

**THEORITICAL INVESTIGATION OF MOLECULAR AND ELECTRONIC
STRUCTURE OF ANTIFUNGAL ACTIVITY OF KETONE COMPOUNDS:
COMBINING QSAR AND DFT STUDIES.**

ABSTRACT

Quantitative structure-activity relationship (qsar) models were displayed for modeling and predicting anti-fungi activities . The accomplishment of any quantitative structure–activity relationship model relies on upon the precision of the info information, choice of proper descriptors, measurable devices and the acceptance of the created model. The essential descriptors were chosen with the guide of the Genetic function approximation(GFA) method.The got model was approved utilizing $R^2_{(cv)}=0.732$, $R^2_{(pred)} = 0.74$, $LOF = 0.05023$, $R^2 = 0.9064$, $R^2_{(adj)} = 0.8831$, $F = 38.7494$, $df, sdep, SSY, RMSE(cv), RMSEP = 0.13807, RMSECV = 0.09205$, $RMSEC = 0.105457$, $PRESS = 0.133447$ $Q = 9.028$, $RP^2 = 0.8119$.Leave-N-out(LNO) and Y-randomization techniques affirmed the model robustness.The proposed model may give a superior comprehension of the counter parasitic action of ketones and it analogs and can be utilized as direction for recommendation of new chemopreventive operator.

Key words: gfa, dft, ketone analogues, qsar, antifungal.

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 pores 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 immunosuppressive drugs, intensive chemotherapy, suffering AIDS and neonates is increasing at an alarming rate [1,2]. These mycoses are very difficult to eradicate constituting an enormous challenge for healthcare 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 kinds of 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, genetic function approximation (GFA) which is a statistical modeling algorithm which builds functional models of experimental data. Since its inception, several applications of this algorithm in the area of quantitative structure–activity relationship modeling have been reported [12].

MATERIALS AND METHODS

The purpose of the present work is to perform a quantum chemical QSAR study of the Ketone derivatives on the series [13], to investigate the experimental activities of the compounds as Antifungal Agents and obtain a linear model by using Genetic function Approximation (GFA) method.

Chemical Data:

Biological data on the activity of Ketone derivatives has been obtained [13]. (Table 1). The activity data refers pMIC, which indicates the biological activity of compounds experimentally determined, necessary for the inhibition of *candidaalbicans* resistant. The $-\log$ MIC(molar) scale refers pMIC. Fluconazole for Table 1: was used as controls in the assays

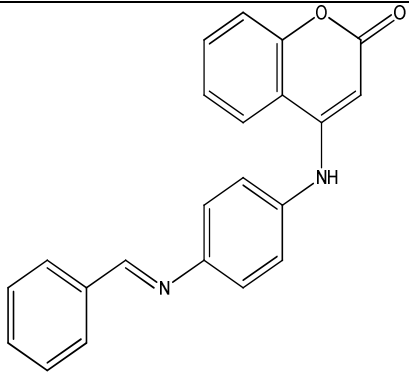
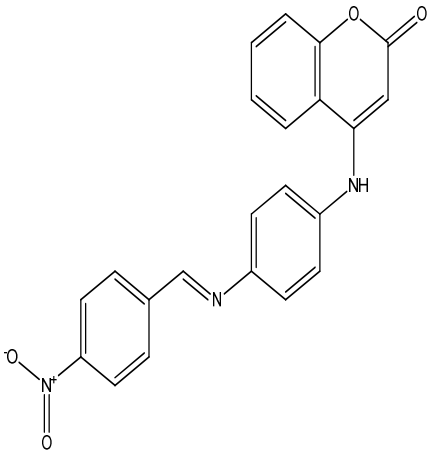
56 Table 1.

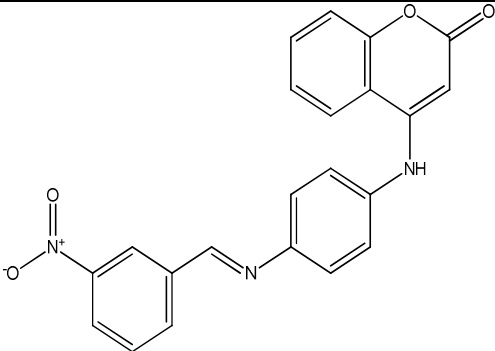
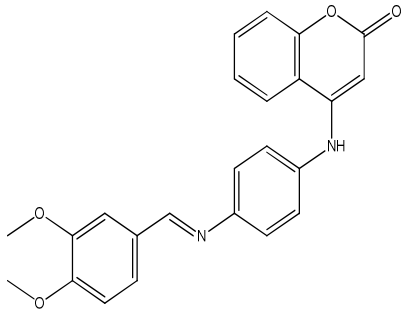
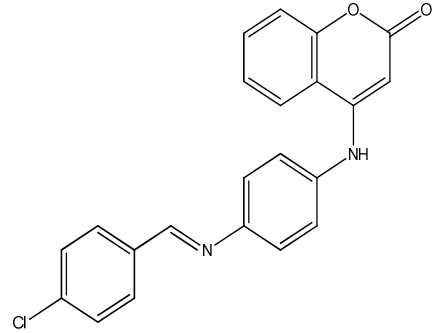
57 Data set from the literature[13] used in the Quantum Chemical QSAR analysis;

