

Theoretical Investigation of Correlations Between Molecular and Electronic Structure and Antifungal Activity in Coumarin Derivatives: Combining Qsar and Dft Studies

Abduljelil Ajala^{1*}, Adamu Uzairu¹, Idris O. Suleiman¹ and Ahmed Jibrin Uttu²

¹*Department of Chemistry, Ahmadu Bello University, Zaria, Kaduna, Nigeria.*

²*Department of Chemistry, Federal University Gashua, Yobe State, Nigeria.*

Authors' contributions

This work was carried out in collaboration among all authors. Author AU designed the study and wrote the protocol. Author AA managed the animals, collected all data, performed the statistical analysis and wrote the first draft of the manuscript. Authors IOS and AA did the literature search and also wrote part of the manuscript. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/JAMPS/2018/22801

Editor(s):

(1) Jinyong Peng, Professor, College of Pharmacy, Dalian Medical University, Dalian, China.

Reviewers:

(1) Joseph Sloop, School of Science and Technology, Georgia Gwinnett College, USA.

(2) Vishal W. Banewar, Government Vidarbha Institute of Science & Humanities, India.

(3) Bharat Singh, Institute of Biotechnology, Amity University Rajasthan, India.

Complete Peer review History: <http://www.sciencedomain.org/review-history/24042>

Original Research Article

Received 28th October 2015

Accepted 18th January 2016

Published 7th April 2018

ABSTRACT

Quantitative structure-activity relationship (QSAR) models were combined with density functional computations and used to predict anti-fungi activities in a series of coumarin derivatives. Essential descriptors employed in this study were chosen based on the use of the Genetic Function Approximation (GFA) method. Leave-N-Out (LNO) and Y-randomization techniques affirmed the model's robustness and validity. Computed pMIC values were found to be in good agreement (+/- XX%) with experimentally determined values. The proposed model may be a superior predictor of the counter-parasitic action of coumarin analogs and can be utilized for recommendation of new chemopreventive species.

*Corresponding author: E-mail: abdulajala39@gmail.com;

Keywords: Gfa; dft; ketone analogues; qsar; antifungal.

1. INTRODUCTION

Numerous skin illnesses, for example, tinea and ringworm brought on by dermatophytes exist in tropical and semitropical areas. In general, these parasites live in the dead, top layer of skin cells in soggy ranges of the body, for example, between the toes, the crotch, and under the bosoms. These contagious diseases cause just a minor aggravation. Different sorts of contagious contaminations could be more genuine. They can enter into the toes and cause tingling, swelling, rankling and scaling. Now and again, parasitic diseases can bring about responses somewhere else in the body. For instance, a man may build up a rash on the finger or hand in the wake of coming into contact with a tainted foot.

The incidence of opportunistic fungal infections in patients treated with immuno suppressive drugs, intensive chemotherapy, suffering from AIDS and neonates is increasing at an alarming rate [1,2]. These mycoses are very difficult to eradicate constituting an enormous challenge for health care providers [3]. Although there appear to be an array of drugs for the treatment of systemic and superficial mycoses, none of them is ideal in terms of efficacy, safety or antifungal spectrum [4,5]. Many of the drugs have undesirable effects or are very toxic (amphotericin B), produce recurrence, show drug-drug interactions (azoles) or lead to the development of resistance (fluconazole, 5-flucytosine) [6].

The QSAR is one of the most vital areas in chemometrics and is a valuable tool that is used

extensively in drug design and medicinal chemistry [7-10]. Chemical and Biological effects are related closely to molecular properties which can be calculated or predicted by their structure using various methods [11]. Once a reliable QSAR model is established, we can predict the activities of molecules and know which structural features play a significant role in the biological processes. In this study, we use genetic function approximation (GFA), a statistical modeling algorithm to build a functional model of experimental data. Since its inception, several applications of this algorithm in the area of quantitative structure activity relationship modeling have been reported [12].

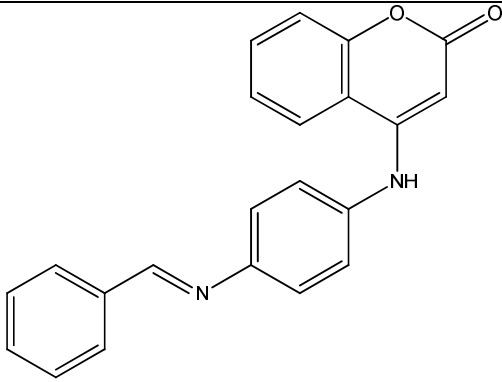
The purpose of the present work is to perform a quantum chemical QSAR study on the series of coumarin derivatives in Table 1 [13] to compare the computed values with the experimental activities of the compounds as Antifungal Agents..." and obtain a linear model by using Genetic function Approximation (GFA) method.

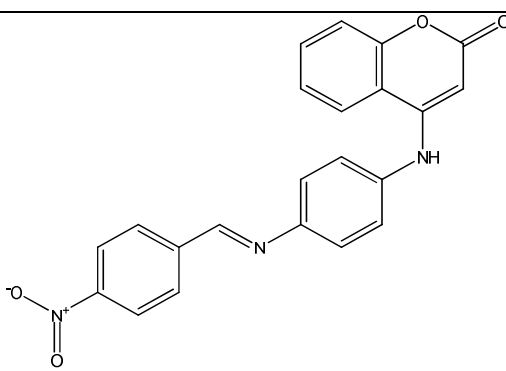
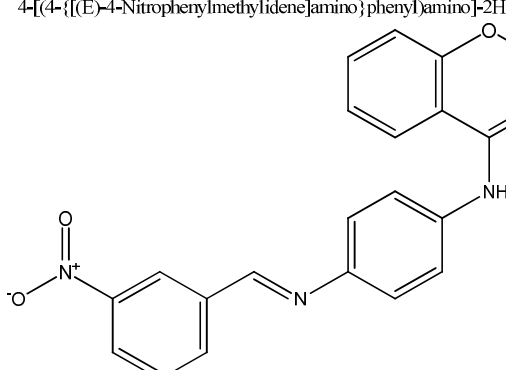
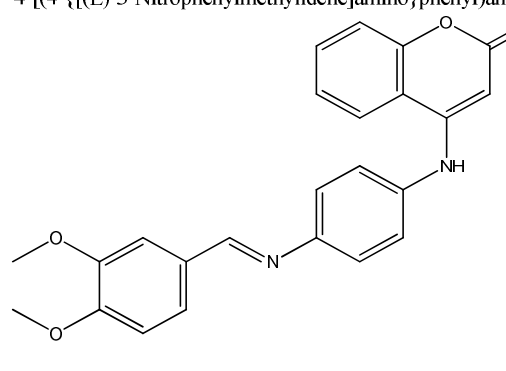
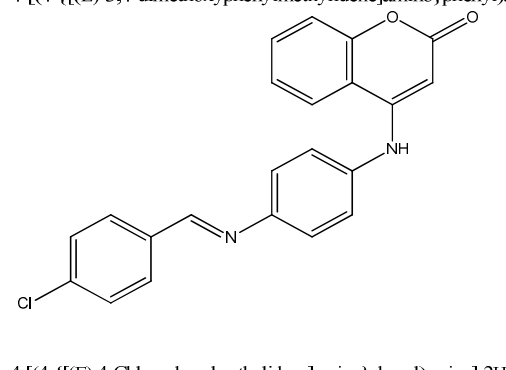
2. MATERIALS AND METHODS

