

APPLICATION OF GENETIC ALGORITHM-MULTIPLE LINEAR REGRESSION (GA-MLR) FOR PREDICTION OF ANTI-FUNGAL ACTIVITY

Stephen Eyije Abechi and Emmanuel Israel Edache*

Department of Chemistry, Ahmadu Bello University, Zaria-Nigeria

*E-mail: edacheson2004@gmail.com. Tel.: +2348066776802

ABSTRACT

Aim: To develop good and rational Quantitative Structure Activity Relationship (QSAR) mathematical models that can predict to a significant level the anti-tyrosinase and anti-*Candida Albicans* Minimum inhibitory concentration (MIC) of ketone and tetra-ketone derivatives.

Place and Duration of Study: Department of Chemistry (Mathieson Laboratory (3)-Physical Chemistry unit), Ahmadu Bello University, Zaria, Nigeria, between December 2015 and March 2016.

Methodology: A set of 44 ketone and tetra-ketone derivatives with their anti-tyrosinase and anti-*Candida Albicans* activities in terms of minimum inhibitory concentration (MIC) against the gram-positive fungal and hyperpigmentation were selected for 1D-3D quantitative structure activity relationship (QSAR) analysis using the parameterization method 6 (PM6) basis set. The computed descriptors were correlated with their experimental MIC. Genetic Function Approximation (GFA) method and Multi-Linear Regression analysis (MLR) were used to derive the most statistically significant QSAR model.

Results: The result obtained indicates that the most statistically significant QSAR model was a five- parametric linear equation with the squared correlation coefficient (R^2) value of 0.9914, adjusted squared correlation coefficient (R^2_{adj}) value of 0.9896 and Leave one out (LOO) cross validation coefficient (Q^2) value of 0.9853. An external set was used for confirming the predictive power of the model, its $R^2_{pred} = 0.9618$ and $rm^2 = 0.8981$.

Conclusion: The QSAR results reveal that molecular mass, atomic mass, polarity, electronic and topological predominantly influence the anti-tyrosinase and anti-*Candida Albicans* activity of the complexes. The wealth of information in this study will provide an insight to designing novel bioactive ketones and tetra-ketones compound that will curb the emerging trend of multi-drug resistant strain of fungal and hyperpigmentation

Keywords: *Candida Albicans*; Tyrosinase; Hyperpigmentation; Melanogenesis; QSAR,

1.0 Introduction

Tyrosinase also known as polyphenol oxidase (PPO), is a copper-containing monooxygenase enzyme involved in melanogenesis [1]. The enzyme is widely distributed in fungi, higher plants and animals [2], and is involved in the first two steps of the melanin biosynthesis, in which L-tyrosine is hydroxylated to 3,4-dihydroxyphenylalanine (L-DOPA, monophenolase activity) and the latter is subsequently oxidated to dopaquinone (diphenolase activity) [3]. For the past few decades, tyrosinase inhibitors have been a great concern solely due to the key role of tyrosinase in both mammalian melanogenesis and fruit or fungi enzymatic browning. Melanogenesis has been defined as the entire process leading to the formation of dark macromolecular pigments, i.e., melanin [4]. Melanin is essential for protecting human skin against radiation, but the accumulation of abnormal melanin induces pigmentation disorders, such as melasma, freckles, ephelides, and senile lentiginos[5]. Melanogenesis is conducted in melanocytes, located in the basal layer of the epidermis and controlled by tyrosinase [3]. The study of tyrosinase inhibitory activity became of interest in recent years because of the significant industrial and economic impact of the inhibitors of this protein. Recently, different inhibitory compounds derived from natural sources or partly/fully synthetic have been tested [6]. On the other hand, knowledge of melanocyte biology and the processes underlying melanin synthesis has made remarkable progress over the last few years, opening new paths in the pharmacologic approach to the treatment of skin hyperpigmentation. In addition to inhibition of tyrosinase catalytic activity, other approaches to treat hyperpigmentation include inhibition of tyrosinase mRNA transcription, aberration of tyrosinase glycosylation and maturation, acceleration of tyrosinase degradation, interference with melanosome maturation and transfer, inhibition of inflammation-induced melanogenic response, and acceleration of skin turnover. Accordingly, a

huge number of depigmenting agents or whitening agents developed by those alternative approaches have been successfully identified and deeply reviewed in many articles [7-9].

Studies on tyrosinases, their substrates and inhibitors, are needed to better understand the details of its biological activity and to know how to control its [10]. The anti-tyrosinase activity can be achieved by several ways:

- (i) By reducing the intermediate o-dopaquinone to dopa with suitable reducing agents, such as ascorbic acid;
- (ii) By introducing o-dopaquinone scavengers, such as alkyl thiols which can react with dopaquinone to form colorless products;
- (iii) By employing alternative tyrosinase substrates, such as certain phenols whose enzymatic reaction products do not further undergo the next step;
- (iv) By denaturing the enzyme with non-specific enzyme inactivators, such as acids or bases; or by specific tyrosinase activators or inhibitors [11].

Candida Albicans is one of the many bugs which is to be found living in and on all of us. It is a fungus organisms, best known for causing thrush in the mouths of babies - sore white, moist plaques in the mouth and on the tongue, and thrush in the vaginas of women. It can also cause nappy rash, and soreness and itching around the anus and genitals in adults [12-14]. *Candida Albicans* is an opportunistic fungus (or form of yeast) that is the cause of many undesirable symptoms ranging from fatigue and weight gain, to joint pain and gassing [15]. The *Candida Albicans* yeast is a part of the gut flora, a group of microorganisms that live in the mouth and intestine. When the *Candida Albicans* population starts getting out of control it weakens the intestinal wall, penetrating through into the bloodstream and releasing its toxic byproducts throughout the body [16]. As they spread, these toxic byproducts cause damage to your body tissues and organs, wreaking havoc on your immune system [17]. The major waste product of yeast cell activity is acetaldehyde, a poisonous toxin that promotes free radical activity in the body [18]. Acetaldehyde is usually broken down into acetic acid within the liver [19]. However, if this process is not working efficiently then it can circulate through your body and cause unpleasant symptoms like headaches and nausea [18, 20].

The number of clinical infections worldwide by *Candida albicans* has risen considerably in recent years, and the incidence of resistance to traditional antifungal therapies is also increasing [21]. In addition, drug-related toxicity, significant drug interactions and insufficient bioavailability of the conventional antifungals, have encouraged the search for new alternatives among natural products [22].

The quantitative structure–activity relationship (QSAR) approach helps to correlate the specific biological activities or physical properties of a series of compounds with the measured or computed molecular properties of the compounds, in terms of descriptors [23]. QSAR methodologies save resources and hasten the process of the development of new molecules and drugs. There have been many QSAR researches related to design of anti-tyrosinase and anti-*Candida Albicans* drugs so far [24, 25] but a systematic QSAR study is yet to be carried out for series of ketones and tetraketones derivatives carrying a branched one to three amino functions. The aim of present work is to derive some statistically significant QSAR models for side chain modified ketones and tetraketones derivatives for their anti- tyrosinase and anti-*Candida Albicans* activities and to relate anti-tyrosinase and anti-*Candida Albicans* activity to its physicochemical properties.

2.0 Materials and Methods

Forty-four experimental data sets were used in the current study, taken from literature were used for the present study [26 27]. The structures of these forty-four compound libraries are shown in Figure 1-3, and detailed information for each data set is listed in Table 1. Each data set was split into training and test sets using Kennard-stone method with a ratio of 70:30 percent respectively for model validation purposes. The QSAR models were generated using a training set of 30 molecules. The structures observed and predicted biological activities of the training set molecules are presented in Table 1. Predictive power of the resulting models was evaluated by a test set of 14 molecules with uniformly distributed biological activity using Kennard-Stone method. The structures observed and predicted biological activities of the test set molecules are presented in Table 1. Selection of test set molecules was made by considering the fact that, test set molecules represent range of biological activity similar to training set. The mean of biological activity of training and test set was 1.1547 and 1.4192 respectively. Therefore test set is the true representative of the training set.

