecular docking studies of quinazoline derivatives as tyrosine (EGFR) inhibitors: A rational approach to anticancer drug design. J. Saudi Chem Soc, 17:361-379.

[39]. Srivastava V, Gupta SP, Siddiqi MI, Mishra BN (2010). Molecular docking studies on quinazoline antifolate derivatives as human thymidylate synthase inhibitors, Bioinformation 4357-365.

[40]. Mittal RR, McKinnon RA, Sorich MJ (2009). Comparison data sets for benchmarking QSAR methodologies in lead optimization, J. Chem. Inf. Model. 49, 1810-1820.

[41]. Sindhu T, Rajamanikandan S, Durgapriya D, Anitha JR, Akila S, Gopalakrishnan VK (2011). Molecular docking and QSAR studies on Plant derived bioactive compounds as potent inhibitors of DEK oncoprotein, Asian J. Pharm. Clin. Res. 4: 67-71.

[42]. Pereanez JA, Nunez V, Patino AC, Londono M, Quintana JC (2011) Inhibitory effects of plant phenolic compounds on enzymatic and cytotoxic activities induced by a snake venom phospholipase $A_2$, Vitae, 18: 295-304.