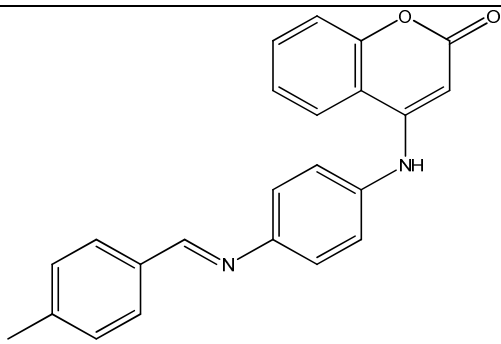
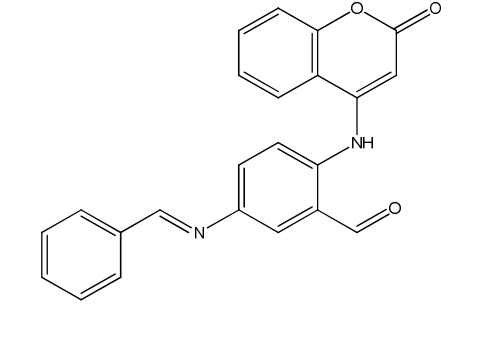
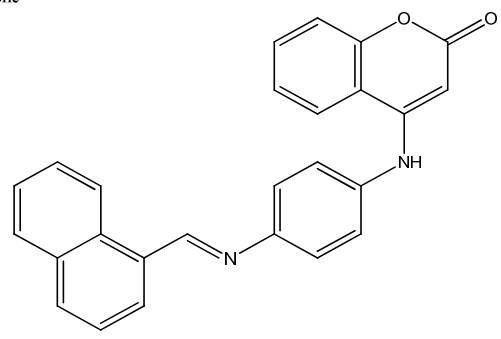
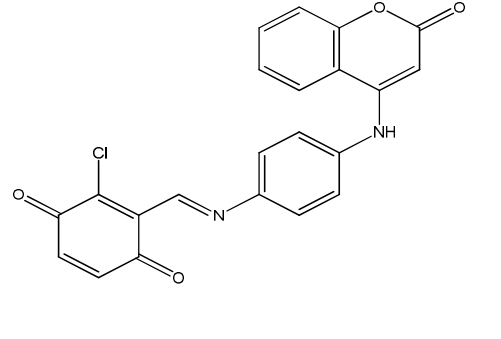
2.1 Chemical Data

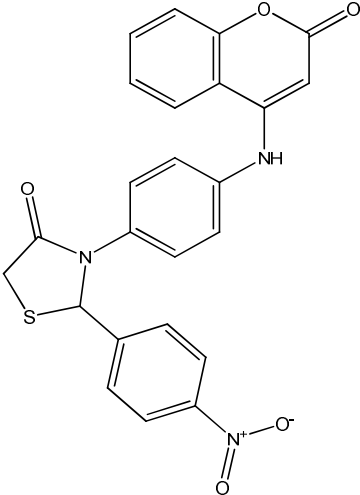
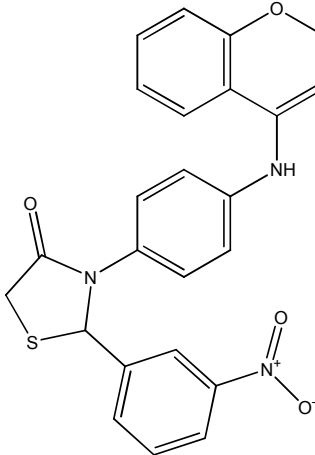
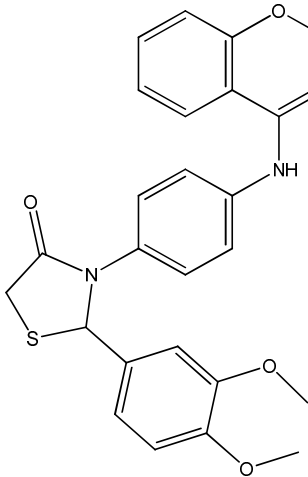
Biological data on the activity of coumarin derivatives has been obtained [13] and is reported in Table 1. The activity data refers to pMIC (-log MIC), which indicates the experimentally determined biological activity of the compounds necessary for the inhibition of *candida albicans*. Fluconazole was used in Table 1 as an assay control.

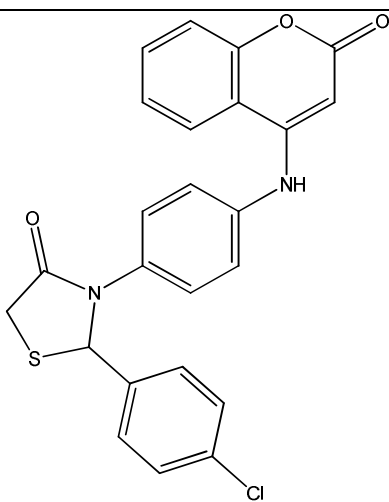
Table 1. Data set from the literature [13] used in the Quantum Chemical QSAR analysis

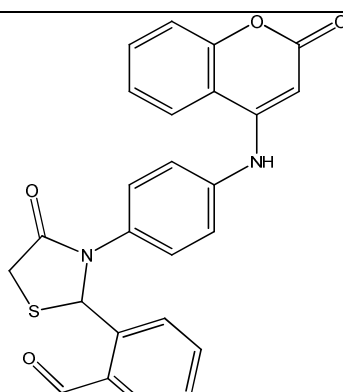
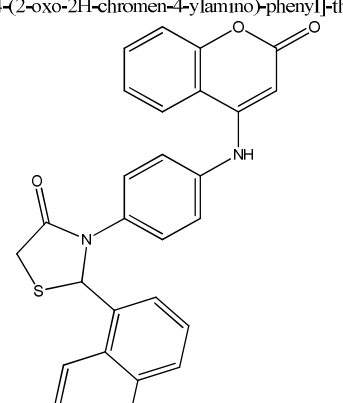
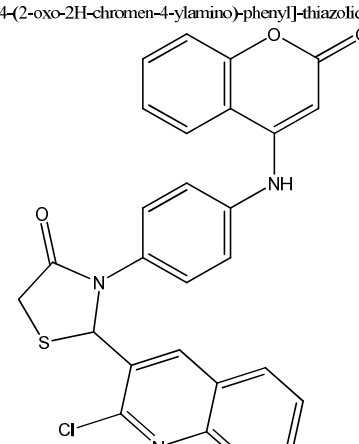
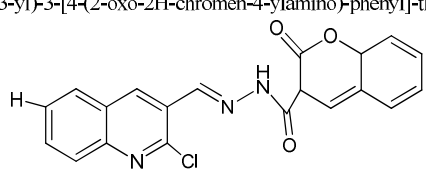
S/no	Structures	MIC ($\mu\text{g/mL}$)	pMIC
1	 <p>4-[(4-[(E)-phenylmethylidene]amino]phenyl)amino]-2H-chromen-2-one</p>	500	2.83