Figure 1-3: Data set (training and test set) from literature used in the Quantum Chemical QSAR analysis;

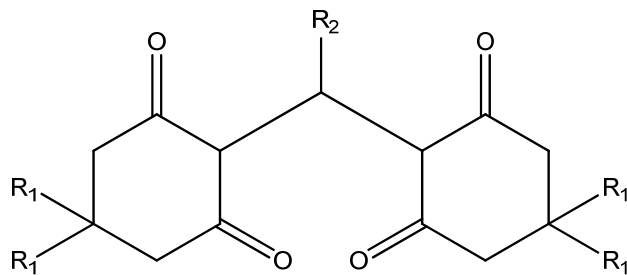


Figure 1: Compound 1-24

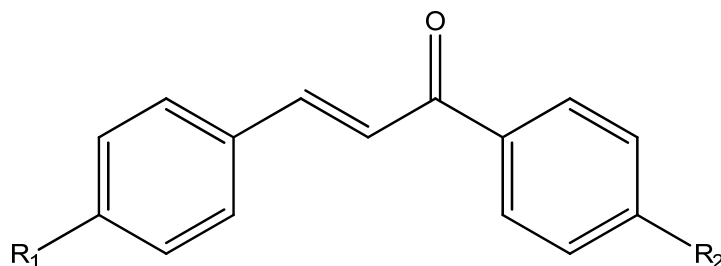


Figure 2: Compound 25-36

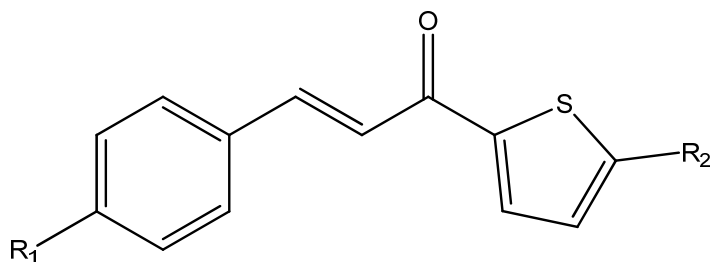
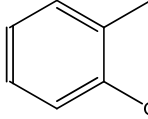
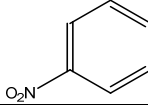
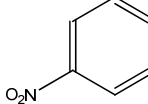
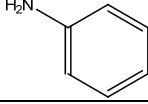
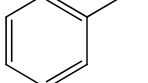
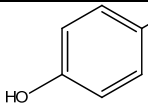
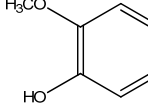
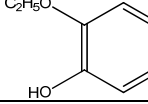
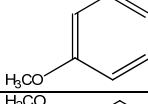
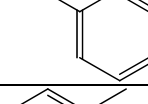
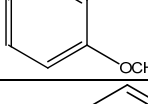
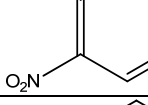
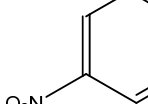
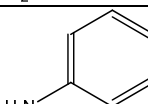
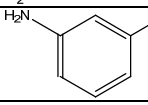
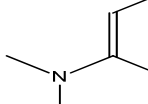
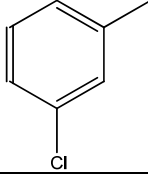
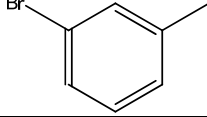


Figure 3: Compound 37-44

Table 1: Experimental and predicted activity of ketones/tetra-ketones derivatives

Comp'd No.	R1	R	pMIC ₅₀	Pred.pMIC ₅₀	Residual
1*		H	-0.81624	-0.722	-0.094
2		H	-1.425	-0.819	-0.607
3		H	-1.090	-1.223	0.133
4		H	-1.230	-1.245	0.015
5		H	-1.071	-1.119	0.048
6		H	-0.684	-0.909	0.225

7		H	-1.295	-1.058	-0.237
8		H	-0.681	-1.026	0.345
9		H	-0.831	-1.094	0.263
10		H	-0.320	-0.756	0.436
11		CH ₃	-0.417	-0.323	-0.094
12		CH ₃	-0.616	-0.575	-0.041
13		CH ₃	-1.164	-1.046	-0.118
14*		CH ₃	-0.957	-1.021	0.064
15		CH ₃	-0.568	-0.635	0.066
16*		CH ₃	-1.108	-0.780	-0.327
17*		CH ₃	-1.186	-0.905	-0.282
18		CH ₃	-0.819	-0.718	-0.101
19*		CH ₃	-1.854	-0.832	-1.022
20*		CH ₃	-0.603	-0.051	-0.552
21*		CH ₃	-0.314	-0.057	-0.257
22		CH ₃	-1.127	-0.578	-0.549

23		CH ₃	-0.504	-0.651	0.148
24		CH ₃	-1.103	-1.467	0.364
25*	4-SCH ₃	4-F	4.531	4.000	0.531
26*	4-SCH ₃	4-Cl	4.159	3.828	0.331
27*	4-SCH ₃	4-Br	3.522	3.334	0.188
28*	4-SCH ₃	2,4-Cl	3.208	2.420	0.788
29	4-SCH ₃	4-NO ₂	3.477	3.743	-0.266
30	4-SCH ₃	4-OCH ₃	4.152	4.018	0.134
31	4-SCH ₃	H	3.804	4.140	-0.336
32	4-SCH ₃	4-OH	3.829	3.799	0.030
33	4-SCH ₃	2-OH	4.130	3.770	0.360
34	4-SCH ₃	3-OH	3.829	3.966	-0.136
35	4-SCH ₃	4-phenyl	3.120	3.292	-0.172
36	2,3-OCH ₃	4-OCH ₃	4.474	3.878	0.596
37*	4-SCH ₃	H	4.716	4.052	0.664
38	4-SCH ₃	Br	3.832	3.390	0.442
39	3,4-OCH ₃	H	4.136	4.267	-0.134
40	3,4-OCH ₃	Br	3.548	3.832	-0.284
41*	4-phenyl	H	3.064	3.433	-0.369
42	4-phenyl	Br	3.169	2.965	0.204
43	4-OCH ₃	H	4.086	4.161	-0.075
44*	4-OCH ₃	Br	3.508	3.633	-0.125

*test set

2.1 Biological Activity Data

The QSAR models of anti-*Candida albicans* and anti-tyrosinase were developed in terms of half maximal inhibitory concentration IC₅₀ (μM) were taken from the literatures [26, 27]. The IC₅₀ summary data contains only molecules that have at least exhibited some activity. The biological activity data (IC₅₀) were converted into pIC₅₀ according to the formula; $pIC_{50} = (-\log (IC_{50} \times 10^{-6}))$ and was used as dependent variable, and correlated with the free energy change.

2.2 Computational Methodology

Chemdraw ultra [28] version 12.0.2 software was used to draw the structure of the compounds in the data set and each structure was saved as *cdx* file. The Spartan'14 [29] version 1.1.2 software was used for the energy minimization of the molecules. The PM6 semi-empirical model has been implemented. The molecules were first pre-optimized with the Semi-empirical (PM6) procedure included in Spartan'14 version 1.1.2 software [29]. In order to calculate the theoretical descriptors, PaDEL-Descriptor package version 2.20 was used [30]. For this purpose the output of the Spartan'14 software [29] for each compound was fed into the PaDEL-Descriptor program [30]. For the calculations, MMFF94s - a static variant of Merck Molecular Force Field 94 (MMFF94) [31] was used and the descriptors were calculated.

2.3 Descriptor Calculation

The PaDEL-descriptors version 2.20 [30] tools was employed for the calculation of different descriptors including Electrostatic descriptors; Topological Descriptors; Constitutional descriptors; Geometrical descriptors and Physicochemical descriptors as shown in Table 2. The calculated descriptors were gathered in a data matrix. The preprocessing of the independent variable (i.e., descriptors) was done by removing invariable (constant column) and cross-correlated (with R = 0.60) which resulted in 152 descriptors for GA-MLR to be used for QSAR analysis.

Table 2: Molecular descriptors used in this study

Descriptor classes	Descriptor names
Electrostatic descriptors	Max positive charge, Max negative charge, Max positive hydrogen charge, Total negative charge, Total positive charge, Total absolute atomic charge, Charge polarization, Local dipole index, Polarity parameter, Relative positive charge, Relative negative charge, PPSA1 (Partial Positive Surface Area 1st type), PPSA2, PPSA3, PNSA1 (Partial Negative Surface Area 1st type), PNSA3, DPSA1 (Difference in Charged Partial Surface Area), DPSA2, DPSA3, FPSA1 (Fractional charged partial positive surface area 1st type), FPSA2, FPSA3, FNSA1 (Fractional charged partial negative surface area 1st type), FNSA3, WPSA1 (Surface weighted charged partial positive surface area 1st type), WPSA2, WPSA3, WNSA1 (Surface weighted charged partial negative surface area 1st type), WNSA3, RPCS (Relative positive charge surface area), RNCS (Relative negative charge surface area), Hydrophobic SA – MPEOE, Positive charged polar SA – MPEOE, Negative charged polar SA – MPEOE, SADH1 (Surface area on donor hydrogens 1st type), SADH2 (Surface area on donor hydrogens 2nd type), SADH3 (Surface area on donor hydrogens 3rd type), CHDH1 (Charge on donatable hydrogens 1st type), CHDH2, CHDH3, SCDH1 (Surface weighted charged area on donor hydrogens 1st type), SCDH2, SCDH3, SAAA1 (Surface weighted charged area on acceptor atoms 1st type), SAAA2, SAAA3, CHAA1 (Charge on acceptors atoms 1st type), CHAA2, CHAA3, SCAA1 (Surface weighted charged area on acceptor atoms 1st type), SCAA2, SCAA3, HRNCS, HRNCG.
Topological Descriptors	Total structure connectivity index, Chi 0 (Simple zero-order chi index), Chi 1, Chi 2, Chi 3 path (Simple third order path chi index), Chi 3 cluster (Simple 3rd order cluster chi index), Chi 4 path, Chi 5 path, Chi 4 path/cluster (Simple 4th-order path/cluster chi index), VChi 0 (Valance zero order chi index), VChi 1, VChi 2, VChi 3 path (Valance 3rd order path chi index), VChi 4 path, VChi 3 cluster, VChi 4 path/cluster, VChi 5 path, Kier shape 1 (encodes the degree of cyclicity in the graph, decreases as graph cyclicity increases), Kier shape 2 (encodes the degree of central branching in the graph, decreases as the degree of central branching increases.), Kier shape 3 (encodes the degree of separated branching in the graph, increases as the degree of separation in branching increases.), Kier alpha 1 (1 st - Order Kappa Alpha Shape Index), Kier alpha 2, Kier alpha 3, Kier flexibility, Kier symmetry index, Kier steric descriptor, Delta Chi 0 (Delta zero-order chi index), Delta Chi 1, Delta Chi 2, Delta Chi 3 path, Delta Chi 3 cluster, Delta Chi 4 path, Chi 4 path/cluster, Delta Chi 5 path, Difference chi 0 (Difference simple zero-order chi index), Difference chi 1, Difference chi 2, Difference chi 3, Difference chi 4, Difference chi 5, IC (information content index), BIC (bond information content), CIC (complementary information content), SIC (structural information content), IAC total (total information index of atomic composition), I_adj_equ (Information index based on the vertex adjacency matrix equality), I_adj_mag (Information index based on the vertex adjacency matrix magnitude), I_adj_deg_equ (Information index based on the degree adjacency matrix equality), I_adj_deg_mag, I_dist_equ (Information index based on the distance matrix equality), I_dist_mag (Information index based on the distance matrix magnitude), I_edge_adj_equ (Information index based on the edge adjacency matrix equality), I_edge_adj_mag (Information index based on the edge adjacency matrix magnitude), I_edge_adj_deg_equ, I_edge_adj_deg_mag, I_edge_dist_equ, I_edge_dist_mag, Wiener index (Half-sum of the off-diagonal elements of the distance matrix of a graph), Hyper Wiener index, Harary index (Half-sum of the off-diagonal elements of the reciprocal molecular distance matrix), 1st Zagreb (1st Zagreb index), 2nd Zagreb, Quadratic index, Rouvray index, 2-MTI (Schultz Molecular Topological Index (MTI)), 2-MTI prime (Schultz MTI by valence vertex degrees), Gutman MTI, Graph diameter, Graph radius, Graph Petitjean, Eccentric connectivity index, Eccentric adjacency index, Platt number, Odd–even index, Vertex degree–distance index, Ring degree–distance index, Balaban index JX, Balaban index JY, Xu (Xu index), Superpendentic index, Unipolarity_distance_matrix, Centralization_distance_matrix,

Descriptor classes	Descriptor names
	Dispersion_distance_matrix, SC-0 (Subgraph Count Index of order 0), SC-1, SC-2, SC-3 path, SC-3 cluster, SC-4 path, Solvation chi 4 path/cluster, Solvation chi 5 path, VS-0 (Valence Shell Count of order 0), VS-1, VS-2, VS-3, VS-4, VS-5, Molecular walk count 2, Molecular walk count 3, Molecular walk count 4, Molecular walk count 5, Path/walk 2, Path/walk 3, Path/walk 4, Path/walk 5, Narumi ATI (Narumi simple topological index (log)), Narumi HTI (Narumi harmonic topological index), Narumi GTI (Narumi geometric topological index), Galvez topological charge indices, Difference connectivity indice, BCUT descriptors, Pogliani index, Ramification index, Degree complexity, Graph vertex complexity, Graph distance complexity, Graph distance index, Mean square distance index, Mean distance deviation, Edge Wiener index, Edge Hyper Wiener index, Edge MTI, Edge Gutman MTI, Edge connectivity index, E-state SsCH3, E-state SssCH2, E-state SdsCH, E-state SsssCH, E-state SdO, E-state S_hydrophobic, E-state S_hydrophobic_unsat, E-state S_polar, E-state S_hbond_donor, E-state SHdsCH, E-state SHCHnX, E-state SH_hydrophobic, E-state SdssC, E-state SsssN, E-state SsOHI, E-state SsF, E-state S_hydrophobic_sat, E-state S_none, E-state S_hbond_acceptor, E-state SsNH2, E-state SssNH, E-state SssO, E-state SsCl, E-state SsBr, E-state ShsOHI, E-state SHsNH2, E-state SHssNH, E-state SHCsats, E-state SHCsatu, , E-state SH_hbond_donor.
Constitutional descriptors	No. amino groups tertiary, no. of amide groups, no. of ester groups, no. of halogen atoms, molecular mass, no. of total atoms, no. of rotatable bonds, fraction of rotatable bonds, no. of rigid bonds, no. of rings, no. of single bonds, no. of double bonds, no. of H-bond acceptors, no. of H-bond donors, ratio of donors to acceptor.
Geometrical descriptors	2D-VDW surface, 2D-VDW volume, 2D-VSA hydrophobic, Fraction of 2D-VSA hydrophobic, 2D-VSA hydrophobic_sat, 2D-VSA hydrophobic_unsat, 2D-VSA other, 2D-VSA polar, Fraction of 2D-VSA polar, 2D-VSA Hbond 2D-VSA Hbond donor, 2D-VSA Hbond all, Fraction of 2D-VSA Hbond, Topological PSA.
Physicochemical descriptors	Polarizability_Miller, SKlogP value, Water solubilityl, Vapor pressure, Buffer solubility, SK_MP, AMR value (Calculated molecular refractivity index), Polarizability_MPEOE, SKlogS value, SKlogPvp, SKlogS_buffer, SK_BP, Solvation Free Energy, AlogP98 value, AlogP98 002C, AlogP98 006C, AlogP98 008C, AlogP98 016C, AlogP98 017C, AlogP98 019C, AlogP98 041C, AlogP98 047H, AlogP98 067N, AlogP98 073N, AlogP98 075N, AlogP98 094Br, AlogP98 084F, AlogP98 089Cl, AlogP98 094Br, AlogP98 001C, AlogP98 003C, AlogP98 005C, AlogP98 046H, AlogP98 050H, AlogP98 052H, AlogP98 056O, AlogP98 058O, AlogP98 059O, AlogP98 060O, AlogP98 068N, AlogP98 072N.