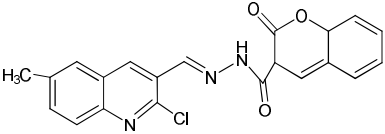
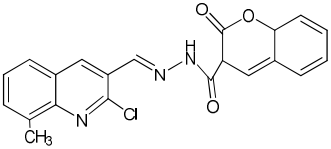
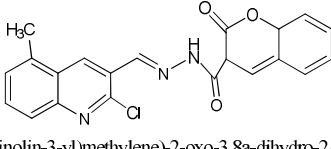
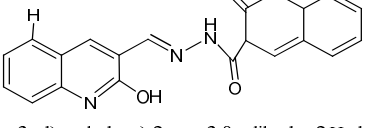
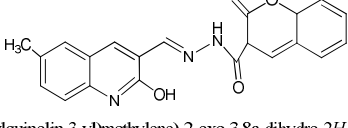
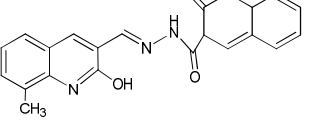
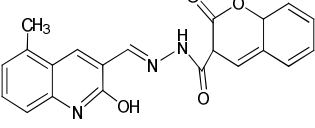
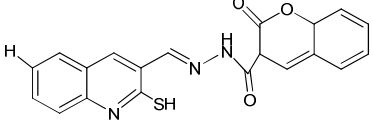
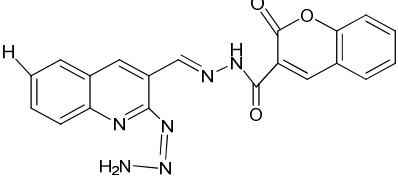
2		800	2.68
3	<p>4-[(4-{(E)-4-Nitrophenylmethylidene}amino)phenyl]amino]-2H-chromen-2-one</p> 	500	2.89
4	<p>4-[(4-{(E)-3-Nitrophenylmethylidene}amino)phenyl]amino]-2H-chromen-2-one</p> 	100	3.60
5	<p>4-[(4-{(E)-3,4-dimethoxyphenylmethylidene}amino)phenyl]amino]-2H-chromen-2-one</p> 	500	2.87
	<p>4-[(4-{(E)-4-Chlorophenylmethylidene}amino)phenyl]amino]-2H-chromen-2-one</p>		

6		200	3.25
7	4-[(4-[(E)-4-Methylphenylmethylidene]amino)phenyl]amino]-2H-chromen-2-one 	500	2.87
8	4-[(4-[(E)-phenylmethylidene]amino)phenyl-2-Carboxaldehyde]amino]-2H-chromen-2-one 	400	2.99
9	4-[(4-[(E)-Naphthylmethylidene]amino)phenyl]amino]-2H-chromen-2-one 	400	3.01
	4-[(4-[(E)-2-Chloroquinonylmethylidene]amino)phenyl]amino]-2H-chromen-2-one		

10	 <chem>O=C1CN(C1Sc2ccc(cc2[N+](=O)[O-])c3ccc(Nc4c(=O)oc5ccccc45)cc3)C(=O)O</chem>	800	2.71
11	 <chem>O=C1CN(C1Sc2cccc(c2[N+](=O)[O-])c3ccc(Nc4c(=O)oc5ccccc45)cc3)C(=O)O</chem>	250	3.26
12	 <chem>COC1=CC(OC)=CC=C1C2CN(C2Sc3ccc(Nc4c(=O)oc5ccccc45)cc3)C(=O)O</chem>	1000	2.68

13		200	3.35
14	2-(4-Chloro-phenyl)-3-[4-(2-oxo-2H-chromen-4-ylamino)-phenyl]-thiazolidin-4-one	250	3.23
15	3-[4-(2-Oxo-2H-chromen-4-ylamino)-phenyl]-2-p-tolyl-thiazolidin-4-one	200	3.30
3-[4-(2-Oxo-2H-chromen-4-ylamino)-phenyl]-2-p-tolyl-thiazolidin-4-one			

16		200	3.34
17	<p>2-(4-oxo-3-[4-(2-oxo-2H-chromen-4-ylamino)-phenyl]-thiazolidin-2-yl)-benzaldehyde</p> 	250	3.27
18	<p>2-Naphthalen-1-yl-3-[4-(2-oxo-2H-chromen-4-ylamino)-phenyl]-thiazolidin-4-one</p> 	200	3.40
19	<p>2-(2-Chloro-quinolin-3-yl)-3-[4-(2-oxo-2H-chromen-4-ylamino)-phenyl]-thiazolidin-4-one</p>  <p><i>(E)</i>-<i>N</i>-((2-chloroquinolin-3-yl)methylene)-2-oxo-3,8a-dihydro-2<i>H</i>-chromene-3-carbohydrazide</p>	1000	2.58

20		1000	2.59
21	<i>(E)-N'-(2-chloro-6-methylquinolin-3-yl)methylene)-2-oxo-3,8a-dihydro-2H-chromene-3-carbohydrazide</i>	1000	2.59
22		500	2.89
23	<i>(E)-N'-(2-chloro-8-methylquinolin-3-yl)methylene)-2-oxo-3,8a-dihydro-2H-chromene-3-carbohydrazide</i>	500	2.89
24		250	3.16
25	<i>(E)-N'-(2-chloro-5-methylquinolin-3-yl)methylene)-2-oxo-3,8a-dihydro-2H-chromene-3-carbohydrazide</i>	250	3.16
26		250	3.17
27	<i>(E)-N'-(2-hydroxyquinolin-3-yl)methylene)-2-oxo-3,8a-dihydro-2H-chromene-3-carbohydrazide</i>	250	3.17
28		200	3.00
29	<i>(E)-N'-(2-hydroxy-6-methylquinolin-3-yl)methylene)-2-oxo-3,8a-dihydro-2H-chromene-3-carbohydrazide</i>	200	3.00
30		500	2.87
31	<i>(E)-N'-(2-hydroxy-8-methylquinolin-3-yl)methylene)-2-oxo-3,8a-dihydro-2H-chromene-3-carbohydrazide</i>	500	2.87
32		1000	2.57
33	<i>(E)-N'-(2-hydroxy-5-methylquinolin-3-yl)methylene)-2-oxo-3,8a-dihydro-2H-chromene-3-carbohydrazide</i>	1000	2.57
34		1000	2.58
35	<i>(E)-N'-(2-mercaptoquinolin-3-yl)methylene)-2-oxo-3,8a-dihydro-2H-chromene-3-carbohydrazide</i>	1000	2.58
36			
37	<i>(E)-2-oxo-N'-(2-((Z)-triaz-1-en-1-yl)quinolin-3-yl)methylene)-2H-chromene-3-carbohydrazide</i>		

$pMIC = -\log MIC = \text{Minimum inhibitory concentration in molar}$

Biological data on the activity of Coumarin derivatives has been obtained from Jayakumar Swamy et al, Divyesh Patel et al

2.2 Computational and Statistical Details

QSAR studies of coumarin derivatives was carried out on windows 7, Intel CORE i5 operating system by Spartan' 14v112 for windows, Macintosh and Linux. PaDEL-Descriptor (A software to calculate molecular descriptors and fingerprints), version: 2.21 and Chem3D pro, version 12.0.2 1076. The molecular structures of the dataset was sketched using Chem Draw Ultra, version 12.0.2.1076 developed by CambridgeSoft and Materials Studio V8.0.0.843 copyright(c) 2014 for the statistical analysis.