2.4 Selection of descriptors and development of the QSAR model

A set of 1867 molecular descriptors was calculated using the PaDEL-descriptor software package (version 2.20) [30]. A systematic search in the order of missing value test, zero test, correlation coefficient, multicollinearity, and genetic algorithm was performed to determine significant descriptors using the material studio (version 7.0) software package [32]. Any parameter that was not calculated (missing value) for any number of the compounds in the data set was rejected (deleted) in the first step. Some of the descriptors were rejected (deleted) because they contained a zero value for all the compounds (zero tests).

2.5 Data Pre-treatment Tool (V-WSP)

To remove the constant and highly inter-correlated descriptors based on user specified variance and correlation coefficient cut-off values using V-WSP (version 1.2) algorithm proposed by Ballabio *et al.*, [33]. It is an unsupervised variable reduction method, which is a modification of the recently proposed WSP algorithm for design of experiments (DOE). A cutoff value of 0.6 and the variables physically removed from the analysis that showed exact linear dependencies between subsets of the variables and multicollinearity (high multiple correlations between subsets of the variables). From the descriptors, the set of descriptors that would give the statistically best QSAR models was selected by using a genetic function approach implemented in the materials studio (version 7.0) software package [32]. The genetic algorithm (GA) starts with the creation of a population of randomly generated parameter sets. The usage probability of a given parameter from the active set is 0.5 in any of the initial population sets. The sets are then compared according to their objective functions. The parameters set used for the GA includes mutation 0.1, crossover 0.9, population 3000 and number of generations

2000. The form of the objective function favors sets that have *LOF*s low as possible while minimizing the number of parameters used as descriptors. The lower the score, the higher the probability that a given set will be used for the creation of the next generation of sets. Creation of a consecutive generation involves crossovers between set contents, as well as mutations. The algorithm runs until the desired number of generations is reached. Equations were developed between the observed activity and the descriptors. The best equation was taken based on statistical parameters such as squared regression coefficient (R^2) and leave-one-out cross-validated regression coefficient (Q^2).

2.6 Validation of the QSAR model

The predictive capability of the QSAR equation was determined using the leave-one-out cross-validation method. The cross-validation regression coefficient (Q^2) was calculated by the following equation:

$$Q^2 = 1 - \frac{\sum(Y_{exp} - Y_{pred})^2}{\sum(Y_{exp} - \bar{Y})^2}$$

where Y_{pred} , Y_{exp} , and \bar{Y} are the predicted, experimental, and mean values of experimental activity, respectively. Also, the accuracy of the prediction of the QSAR equation was validated by *LOF*, *F* value, R^2 , and R^2_{adj} . A small value of *LOF* and a large *F* indicates that the model fit is not a chance occurrence. It has been shown that a high value of statistical characteristics is not necessary for the proof of a highly predictive model [34, 35]. Hence, to evaluate the predictive ability of our QSAR model, we used the method described by Golbraikh and Tropsha, [34] and Roy and Roy [35]. The values of the correlation coefficient of predicted and actual activities and the correlation coefficient for regressions through the origin (predicted vs. actual activities and vice versa) were calculated using the regression of analysis Tool-pak option of Excel, and other parameters were calculated as reported by Golbraikh and Tropsha, [34] and Roy and Roy, [35]. The determination coefficient in prediction, Q^2_{test} , was calculated using the following equation:

$$Q^2_{test} = 1 - \frac{\sum(Y_{pred_{test}} - Y_{test})^2}{\sum(Y_{test} - \bar{Y}_{training})^2}$$

for the external test set compounds was done by determining the value of r_m^2 by the following equation:

$$r_m^2 = r_{test}^2 \left(1 - \left| \sqrt{r_{test}^2 - r_{test_0}^2} \right| \right)$$

where r_{test}^2 is the square correlation coefficient between experimental and predicted values and $r_{test_0}^2$ is the squared correlation coefficient between experimental and predicted values without intercept for the external test set compounds. The values of *K* and *K'*, slopes of the regression line of the predicted activity versus actual activity and vice versa, were calculated using the following equations [34]:

$$K = \frac{\sum Y_i \bar{Y}_i}{\sum \bar{Y}_i^2} \text{ and } K' = \frac{\sum Y_i \bar{Y}_i}{\sum Y_i^2}$$

where Y_i and \bar{Y}_i are the predicted and experimental activities, respectively.

Further statistical significance of the relationship between activity and the descriptors was checked by randomization test (Y-randomization) of the models. The Y column entries were scrambled and new QSAR models were developed using same set of variables as present in the unrandomized model. We have used a parameter, R_p^2 [36], which penalizes the model R^2 for the difference between squared mean correlation coefficient (R_r^2) of randomized models and squared correlation coefficient (R^2) of the nonrandomized model. The R_p^2 parameter was calculated by the following equation:

$$R_p^2 = R^2 \times \sqrt{R^2 - R_r^2}$$

This parameter, R_p^2 , ensures that the models thus developed are not obtained by chance. We have assumed that the value of R_p^2 should be greater than 0.5 for an acceptable model.

To check the intercorrelation of descriptors, variance inflation factor (*VIF*) analysis was performed. The *VIF* value is calculated by the following equation:

$$VIF = \frac{1}{1 - R^2}$$

where R^2 is the multiple correlation coefficient of one descriptor's effect regressed on the remaining molecular descriptors. If the *VIF* value is larger than 10, information of descriptors can be hidden by correlation of descriptors [37, 38].

3.0 Results and Discussion

The 44 active compounds with their activity (IC_{50} values in μM) against *Candida albicans* and tyrosinase inhibitors were randomly divided into a training set of 30 compounds and a test set of 14 compounds. With the wide range of difference between the IC_{50} values and the large diversity in the structures, the combined data set of 30 molecules and 14 molecules is ideal as a training and test set, as both sets do not suffer from bias due to the differences of the structures. Table 3 shows a univariate analysis for the inhibition data. Table 3 contains several statistical measures that describe the tyrosinase and *Candida albicans* inhibition data. The most important parameters in Table 3 are the skewness and kurtosis. Skewness is the third moment of the distribution, which indicates the symmetry of the distribution. As the skewness is positive, the distribution of data values within the column is skewed toward positive values. For a symmetrical distribution, the skewness is zero. Kurtosis is the fourth moment of the distribution, which indicates the profile of the column of data relative to a normal distribution. Univariate analysis calculates Fisher kurtosis, which subtracts 3.0 from the definition above. For a normally distributed data set, it gives a value of 0.0. If the kurtosis is positive, the distribution of data in the column is more sharply peaked than a normal distribution. If the kurtosis is negative, the distribution is flatter than a normal distribution [39].

Table 3: A univariate analysis for the inhibition data

Parameter	pIC ₅₀
Number of sample points	30
Range	5.89900000
Maximum	4.47400000
Minimum	-1.42500000
Mean	1.15470000
Median	-0.46050000
Variance	5.53231000
Standard deviation	2.39229000
Mean absolute deviation	2.30499000
Skewness	0.25734300
Kurtosis	-1.91654000

The various molecular descriptors (about 1900 in total) as described in Table 2 were calculated initially. By applying a missing value test, a zero test, a correlation test with a cutoff value of 0.6, and a multicollinearity test with a cutoff value of 0.9, we have discarded the most likely parameters, resulting in 152 parameters. Further additional parameters were discarded by applying the GA, and finally 5 parameters were selected for the development of the QSAR equation. As the squared correlation coefficient, R^2 , can be easily increased by the number of terms in the QSAR equation, we took the cross-validation correlation coefficient, Q^2 , as the limiting factor for a number of descriptors to be used in the final model. It was observed that the Q^2 value increased until the number of descriptors in the equation reached 5. With further addition of parameters to the equation with 5 descriptors, there was a decrease in the Q^2 value of the model. So, the number of descriptors was restricted to 5 in the final QSAR model. The best significant relationship for the activity against tyrosinase and *Candida albicans* inhibitors has been deduced to be:

$$pIC_{50} = 4.4138(\pm 0.4305) + 0.3229(\pm 0.0339) ALogP - 0.0007(\pm 0.00003) ATS4m - 0.0823(\pm 0.01749) AATSC5v + 1.6287(\pm 0.5647) SHsNH2 + 6.5004(\pm 1.0939) RotBtFrac$$

$$(1) N = 30, SEE = 0.2436, R^2 = 0.9914, R^2_{adjusted} = 0.9896, F = 554.7321 (DF = 5, 24), Q^2 = 0.9853, PRESS = 2.4455, SDEP : 0.2855,$$

where N is the number of compounds in the training set, R^2 is the squared correlation coefficient, SEE is the estimated standard deviation about the regression line, $R^2_{adjusted}$ is the square of the adjusted correlation coefficient for degrees of freedom, F test is the measure of variance that compares 2 models differing by 1 or more variables to see if the more complex model is more reliable than the less complex one (the model is supposed to be good if the F test is above a threshold value), and Q^2 is the square of the correlation coefficient of the cross-validation using the leave-one-out cross-validation technique. The QSAR model developed in this study was statistically ($R^2 = 0.9914$, $Q^2 = 0.9853$, F test = 554.7321) best fitted and consequently was used for prediction of activities against strains of *Candida albicans* and tyrosinase (pIC_{50}) of training and test sets of molecules, as reported in Tables 1. The relationships between predicted (both training and test) activities and the corresponding experimental activities are shown in Figures 4 and 5. The R^2 and Q^2 values of 0.9914 and 0.9853, respectively, of the model corroborate with the criteria for a QSAR model to be highly predictive [34]. The

standard error of estimate for the model was 0.2436, which is an indicator of the robustness of the fit and suggested that the predicted pIC_{50} based on equation (1) is reliable.

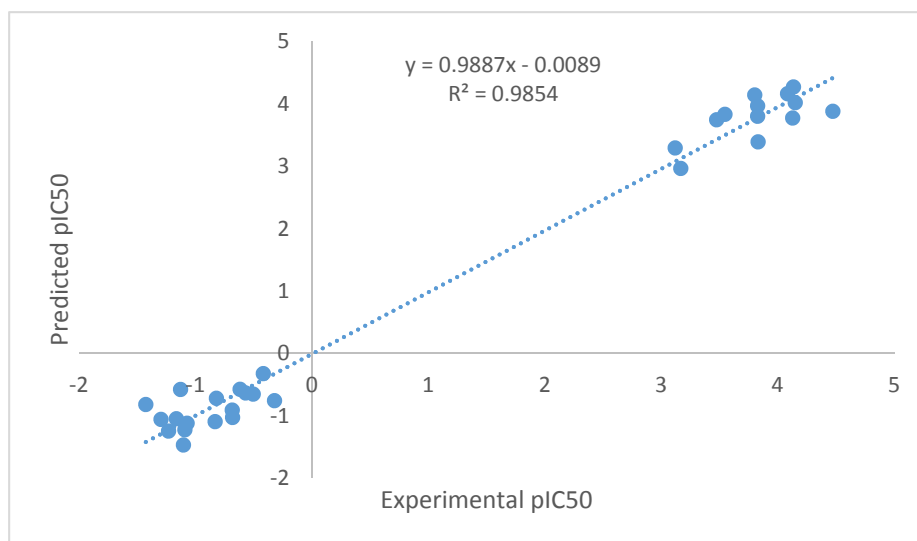


Figure 4: The relationships between predicted (training set) activities and the experimental activities

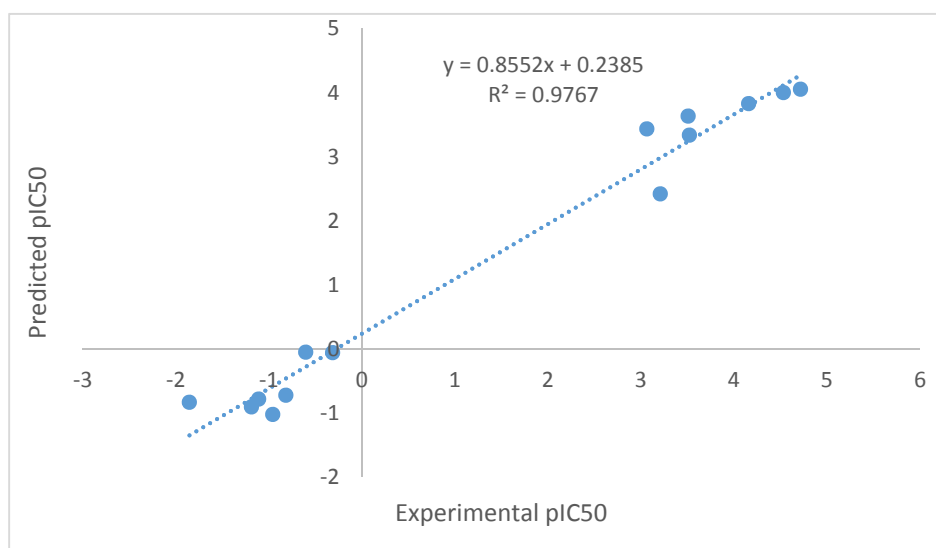


Figure 5: The relationships between predicted (test set) activities and the experimental activities

Figure 4 shows the plots of linear regression predicted versus experimental value of the biological activity of ketones and tetraketones derivatives outlined above. The plots for this model show to be more convenient with $R^2 = 0.9914$. It indicates that the model can be successfully applied to predict the tyrosinase and *Candida albicans* inhibitory activity of these compounds.

PRESS is an important cross-validation parameter as it is a good approximation of the real predictive error of the model. Its value being less than SSY points out that model predicts better than chance and can be considered statically significant. The smaller PRESS value means the better of the model predictability. From the results depicted in Table 4, the model is statistically significant. Also, for reasonable QSAR model, the PREES/SSY ratio should be lower than 0.4 [40]. The data presented in Table 4 indicate that for the developed model this ratio is 0.0420. Our result of Q^2 for this QSAR model has been to be 0.9853. The high value of Q^2 and R^2 adjusted are essential criteria for the best qualification of the QSAR model.

Table 4: Predictive error sum of squares (PRESS) and sum of the squared differences between the experimental responses and the average experimental responses

Parameter	Values	Threshold value
PRESS	0.2326	Low value

SSY	5.5323	Low value
PRESS/SSY	0.2326/5.5323	(0.0420) < 0.4

The developed model was further validated by a randomization technique (Table 5). The values of R^2_r and R^2 were determined, which were then used for calculating the value of R^2_p . Models with R^2_p values greater than 0.5 are considered statistically robust. If the value of R^2_p is less than 0.5, then it may be concluded that the outcome of the model is merely by chance, and it is not at all well predictive for truly external data sets. In this data set, values of R^2_p for all the 10 models were well above the stipulated value of 0.5. Therefore, it can be concluded that besides being robust, the model developed is well predictive.

Table 5: randomization technique

Model	R	R ²	Q ²
Original	0.9957	0.9914	0.9853
Random 1	0.3980	0.1584	-0.2202
Random 2	0.3561	0.1268	-0.4242
Random 3	0.3265	0.1066	-0.3425
Random 4	0.3507	0.1230	-0.3514
Random 5	0.3802	0.1446	-0.1967
Random 6	0.3579	0.1281	-0.2550
Random 7	0.3745	0.1403	-0.3623
Random 8	0.3686	0.1359	-0.3100
Random 9	0.6973	0.4863	0.2060
Random 10	0.5076	0.2577	-0.0753
Random Models Parameters			
Average r :			0.4118
Average r ² :			0.1808
Average Q ² :			-0.2331
cRp ² :			0.9027

The intercorrelation of the descriptors used in the QSAR model (1) was very low (below 0.6), which is in conformity to the study that, for a statistically significant model, it is necessary that the descriptors involved in the equation should not be intercorrelated with each other. 14 To further check the intercorrelation of descriptors, VIF analysis was performed. In this model, the VIF values of these descriptors are shown in Table 6 which are less than the threshold value of 10 [38].