The molecular geometry of all the derivatives from the dataset (Table 1) was determined via energy minimization [14] and geometry optimization using a Merck Molecular Force Field (MMFF) with the B3LYP/6-311+G(d) method. This same density functional level was used to study QSAR [15-18]."

2.3 Calculation

Molecular modeling software programs perform mathematical calculations to determine several physical and chemical properties of molecules. In general, computational chemistry programs attempt to find a solution to the Schrödinger equation, as follows:

$$\hat{H}\Psi=E\Psi \quad (1)$$

This function is an eigensystem in which the Hamiltonian operator applied to the wave function (Ψ) of the molecule is equal to an energy term (E) multiplied by the wave function, which is a mathematical expression that describes the system. The energy term in this function adjusts for the relative locations of electrons and the nucleus, and all interactions between them, while the operator is simply a function that is applied to the wave function. Once solved, the wave function can be used to calculate most properties of an atom or molecule's geometry, behavior, or energy. On the other hand, on account of the complexity of electron communications inside of a particle and the increment in this many-sided quality when atoms consolidate to shape atoms, it is difficult to locate a careful answer for the Schrödinger comparison and in this way fluctuating degrees of guess must be connected. The PC project performs every one of the figurings to tackle such mathematical statements,

and gives the critical information on the atomic framework.

2.4 Density Functional Theory (DFT)

Density function theory is ab initio estimation for a solution to the Schrödinger equation that depends on electron density around the nuclei in a molecule, rather than wave functions that other methods use. DFT is a well known technique for estimating giant molecule in the computational science world on the grounds that it has a high precision for every unit PC processor (CPU) time proportion in examination to different systems [19].

The Genetic Function Approximation (GFA) calculation offers another way to deal with the issue of building quantitative structure activity relationship (QSAR) and quantitative structure property relationship (QSPR) models. Supplanting relapse investigation with the GFA calculation empowers the development of models focused with or better than those delivered by standard procedures and makes accessible extra data not gave by different methods. Not at all like most different investigation calculations, GFA gives various models, where the populaces of the models are made by advancing irregular starting models utilizing a hereditary calculation. GFA can manufacture models utilizing straight polynomials as well as higher request polynomials, splines, and other nonlinear capacities. The Genetic Function Approximation calculations are hunt calculations that take motivation from regular hereditary qualities and development. In this segment, the thoughts hidden Genetic Function Approximation calculations are quickly depicted, underlining the angles important to the genetic function approximation (GFA) way to deal with model building. The GFA calculation itself applies these thoughts to the issue of capacity approximation [20,21] given a substantial number of potential elements impacting a reaction, including a few forces and different elements of the crude inputs, to discover the subset of terms that corresponds best with the reaction. The focal thoughts of calculations are straightforward. The locale to be looked is coded into one or different strings. In the GFA, these strings are sets of terms forces and splines of the crude inputs. Every string speaks to an area in the inquiry space. The calculation works with an arrangement of these strings, called a populace. This populace is advanced in a way that leads it toward the goal of the pursuit. This requires that

a measure of the wellness of every string, comparing to a model in the GFA, be accessible.

Table 2. Parameters for energy minimization

Parameters	Value
Force Field	MMFF(Merck Molecular Force Field)
Maximum no of Cycles	100,000
Convergence Criteria	0.001cal/molÅ
Dielectric constant	1(in a vacuum)
Gradient Type	Analytical

Taking after this, three operations are performed iteratively in progression: determination, hybrid and change. Recently added individuals are scored by wellness paradigm. In the GFA, the scoring criteria for models are all identified with the nature of the relapse fit to the information. The choice probabilities must be re-assessed every time another part is added to the populace.

Steadiness and meeting in the same way as other iterative minimization calculations, there are issues with the strength and union of the GFA calculation. A sign of the solidness of the GFA calculation can be got by producing a plot demonstrating the development of variable utilization withtime. Such a plot demonstrates the quantity of events of every variable in the populace for every era of the development. For pragmatic reasons, to decrease the measure of information that would be gathered, such a plot is produced just for those variables that happen most regularly in the last populace and the information are not ordinarily gathered for each era. The GFA calculation is accepted to have met when no change is found in the score of the populace over a noteworthy time allotment, either that of the best model in every populace or the normal of the considerable number of models in every populace. At the point when this rule has been fulfilled, no further eras are figure.

3. RESULTS AND DISCUSSION

In a general sense, QSAR addresses two inquiries: what auxiliary and electronic properties

of an atom decide its movement and what can be modified to enhance this action?

Computational devices permit analysts to recognize chemicals with ideal physico compound properties in silico, before costly experimentation. This saves both time and money,permitting the exclusion of subpar competitors without expending research facility resources [22].

3.1 QSAR Study

To examine the Observed data, the conveyance of the information must be initially explored. Most relapse calculation depends on the information that is in effect typically examined, on the off chance that the information are not ordinarily disseminated, we ought to think about applying as a numerical change to accomplish an ordinary circulation. Observed data in Table 4 show acceptable normal distribution, so no need to perform a numerical transformation. Table 4 shows a univariate analysis for the actual data. Table 4 contains several statistical measures that describe the actual data. The most important parameters in Table 4 are the skewness and kurtosis. Skewness is the third moment of the distribution, which indicates the symmetry of the distribution.

Developing a QSAR model is a procedure that takes an arrangement of inputs and gives an arrangement of yields. For instance, a vitality minimization is a model which takes a structure as data and gives an enhanced structure as yield. In a normal QSAR study, estimation of descriptors happens. These are models which take a solitary structure as a data and give a solitary number or gathering of firmly related numbers as yields. Table 5 Shows the experimental pMIC and the predicted pMIC using the GFA approach of the training set. This shows how the GFA method predicted the pMIC.

Table 3. List of descriptors used in this study

Descriptors	Type	Significance	VIF
SHBint5	Electrotopological state	Strength for potential hydrogen bonds of path lenght5	1.099
WD.Unity	WHIM Descriptor	A non-directional WHIM,weighted by unit weights	2.144
Wlambda2.eneg	WHIM Descriptor	A non-directional WHIM,weighted by mulliken atomic electronegativities.	1.188
Wlambda3.polar	WHIM Descriptor	Non-directional WHIM, weighted by atomic polarizabilities	1.267

Table 4. Univariate analysis of the observed data

		Column A
1	Number of sample points	21
2	Range	1.03000000
3	Maximum	3.60000000
4	Minimum	2.57000000
5	Mean	2.99380952
6	Median	2.89000000
7	Variance	0.09055690
8	Standard deviation	0.30835800
9	Mean absolute deviation	0.26054400
10	Skewness	0.23096600
11	Kurtosis	-1.27567000

Table 6 shows the GFA guess investigation which gives a synopsis of the information parameters utilized for the count. Additionally, it reports whether the GFA calculation united in the predetermined number of eras. Merging is accomplished when there has been no change in the scoring capacity for various eras. It can be seen from Table 6 that the accuracy of the model, indicated by the R^2 value, is reasonably high therefore the predictive power of the model, as indicated by the adjusted R^2 and cross validated R^2 values, is also, high, even though the regression is significant according to F-test. In Table 6 the Friedman's lack-of-fit (LOF) score [23-25], which evaluates the QSAR model by considering the number of descriptors as well as the quality of fitness, is chosen: the lower the LOF, the less likely it is that GFA model will fit the data.