Table 6: Specification of entered descriptors in genetic algorithm

Descriptors	Definition	MF	VIF
ALogP	Ghose-CrippenLogKow	-0.1038	1.7339
ATS4m	Broto-Moreau autocorrelation of lag 4 (log function) weighted by mass	1.5977	2.5526
AATSC5v	Average Broto-Moreau autocorrelation - lag 5 / weighted by van der Waals volumes	0.0382	2.4791
SHsNH2	Sum of atom-type H E-State: -NH2	-0.0079	1.1764
RotBtFrac	Fraction of rotatable bonds, excluding terminal bonds	-0.5241	2.4738

Satisfied with the robustness of the QSAR model developed using the training set, we have applied the QSAR model to an external data set of ketones and tetra-ketones 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. Table 7 represents the predicted pIC_{50} values of the test set based on equation (1). The overall root mean square error of prediction (RMSEP) between the experimental and predicted pIC_{50} values was 0.4823, which reveals good predictability. The estimated correlation coefficients between experimental and predicted pIC_{50} values with intercept (r^2) and without intercept (r'^2) were 0.9702 and 0.9767, respectively. The value of $[(r^2 - r'^2)/r^2] = 0.0097$, which is less than 0.1 stipulated value [34] 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 1.0896 and 0.8973, which are well

within the specified ranges of 0.85 and 1.15 [34]. The values of $R^2_{pred} = 0.9618$ and $rm^2(test) = 0.8981$ were found to be in the acceptable range [35], thereby indicating the good external predictability of the QSAR model.

Table 7: The predicted pIC_{50} values of the test set

Parameters	Values	Parameters	Values
r^2	0.9767	RMSEP	0.4823
r_0^2	0.9702	R^2_{pred}	0.9618
reverse r_0^2	0.9672	Q^2_{fl}	0.9618
$rm^2(test)$	0.8981	Q^2_{f2}	0.9614
reverse $rm^2(test)$	0.8818	$ r_0^2 - r^2 $	0.0030
average $rm^2(test)$	0.0163	$[(r^2 - r_0^2)/r^2]$	0.0066
delta $rm^2(test)$	0.0163	$[(r^2 - r_0^2)/r^2]$	0.0097
k	1.0896	k'	0.8973

For selecting the best model, values of $rm^2(overall)$ for the model was determined. As shown in Table 8, 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 really close to the observed values or not since high values of Q^2 and R^2_{pred} 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 0.5 [41]. 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 $rm^2(test)$ in Table 7 and difference between Q^2 and $rm^2(LOO)$ should be low. Large difference between the values will ultimately lead to poor values of $rm^2(overall)$ parameter. For this data set, the difference between Q^2 and $rm^2(LOO)$ is quite less (0.0097) and that between R^2_{pred} and $rm^2(test)$ is also very less (0.0800). Thus indicates that the model obtained for this data set using those descriptors are quite robust and predictive.

Table 8: Some internal and external Validation Parameters

Parameter	Without scaling	After scaling
$rm^2(LOO)$	0.9756	0.9767
$rm^2'(LOO)$	0.9820	0.9833
average $rm^2(LOO)$	0.9788	0.9800
delta $rm^2(LOO)$	0.0064	0.0067
$rm^2(overall)$	0.9641	0.9525
reverse $rm^2(overall)$	0.9540	0.9346
average $rm^2(overall)$	0.9590	0.9435
delta $rm^2(overall)$	0.0101	0.0179

The $rm^2(LOO)$ parameter in Table 8 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 $rm^2(test)$ (Table 8) determines the extent of deviation of the predicted activity from the experimental activity values of test set compounds where the predicted activity is calculated on the basis of the model developed using the corresponding training set. Equation 1 show acceptable values of $rm^2(LOO)$ and $rm^2(test)$ since they are greater than 0.5 [41].

3.1 Applicability domain of the developed QSAR model

Applicability domain (AD) is the physicochemical (e.g. structural or biological space, knowledge or information) on which the training set of the model has been developed. The resulting model can be reliably applicable for only those compounds which are inside this domain. This Euclidean based applicability domain helps to ensure that the compounds of the test set are representative of the training set compounds used in model development. It is based on distance scores calculated by the Euclidean distance norms. At first, normalized mean distance score for training set compounds are calculated and these values ranges from 0 to 1 (0=least diverse, 1=most diverse training set compound). Then normalized mean distance score for test set are calculated, and those test compounds with score outside 0 to 1 range are said to be outside the applicability domain. If the test set compounds are inside the domain/area covered by training set compounds that means these compounds are

inside the applicability domain otherwise not [39, 42, 43]. The training and test set compounds normalized mean distance score ranging from 0 to 1 are shown in Table 9 are all in the applicability domain.

Table 9: Euclidean based applicability domain

Training Set			
Compound Name	Distance Score	Mean Distance	Normalized Mean Distance
2	74121.69	2470.723	0.00447
3	73984.15	2466.138	0.002295
4	76680.14	2556.005	0.044929
5	80287.72	2676.257	0.101979
6	73838.99	2461.3	0
7	75898.53	2529.951	0.032569
8	75373.25	2512.442	0.024262
9	76342.4	2544.747	0.039588
10	73839	2461.3	1.83E-07
11	91510.4	3050.347	0.279451
12	95294.75	3176.492	0.339296
13	115166.2	3838.872	0.653537
15	101759.2	3391.975	0.441523
28	110569.3	3685.644	0.580844
22	109291.3	3643.044	0.560634
23	107087.7	3569.589	0.525786
24	137075.2	4569.172	1
29	94117.15	3137.238	0.320673
30	101626.9	3387.562	0.43943
31	116007.4	3866.913	0.66684
32	109857.9	3661.931	0.569594
33	100189.9	3339.663	0.416706
34	105899	3529.968	0.506989
35	86568.78	2885.626	0.201305
36	84770.2	2825.673	0.172863
38	93605.33	3120.178	0.31258
39	95640.27	3188.009	0.34476
40	83891.64	2796.388	0.15897
42	84360.47	2812.016	0.166384
43	121617.2	4053.905	0.755551
Test Set			
Compound Name	Distance Score	Mean Distance	Normalized Mean Distance
1	75067.74	2502.258	0.019431
14	123584	4119.465	0.786654
16	105058.2	3501.938	0.493692
17	112845.5	3761.517	0.616839
19	114288.1	3809.602	0.639651

20	95267.87	3175.596	0.338871
21	97838.78	3261.293	0.379526
25	109478.1	3649.269	0.563587
26	104103.4	3470.114	0.478594
27	92950.12	3098.337	0.302218
28	79212.74	2640.425	0.084979
37	115359	3845.301	0.656587
41	97042.9	3234.763	0.366941
44	97323.56	3244.119	0.371379

The leverage values can be calculated for every compound and plotted vs. standardized residuals, and it allows a graphical detection of both the outliers and the influential chemicals in a model. Figure 6, shows the Williams plot, the applicability domain is established inside a squared area within ± 3 bound for residuals and a leverage threshold h^* ($h^* = 3(p+1)/N$), where p is the total number of descriptors used for developing of QSAR model, while N is the total number of the training set compounds [44]. It demonstrates that all the compounds of the training set and test set are inside of the square area. It is obvious that all compounds in the test set fall inside the domain of the GA-MLR model (the warning leverage limit is 0.6). There are only one chemicals (No. 10 in the training set and No. 20, 21 and No. 41 in the test set) which have the leverage higher than the warning h^* value, so they can be regarded as structural outliers.

Luckily, in this case the data predicted by the model are good for compound numbers 4, 12, 26 and 44, therefore, they are “good leverage” chemicals. For all the compounds in the training and test sets, their standardized residuals are smaller than three standard deviation units (3δ).