Use of the Friedman lack-of-fit (LOF) measure has several benefits over the regular least square error measure. In Materials Studio [25] LOF is measured using a slight variation of the original Friedman formula [26] The revised formula is:

$$LOF = \frac{SSE}{\left(1 - \frac{c+dp}{M}\right)^2} \quad (2)$$

Where SSE is the sum of squares of errors, c is the number of terms in the model, other than the constant term, d is a user defined smoothing parameter, p is the total number of descriptors contained in all model terms (again ignoring the constant term) and M is the number of samples in the training set. Unlike the commonly used least squares measure, the LOF measure cannot always be reduced by adding more terms to the regression model. While the new term may reduce the SSE, it also increases the values of c and p, which tends to increase the

LOF score. Thus, adding a new term may reduce the SSE, but actually increases the LOF score. By limiting the tendency to simply add more terms, the LOF measure resists overfitting better than the SSE measure [26]. The Friedman's lack-of-fit (LOF) score in Table 6 evaluates the QSAR model. The lower the LOF, the less likely it is that GFA model will fit the data. The significant regression is given by F-test and the higher the value the better the model.

$$\text{Best model: } Y = 0.04703a + 3.017924b - 0.05263c + 0.01986d - 1.13568 \quad (3)$$

a = SHBint5, b = WD.unity, c = wlambda2.eneg and d = wlambda3.polar

The multi-collinearity between the above five descriptors was detected by calculating their variation inflation factors (VIF), which can be calculated as follows:

$$VIF = \frac{1}{1-R^2} \quad (4)$$

Where R^2 is the correlation coefficient of the multiple regression between the variables within the model. If VIF equals to 1, then no inter-correlation exists for each variables; and 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 [27]. The corresponding VIF values of the five descriptors are presented in Table 3 As can be seen from this table, all the variables have VIF values of less than five, indicating that the obtained model has statistical significance and the descriptors were found to be reasonably orthogonal [27].

3.2 QSAR Model Validation

The real usefulness of QSAR models is not just their ability to reproduce known data, verification by their fitting power (R^2), but mainly is their potential for predictive application. For this reason, the internal consistency of the training set was confirmed by using leave-oneout (LOO) cross-validation method to ensure the robustness of the model. The high calculated Q^2_{LOO} value, 0.7321 suggests a good internal validation. A second validation method was also developed on the basis of a leave-group-out (LGO) internal cross-validation method. In this case, a group of compounds in. Leave-N-out (LNO) crossvalidation, [28,30,31] known also as leave-many-out, is highly recommended to test the robustness of a model. The training set of M samples is divided into consecutive blocks of N

samples, where the first N define the first block, the following N samples is the second block, and so on. This way, the number of blocks is the integer of the ratio M/N if M is a multiple of N ; otherwise the left out samples usually make the last block. This test is based on the same basic principles as LOO: each block is excluded once, a new model is built without it, and the values of the dependent variable are predicted for the block in question. LNO is performed for $N = 2, 3$, etc., and the leave- N -out crossvalidated correlation coefficients Q^2_{LNO} are calculated in the same way as for LOO. LNO can be performed in two modes: keeping the same number of factors for each value of N (determined by LOO for the real model) or with the optimum number of factors determined by each model. LNO is sensitive to the order of samples in the data set. For example, leave-two-out crossvalidation for even M means that $M/2$ models are obtained, but this is only a small fraction $(0.5 \cdot (M - 1) - 1)$ of all possible combinations of two samples $M!/(M - 2)! = M(M - 1)$. To avoid any systematic variation of descriptors through a data set or some subset what would affect LNO, the samples should be randomly ordered (in X and Y simultaneously). It is recommended that N represents a significant fraction of samples (like leave-20 to 50% - out for smaller data sets [35]. In this research, we have done leave-10-out which yield $Q^2 = 0.9295$ and $SDEP = 0.1978$.

3.3 y-Randomization

The purpose of the y-randomization test [28-34] is to distinguish and evaluate chance connections between's the reliant variable and descriptors. In this setting, the term chance connection implies that the genuine model may contain descriptors which are factually very much corresponded to y however as a general rule there is no reason impact relationship encoded in the separate relationships with y on the grounds that they are not identified with the component of activity. Two fundamental inquiries can be raised in regards to y-randomization: how to break down the outcomes from every randomization run and what number of runs ought to be carried out? There are different ways to judge whether the genuine model is described by a chance relationship. The straightforward methodology of Eriksson and Wold [33] can be summarized as a set of decision inequalities based on the values of Q^2_{yrand} and R^2_{yrand} and their relationship $R^2_{yrand} > Q^2_{yrand}$

$Q^2_{yrand} < 0.2$ and $R^2_{yrand} < 0.2 \rightarrow$ no chance correlation;

Any Q^2_{yrand} and $0.2 < R^2_{yrand} < 0.3 \rightarrow$ negligible chance correlation;

any Q^2_{yrand} and $0.3 < R^2_{yrand} < 0.4 \rightarrow$ tolerable chance correlation;
any

Table 5. Experimental pMIC and GFA Predicted pMIC for the training set

	Actual values	Predicted values	Residual values
1	2.68000000	2.74047800	-0.06047800
2	2.89000000	2.90442700	-0.01442700
3	3.60000000	3.55309600	0.04690400
4	2.87000000	2.85161400	0.01838600
5	3.25000000	3.23910800	0.01089200
6	2.87000000	2.96696900	-0.09696900
7	2.99000000	2.99727500	-0.00727500
8	3.01000000	2.94309900	0.06690100
9	2.71000000	2.69563800	0.01436200
10	2.68000000	2.58562300	0.09437700
11	3.35000000	3.16478300	0.18521700
12	3.30000000	3.38182300	-0.08182300
13	3.34000000	3.43747900	-0.09747900
14	3.27000000	3.28108900	-0.01108900
15	3.40000000	3.35432400	0.04567600
16	2.58000000	2.59206700	-0.01206700
17	2.59000000	2.77082600	-0.18082600
18	2.89000000	2.72726800	0.16273200
19	3.16000000	3.04156800	0.11843200
20	2.87000000	3.00249100	-0.13249100
21	2.57000000	2.63895400	-0.06895400

Q^2_{yrand} and $R^2_{\text{yrand}} > 0.4 \rightarrow$ recognized chance correlation.

Therefore, the correlation's frequency is counted as the number of randomizations which resulted in models with spurious correlations (falsely good), which is easily visible in a Q^2_{yrand} against R^2_{yrand} plot.

3.4 Model Applicability Domain and Determination of Outlier towards the Model

Applicability domain is simply defined as the response and chemical space in which a QSAR/QSPR model makes prediction with a given dependability. It includes the physicochemical domain, the descriptor domain, chemical domain, which is termed the chemical space or structural domain and metabolic domain. A model can only be put to use for screening compound (Insilico screening) if its applicability domain is defined and prediction for only those compounds that fall in this domain can be considered reliable [36]. The applicability domain of the model in this work will be defined using the leverage approach. It involves the construction of a hat matrix **H**

$$H_{tr} (n \times n) = X_{tr} \cdot (X_{tr}^T X_{tr})^{-1} \cdot X_{tr}^T \quad (5)$$

$$H_{tr} = X_{tr} (X_{tr}^T X_{tr})^{-1} X_{tr}^T \rightarrow \hat{y}_{tr} = H_{tr} y_{tr} \leftrightarrow \hat{y}_{tr} = X_{tr} (X_{tr}^T X_{tr})^{-1} X_{tr}^T y_{tr} \quad (6)$$

$$H_{ext} = X_{ext} (X_{tr}^T X_{tr})^{-1} X_{tr}^T \rightarrow \hat{y}_{ext} = H_{ext} y_{ext} \leftrightarrow \hat{y}_{ext} = X_{ext} (X_{tr}^T X_{tr})^{-1} X_{tr}^T y_{ext} \quad (7)$$

which maps the vector of observed values to the vector of fitted values. It is an $n \times n$ symmetric matrix which diagonal elements h_{ii} (known as leverage values) which directly reflect the structural influence of a compound to the values predicted by the model (i.e., a distance metric which shows how far a compound is from the model experimental space) [35]. In linear modelling, the leverage ranges between $1/n$ and 1 and the average leverage for all compound in the training set is $(k+1)/n$. Therefore, warning leverage (cut-off leverage) value will be calculated using the relation below:

$$h^* = \frac{3(k+1)}{n} \quad (8)$$

Where h^* the warning leverage, k is the number of descriptor in the model and n is the number of observation that make up the training set. When h_{ii} values of any molecule is lower than h^* the

molecule is said to be structurally similar to all molecule that made up the test set, but if h_{ii} is greater than h^* it means the molecule is structurally distant from all other molecule in the training set and it predicted data by the model may be unreliable [35]. However if a compound has h_{ii} greater than h^* it will reinforce the model if the compound is in the training set. If such compound is in test set it may not appear to be an outlier because its residual may be low. To determine vividly that a compound is an outlier, leverage values and standardized residual are used together in what is known as Williams graph to describe the applicability domain of a given model. Williams graph is a plot of standardized residual as the ordinate (y-axis) against the leverage as the abscissa (x-axis). The standardized residual of all (training, test and evaluation set) the compounds will be calculated using the equation below

$$SDR = \frac{\hat{y}_i - y_i}{RMSE} \quad (9)$$

All ($h_{ii,te}$ or ext , $h_{ii,tr}$) leverages will be extracted and a plot of the standardized residual (SDR) against the leverages will be made. On that plot along the abscissa the cut-off leverage value will be made the boundary. Along the ordinate $\pm 2.5 - 3.5$ standard deviation units will be used as boundary because points that lie within a ± 3 standardised residual from the mean covers 99% of the normally distributed data [37-39]. Any compounds with cross-validated standard residual larger than 2.5 standard deviation unit are considered as outlier.

The Fig. 1 shows, all compounds of the training set and test set are inside an area bounded by 3 standardized residuals (Y-axis) and the warning h^* value of 0.714 (X-axis) with the exception of compound X. This means that all evaluated compounds except one have a leverage lower than the warning h value of 0.714.

Table 6. Validation of the genetic function approximation

		Equation 1:
1	Friedman LOF	0.05023100
2	R^2	0.90643100
3	R^2_{adi}	0.88303900
4	$R^2(\text{cv})$	0.73206600
5	Significant Regression	Yes
6	Significance of F-value	38.74939300
7	Critical SOR F-value (95%)	3.05581800
8	Lack-of-fit points	16
9	Min expt. error for non-significant LOF (95%)	0.08216200

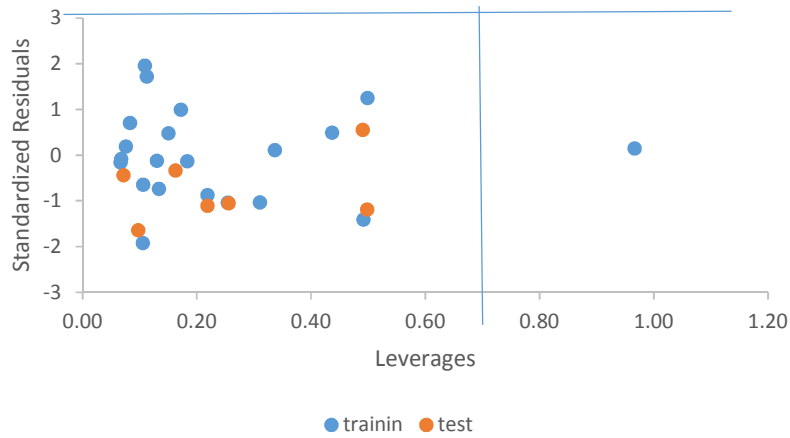


Fig. 1. Williams plot of the standardized residuals versus the leverage values

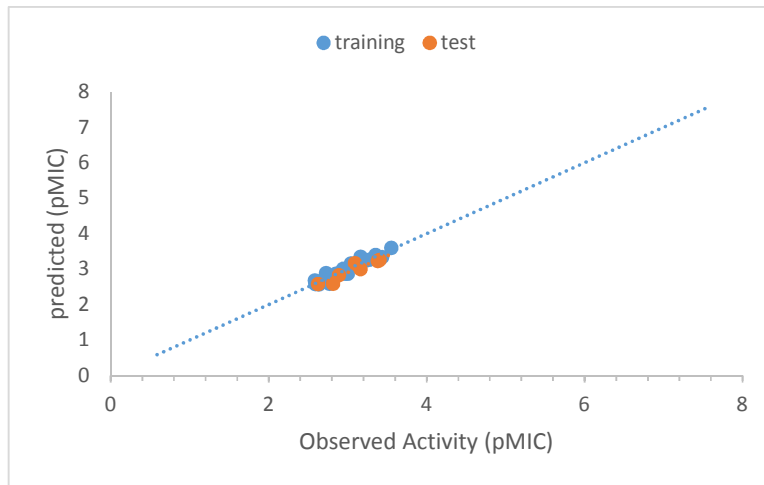


Fig. 2. Predicted pMIC versus Observed pMIC for the training and test sets

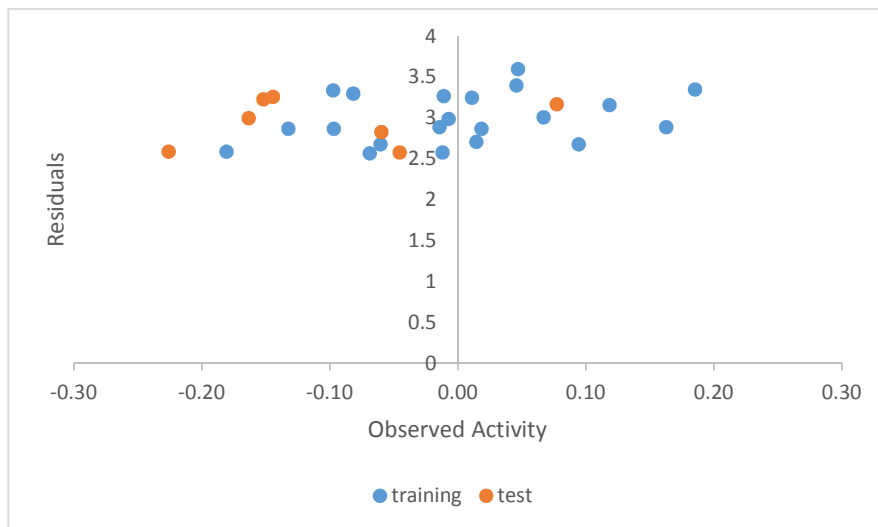


Fig. 3. The residuals versus observed activity values for the training and test sets

Table 7. Experimental pMIC and GFA predicted pMIC for test set

Compound	Observed activity	Predicted activity	Residuals
1	2.83	2.889829	-0.05983
11	3.26	3.404563	-0.14456
14	3.23	3.382011	-0.15201
20	2.59	2.816044	-0.22604
24	3.17	3.092799	0.077201
25	3	3.163509	-0.16351
27	2.58	2.625505	-0.04551

Fig. 2 shows a relation between the predicted values against observed activity for training and tests set using the equation 3.