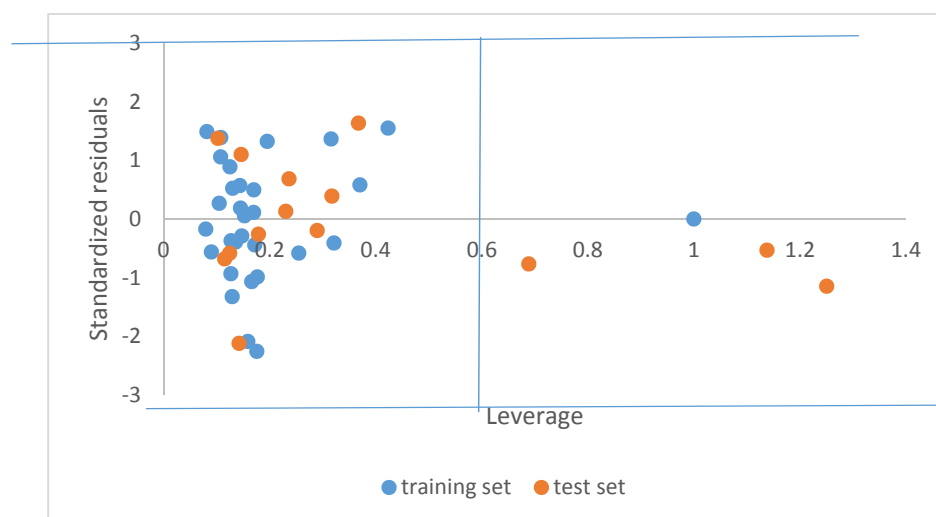


Figure 6: The Williams plot, the plot of the standardized residuals versus the leverage value

3.2 Molecular descriptors interpretations

By interpreting the descriptors contained in the QSAR model, it is possible to gain some insights into factors, which are related to the anti-*Candida albicans* and tyrosinase inhibitors activity. For this reason, an acceptable interpretation of the selected descriptors is provided below. The brief descriptions of descriptors are shown in Table 6. 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 [45, 46]. The MF value indicates the relative importance of a descriptor, compared with the other descriptors in the model. Its sign indicates the variation direction in the values of the activities as a result of the increase or decrease of the descriptor values. The mean effect (MF) values are given in Table 6.

Ghose-CrippenLogKow (ALogP) is a thermodynamic descriptor. ALogP is the partition coefficient calculated using atom based approach and represents hydrophobicity of the molecules [47]. A positive mean effect of this descriptor illustrate that the activity increases with increasing the value of ALogP, which mean that the anti-*Candida albicans* and tyrosinase inhibitors activity is directly related to this descriptor. This property assumes significances in the present case because of the fact that the molecules under study contain lipophilic groups.

ATS4m (Broto-Moreau autocorrelation of lag 4 (log function) weighted by mass) which is a 2D autocorrelation descriptor. In this descriptor the Moran coefficient is a distance-type function, and is any physicochemical property calculated for each atom of the molecule, for example atomic mass, polarizability, etc. The Moran coefficient usually takes a value in the interval [-1, +1]. Positive autocorrelation corresponds to positive values of the coefficient whereas negative autocorrelation produces negative values. Therefore, the molecule atoms represent a set of discrete points in space and the atomic property is the function evaluated at those points. The physicochemical property in this case is the atomic mass. ATS4m has a positive sign, illustrating a greater mean effect value than that of the other descriptors, which indicates that this descriptor had a significant effect on the activity and that the pIC_{50} value is directly related to this descriptor. Hence, it was concluded that by increasing the molecular mass the value of this descriptor increased, causing an increase in its pIC_{50} value.

AATSC5v (Average Broto-Moreau autocorrelation - lag 5 / weighted by van der Waals volumes) belong to 2D autocorrelation descriptors. The 2D autocorrelation descriptors have been successfully employed by Fernandez and coworker [48, 49]. In these descriptors, the molecule atoms represent a set of discrete points in space, and the atomic property and function are evaluated at those points. The physicochemical property for AATSC5v descriptor is atomic van der Waals volumes, which relate to volume of the molecule. Therefore, increasing the volume of a molecule increases AATSC5v value. Mean effect of AATSC5v has the positive sign, which indicates that an increase in the volume of molecule leads to an increase in its activity.

SHsNH2 (Sum of atom-type H E-State: -NH2) is one of the Electro-topological state (E-state) descriptors. The electro-topological state is a method for describing and encoding molecular structure at the atom level. It gives information related to the electronic and topological state of the atom in the molecule. SHsNH2 is the sum of Hydrogen E-states for the Nitrogen atom with two hydrogens attached and one single bond to a non-hydrogen atom. The SHsNH2 mean effect has a positive sign. This sign suggests that the activity is directly related to this descriptor.

RotBtFrac (Fraction of rotatable bonds, excluding terminal bonds) belong to the 2D PaDEL Rotatable Bonds Count Descriptor. This descriptor signifies the number of rotatable bonds. RotBtFrac is the number of bonds in the molecule having rotations that are considered to be meaningful for molecular mechanics. All terminal H-atoms are ignored. The RotBtFrac mean effect has a negative sign. This sign suggests that the activity is indirectly related to the descriptor.

4. Conclusion

The present work shows how a set of antifungal (anti-*Candida albicans*) and anti-tyrosinase activities of various ketones and tetraketones may be treated statistically to uncover the molecular characteristics which are essential for high activity. The generated models were analyzed and validated for their statistical significance and external prediction power. The awareness and understanding of the descriptors involved in antifungal and anti-tyrosinase activity of these compounds could provide a great opportunity for the ligand structures design with appropriate features, and for the explanation of the way in which these features affect the biological data upon binding to the respective receptor target. The results derived may be useful in further designing more novel anti-tyrosinase and anti-*Candida Albicans* agents in series.

5. Recommendations

It is suggested that further in-vivo research on ketone and tetraketone derivatives to know the best compounds that have a lower dose with the highest activity as antifungal (anti-tyrosinase and anti-*Candida Albicans* agent). Future work may include examining the approach with more datasets with different activities. Taking into account the multidimensional model proposed, the selected properties and the proposed of one possible mechanism of action for the compounds studied, we propose for future research on Docking and synthesis of the organic compounds.

Conflict of interest

The authors confirm that this article content has no conflicts of interest.

References

- [1] Parveen I, Threadgill MD, Moorby JM, Winters A (2010). Oxidative phenols in forage crops containing polyphenol oxidase enzymes. *J Agric Food Chem.* 58:1371–82.
- [2] van Gelder CW, Flurkey WH, Wichers HJ (1997). Sequence and structural features of plant and fungal tyrosinases. *Phytochemistry.* 45: 1309–23.
- [3] Hearing VJ (2011). Determination of melanin synthetic pathways. *J Invest Dermatol.* 131:8–11.
- [4] Chang T (2009). An Updated Review of Tyrosinase Inhibitors. *Int. J. Mol. Sci.* 10, 2440-2475.
- [5] Slominski A, Tobin DJ, Shibahara S, Wortsman J (2004). Melanin pigmentation in mammalian skin and its hormonal regulation. *Physiol Rev.* 84:1155–228.
- [6] Khan MTH (2007). Molecular design of tyrosinase inhibitors: A critical review of promising novel inhibitors from synthetic origins. *Pure Appl. Chem.* 2007, 79: 2277-2295.
- [7] Briganti S, Camera E, Picardo M (2003). Chemical and instrumental approaches to treat hyperpigmentation. *Pigment Cell Res.* 16: 101-110.
- [8] Rendon MI, Gaviria JI (2005). Review of skin-lightening agents. *Dermatol. Surg.* 31: 886-889