Also, Fig. 3 shows the distribution of the residual values against the Observed activity values for training and tests set. A residual can be defined as the difference between the predicted value in the generated model and the measured value for the Observed activity.

To test the constructed QSAR model, potential outliers have been identified in Fig. 4a-b. An

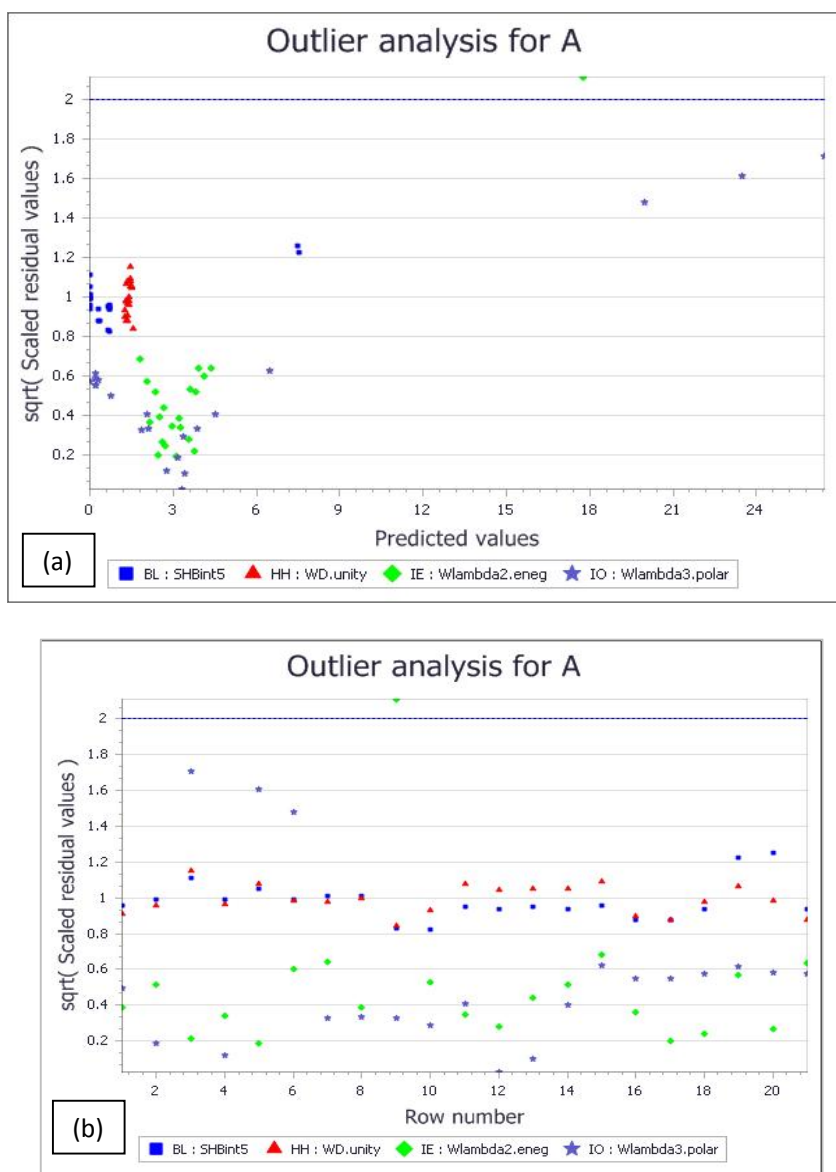


Fig. 4. Outlier analysis

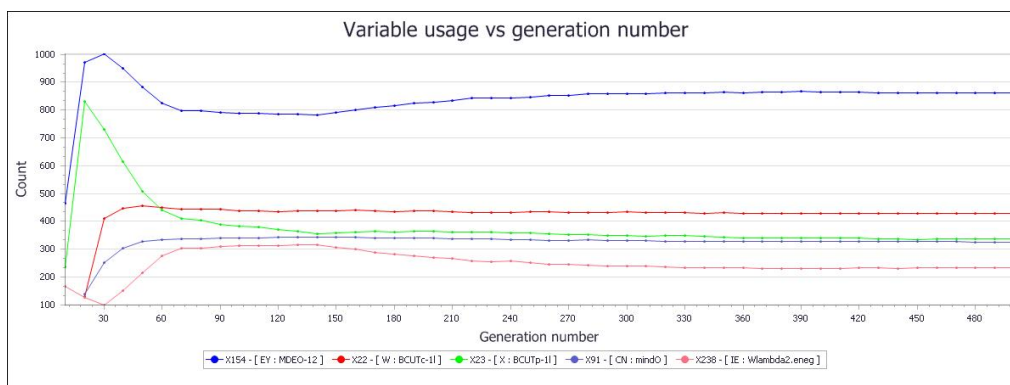


Fig. 5. Variable usage count against generation number

outlier can be characterized as an information point whose remaining worth is not inside of two standard deviations of the mean of the leftover qualities. Despite the fact that the quantity of exceptions can differ contingent upon the nature of the dataset (e.g., erroneous estimations of physical properties or errors in molecular structures will diminish the information set quality), it still a decent test of QSAR model is to distinguish potential outliers. Fig. 4a-b contains two charts. One contains the residual values (a) plotted against the Observed pMIC and others shows the remaining qualities (b) plotted against Table 1 raw number. Each chart contains a dotted line that indicates the critical threshold of two standard deviations beyond which a value may be considered to an outlier. Inspection of Fig. 4a-b shows that there is no points appeared outside the dotted lines which make the QSAR model acceptable.

In Fig. 5, the Y-axis represents the different molecular descriptors used in this study as shown on the left side of the graph. On the other hand, the X-axis represents the number of the generations we could generate for each of these molecular descriptors. According to Fig. 5, at each step, the GFA uses the current population to create the children that makes up the next generation.

The algorithm selects a group of individuals in the current population called parents who contribute their genes the entries of their vectors to their children. The algorithm usually selects individuals that have better fitness values as parents. User can specify the function that the algorithm uses to select the parents. The GFA creates three types of children for the next generation: Elite Children, Crossover

Children and Mutation Children. In our QSAR study, the algorithm stops when the number of generations reaches the value of 500 Generations.

4. CONCLUSION

A genetic function approximation method was used to run the regression analysis and establish correlation's between different types of descriptors and measured Chemotherapy activities of Coumarin derivatives. These models were validated by means of leaveoneout cross-validation, leave-*N*-out crossvalidation, external validation, y-randomization and applicability domain.

The constructed model was assessed comprehensively (internal and external validations), and all the validation indicate that the QSAR model we constructed was robust and satisfactory. Selection of four variables showed that the SHBint5, WD.Unity, wlambda2.eneg and wlambda3.polar of the molecule play a main role in the predicting the antifungal activity of Coumarin derivatives.

CONSENT

It is not applicable.

ETHICAL APPROVAL

It is not applicable.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

- Groll A, Shah P, Mentzel C, Schneider M, Just-Neubling G, Huebner K. Trends in the postmortem epidemiology of invasive fungal infections at a university hospital. *Journal of Infection*. 1996;3:23–32.
- Denning D, Evans E, Kibbler C, Richardson MD, Roberts MM, Rogers TR, Warnock DW, Warren RE. Guidelines for the investigation of invasive fungal infections in haematological malignancy and solid organ transplantation. *European Journal of Clinical Microbiology & Infectious Diseases*. 1997;16:424–436.
- Meyers JD. Fungal infections in bone marrow transplant patients. In *Seminars in oncology*. 1990;17(3) (Suppl 6):10-13.
- DiDomenico B. Novel antifungal drugs. *Current opinion in microbiology*. 1999;2(5): 509-515.
- Ablordeppey SY, Fan P, Ablordeppey JH, Mardenborough L. Systemic antifungal agents against AIDS-related opportunistic infections: Current status and emerging drugs in development. *Current medicinal chemistry*. 1999;6(12):1151-1196.
- White TC, Marr KA, Bowden RA. Clinical, cellular and molecular factors that contribute to antifungal drug resistance. *Clinical microbiology reviews*. 1998;11(2): 382-402.
- Manly CJ, Louise-May S, Hammer JD. The impact of informatics and computational chemistry on synthesis and screening. *Drug discovery today*. 2001;6(21):1101-1110.
- Pourbasheer E, Riahi S, Ganjali MR, Norouzi P. Quantitative structure–activity relationship (QSAR) study of interleukin-1 receptor associated kinase 4 (IRAK-4) inhibitor activity by the genetic algorithm and multiple linear regression (GA-MLR) method. *Journal of enzyme inhibition and medicinal chemistry*. 2010;25(6):844-853.
- Pourbasheer E, Riahi S, Ganjali MR, Norouzi P. QSAR study of C allosteric binding site of HCV NS5B polymerase inhibitors by support vector machine. *Molecular diversity*. 2011;15(3):645-653.
- Delley B. An all-electron numerical method for solving the local density functional for polyatomic molecules. *The Journal of chemical physics*. 1990;92(1):508-517.
- Burger A, Abraham DJ. *Burger’s medicinal chemistry and drug discovery*. Wiley, Hoboken, New Jersey; 2003.
- Rogers D. Approximation with comparison to evolutionary techniques. *Genetic algorithms in molecular modeling*. 1996;87.
- Seema B, Ramar S, Degani MS. Synthesis and biology evaluation of α , β -unsaturated ketone as potential antifungal agents. *Medicinal Chemistry Research*. 2009;18: 309-316
- Williams DA, Lemke TLF. *Principles of Medicinal Chemistry* (Ed). Lippincott Williams Wilkins, Baltimore. 2002;81.
- Haigren TA, Merck molecular force field-11.MMFF94 van der Walls and electrostatic parameters for intermolecular interactions, *Journal of Computational Chemistry*. 1996;17:520–552.
- Haigren TA. Merck molecular force field-111. Molecular Geometries and vibrational frequencies for MMFF94, *Journal of Computational Chemistry*. 1996;17:553–586.
- Haigren TA. Merck molecular force field-1 V. Conformational energies and geometries for MMFF94, *Journal of Computational Chemistry*. 1996;17:587–615.
- Halgren TA. Merck molecular force field-1. Basis, form, scope, parametrization and performance of MMFF94, *Journal of Computational Chemistry*. 1996;17:490–519.
- Young DC. *Computational chemistry: A practical guide for applying techniques to real-world problems*. John Wiley & Sons, Inc. 605 Third Avenue, New York, NY; 2001.
- Rogers D. *Evolutionary statistics: Using a genetic algorithm and model*. East Lansing, MI, Morgan Kaufmann, San Francisco. for Computational Statistics, Department of Statistics, Stanford University: Stanford; 1997.
- Rogers D. *Evolutionary statistics: Using a genetic algorithm and model*. East Lansing, MI, Morgan Kaufmann, San Francisco. For Computational Statistics, Department of Statistics, Stanford University: Stanford; 1994.
- Friedman JH. Multivariate adaptive regression splines. *The annals of statistics*. 1991;1:67.
- Friedman JH. Multivariate adaptive regression splines Technical Report Laboratory. 1991;102.

24. Khaled KF. Modeling corrosion inhibition of iron in acid medium by genetic function approximation method: A QSAR model. *Corrosion Science*. 2011;53(11):3457-3465.
25. Materials Studio V8.0.0.843. Copyright (c) 2014.
26. Jaiswal M, Khadikar PV, Scozzafava A, Supuran CT. Carbonic anhydrase inhibitors: The first QSAR study on inhibition of tumor-associated isoenzyme IX with aromatic and heterocyclic sulphonamides. *Bioorganic and Medicinal Chemistry Lett*. 2004;14:3283-3290.
27. Shapiro S, Guggenheim B. Inhibition of oral bacteria by phenolic compounds. Part 1. QSAR analysis using molecular connectivity. *Quantitative Structure-Activity Relationships*. 1998;17(04):327-337.
28. Baumann K, Stiefl NJ. *Journal of Computer-Aided Molecular Design*. 2004; 18:549.
29. Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. *Environmental health perspectives*. 2003;111(10):1361.
30. Clark RD, Fox PC. Statistical variation in progressive scrambling. *Journal of Computer-Aided Molecular Design*. 2004; 18(7-9):563-576.
31. Baumann K, Cross-validation as the objective function for variable-selection techniques. *TrAC Trends in Analytical Chemistry*. 2003;22(6):395-406.
32. Teófilo RF, Martins JPA, Ferreira MMC. *Journal of Chemometrics*. 2009;23:32.
33. Wold S, Eriksson L. In chemometric methods in molecular design, H. van water beemd, VCH: Weinheim. 1995;309.
34. Rücker C, Rücker G, Meringer M. *Journal of Chemical Information Model*. 2007;47: 2345.
35. Gramatica P. Principle of QSAR model validation: Internal and external. *QSAR Combinatorial science*. 2007;26(5):694-701.
36. Tropsha A, Golbraith A. Predictive quantitative structure and activities relationship modelling data preparation and general modelling workflow: In handbook of chemoinformatics algorithms, 1st Edition, mathematical and computational biology series, Chapman and Hall/CRC books; 2010.
37. Jawoska JS, Nikola TN, Aldenberg T. QSAR applicability domain estimation by projection of the training set in descriptors space: A review. *ALTA Alteran .Lab Anim*. 2005;33:445-459.
38. Tahereh A, Shayessteh D, Ali-Muhammad HS, Jahan BG, Sarkosh M. QSAR models for CXCR2 receptor antagonist based on genetic algorithm for data processing prior to application of the PLS linear regression method and design of new compound using insilico virtual screening. *Molecules*. 2011;16:1928-1955.
39. Minovski N, Zuperl S, Drgan V, Novic M. Assessment of applicability domain for multivariate counter-propagation artificial neural network predictive models by minimum Euclidean distance space analysis: A case study. *Analytical Chimica Acta*. 2013;759:28-42.

© 2018 Ajalaa et al.; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here:
<http://www.sciencedomain.org/review-history/24042>