- [9] Draelos ZD (2007). Skin lightening preparations and the hydroquinone controversy. *Dermatol. Ther.* 20: 308-313.
- [10] Parvez S, Kang M, Chung HS, Bae H (2007). Naturally occurring Tyrosinase inhibitors: mechanism and applications in skin health, cosmetics and agriculture industries. *Phytother. Res.* 21: 805-816.
- [11] Lee SY, Baek N, and Nam T-G (2015). Natural, semisynthetic and synthetic tyrosinase inhibitors. *J Enzyme Inhib Med Chem*, Early Online: 1–13.
- [12] Ryan KJ, Ray CG (editors) (2004). *Sherris Medical Microbiology* (4th ed.). McGraw Hill.
- [13] Parahitiyawa NB, Samaranyake YH, Samaranyake LP, Ye J, Tsang PW, Cheung BP, Yau JY, Yeung SK (2006). Interspecies variation in *Candida* biofilm formation studied using the Calgary biofilm device. *APMIS*. 114: 298-306.
- [14] Enfert C; Hube B (editors) (2007). *Candida: Comparative and Functional Genomics*. Caister Academic Press.
- [15] Edmond MB, Wallace SE, McClish DK, Pfaller MA, Jones RN, Wenzel RP (1999). Nosocomial bloodstream infections in United States hospitals: a three-year analysis. *Clin Infect Dis.* 29:239–244.
- [16] Mishra NN, Prasad T, Sharma N, Payasi A, Prasad R, Gupta, DK, Singh R (2007). Pathogenicity and drug resistance in *Candida albicans* and other yeast species. *ActaMicrobiol. Immunol. Hung.* 54:201-235.
- [17] Ramage G, Saville SP, Thomas DP, Lopez-Ribot JL (2005). *Candida* biofilms: an update. *Eukar. Cell.* 4: 633-638.
- [18] Baker JA, Beehler GP, Sawant AC, Jayaprakash V, McCann SE, Moysich KB (2006). Consumption of coffee, but not black tea, is associated with decreased risk of premenopausal breast cancer. *J. Nutr.* 136: 166-171.
- [19] Boroujeni HAR, Pirbalouti AG, Hamed B, Abdzadeh R, Malekpoor F (2012). Anti-*Candida* activity of ethanolic extracts of Iranian endemic medicinal herbs against *Candida albicans*. *J. Med. Plant Res.* 6: 2448-2452.
- [20] Zore GB, Thakre AD, Jadhav S, Karuppaiyl SM (2011). Terpenoids inhibit *Candida albicans* growth by affecting membrane integrity and arrest of cell cycle. *Phytomed.* 18: 1181–1190.
- [21] Sardi JCO, Scorzoni L, BernardiFusco-Almeida TAM and Mendes GianniniMJS (2013). *Candida* species: current epidemiology, pathogenicity, biofilm formation, natural antifungal products and new therapeutic options. *Journal of Medical Microbiology*, 62: 10–24.
- [22] Cavaleiro C, Pinto E, Goncalves MJ & Salgueiro LR (2006). Antifungal activity of *Juniperus* essential oils against dermatophyte, *Aspergillus* and *Candida* strains. *J ApplMicrobiol* 100: 1333–1338.
- [23] Hansch C, Kurup A, Garg R, Gao H, (2001). Chem-Bioinformatics and QSAR: A Review of QSAR Lacking Positive Hydrophobic Terms. *Chemical Review.* 101: 619–672.
- [24] Adane L, Bharatam PV, (2008). V. Modelling and informatics in the analysis of P-falciparum DHFR Enzyme inhibitors, *Curr. Med. Chem.* 15: 1552–1569.
- [25] Deshpande S, Solomon VR, Katti SB and Prabhakar YS (2009). Topological descriptors in modelling antimalarial activity: N1-(7-chloro-4-quinolyl)-1,4-bis(3aminopropyl)piperazine as prototype. *Journal of Enzyme Inhibition and Medicinal Chemistry.* 24(1): 94-104.
- [26] Khan KM, Maharvi GM, Khan MTH, Shaikh AJ, Parveen S, Begum S, and Choudhary MI (2006). Tetraketones: A new class of tyrosinase inhibitors. *Bioorganic & Medicinal Chemistry* 14: 344-351.
- [27] Motta LF and Almeida WP (2011). Quantitative structure-activity relationships (QSAR) of a series of ketone derivatives as anti-*Candida albicans*. *International Journal of Drug Discovery* 3(2): 100-117.
- [28] ACD-Lab software for calculating the referred physicochemical parameters: Chemsketech 12.0, www.acdlabs.com
- [29] Wavefunction, (2013). Inc. Spartan'14, version 1.1.2, Irvine, California, USA.
- [30] Yap CW (2011). PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput Chem.* 32(7): 1466-74.
- [31] Karelson M, Karelson G, Tamm T, Tulp I, Jänes J, Tamm K, Lomaka A, Savchenko D, and Dobchev D (2009). QSAR study of pharmacological permeabilities. *ARKIVOC* 2(ii): 218-238.
- [32] Material Studio (modeling and simulation solutions for chemicals and materials research software) version 7.0 which was downloaded from <http://www.accelrys.com/products/materials-studio>.
- [33] Ballabio D, Consonni V, Mauri A, Claeys-Bruno M, Sergent M, Todeschini R (2014). "A novel variable reduction method adapted from space-filling designs." *Chemometrics and Intelligent Laboratory Systems* 13: 147-154.
- [34] Golbraikh A, Tropsha A (2002): Beware of q²! *J Mol Graph Model.* 20:269-276.
- [35] Roy PP, Roy K (2008). On some aspects of variable selection for partial least squares regression models. *QSAR Comb Sci* 27: 302-313.
- [36] Roy K, Paul S (2008). Exploring 2D and 3D QSARs of 2,4-diphenyl-1,3-oxazolines for ovicidal activity against Tetranychusurticae. *QSAR Comb Sci* 28: 406-425.
- [37] Jaiswal M, Khadikar PV, Scozzafava A, Supuran CT (2004). Carbonic anhydrase inhibitors: the first QSAR study on inhibition of tumor-associated isoenzyme IX with aromatic and heterocyclic sulfonamides. *Bioorg Med Chem Lett* 14: 3283-3290.
- [38] Shapiro S, Guggenheim B (1998). Inhibition of oral bacteria by phenolic compounds: Part I. QSAR analysis using molecular connectivity. *Quant Struct Act Relat* 17: 327-337.
- [39] Edache EI, Uzairu A and Abechi SE (2016). Multi-target in-silico study of 5,6-dihydro-2-pyrones, indole β -diketo acid, diketo acid and carboxamide derivatives against various anti-HIV-1 strain at PM3 semi-empirical level. *Ew J Pharm*, 1(1): 1-13.
- [40] Roy PP, Paul S, Mitra I, and Roy K (2009). On two novel parameter for validation of predictive QSAR models. *Molecules*, 14: 1660-1701.
- [41] Tropsha A, Gramatica P, Gombar V (2003). The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR Models. *QSAR & Combinatorial Science.* 22: 69–77.
- [42] Edache EI, Uzairu A, Abechi SE (2015). Multivariate QSAR Study of Indole β -Diketo Acid, Diketo Acid and Carboxamide Derivatives as Potent Anti-HIV Agents. *International Journal of Innovative Research & Development*, vol 4 Issue 9: 374-390.
- [43] Edache EI, Uzairu A, Abechi SE (2016). Development and Estimation of an in Silico Model for Anti-HIV-1 Integrase Inhibitor Using Genetic Function Approximation. *Journal of Advances in Medical and Pharmaceutical*, 5(2): 1-18.
- [44] Minovski N, Zuperl S, Drgan V, Novi M (2013). Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum Euclidean distance space analysis: A case study. *Analytica Chimica Acta* 759: 28–42.
- [45] Riahi S, Pourbasheer E, Ganjali MR, Norouzi P (2009). Investigation of different linear and nonlinear chemometric methods for modeling of retention index of essential oil components: concerns to support vector machine. *J. Hazard. Mater.* 166: 853-859.
- [46] Pourbasheer E, Riahi S, Ganjali MR, Norouzi P (2009). Application of genetic algorithm-support vector machine (GA-SVM) for prediction of BK-channels activity. *Eur. J. Med. Chem.* 44: 5023-5028.
- [47] Todeschini R, and Consonni V (2000). *Handbook of Molecular Descriptors*. Wiley-VCH, Weinheim, Germany.
- [48] Fernandez M, Tudidor-Camba A, Caballero J (2005). Modeling of cyclin-dependent kinase inhibition by 1H-pyrazolo[3,4-d]pyrimidine derivatives using artificial neural network ensembles. *J. Chem. Inform. Model.* 45: 1884-1895.

- [49] Caballero J, Tundidor-Camba A, Fernandez M (2007). Modeling of the inhibition constant (K_i) of some cruzain ketone-based inhibitors using 2D spatial autocorrelation vectors and data-diverse ensembles of Bayesian-regularized genetic neural networks. QSAR Comb. Sci. 26: 27-40.