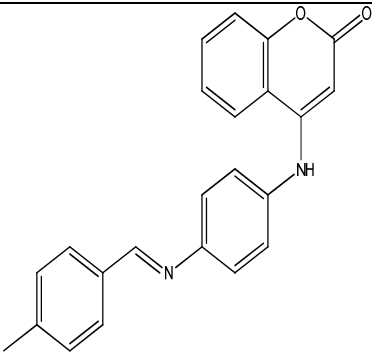
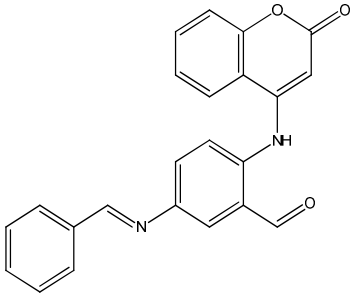
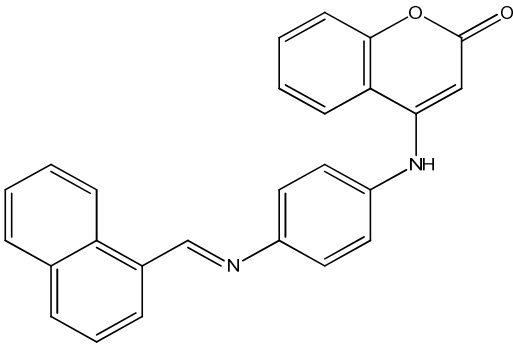
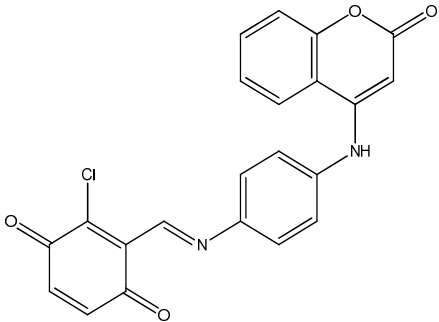
58 $pMIC = -\log MIC$ =Minimum inhibitory concentration in molar

59 Biological data on the activity of Ketone derivatives has been obtained from Jayakumar Swamy et al,

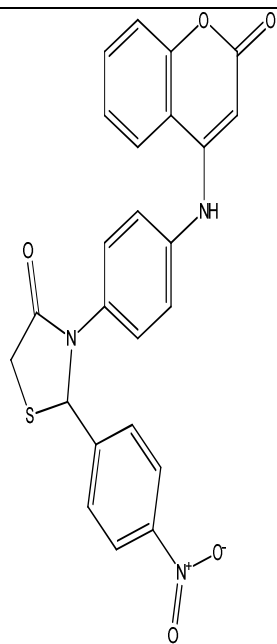
60 Divyesh Patel et al

S/NO	STRUCTURES	pMIC ($\mu\text{g/mL}$)	pMIC
1	 4-[(4-{[(E)-phenylmethylidene]amino}phenyl)amino]-2H-chromen-2-one	500	2.83
2	 4-[(4-{[(E)-4-Nitrophenylmethylidene]amino}phenyl)amino]-2H-chromen-2-one	800	2.68

3		500	2.89
4		100	3.60
5		500	2.87

6		200	3.25
7		500	2.87
8		400	2.99
9		400	3.01

10

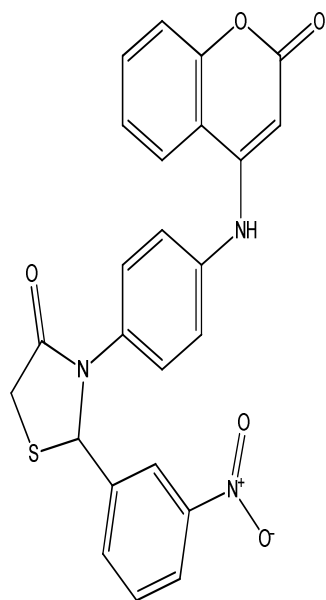


800

2.71

2-(4-Nitro-phenyl)-3-[4-(2-oxo-2H-chromen-4-ylamino)-phenyl]-thiazolidin-4-one

11

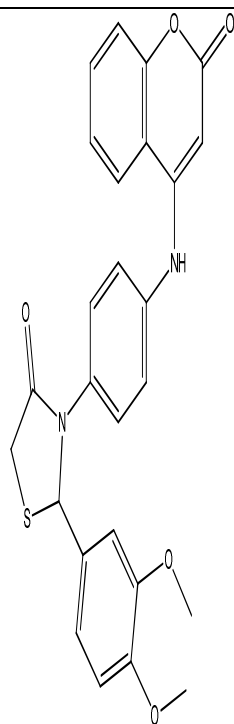


250

3.26

2-(3-Nitro-phenyl)-3-[4-(2-oxo-2H-chromen-4-ylamino)-phenyl]-thiazolidin-4-one

12

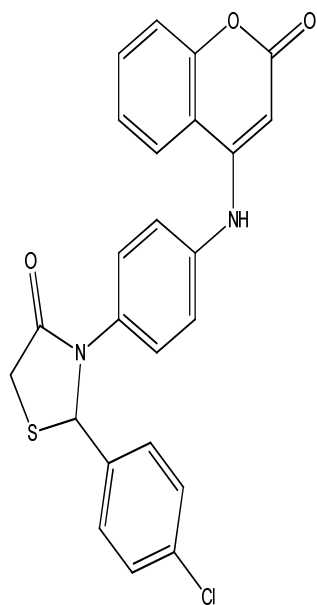


1000

2.68

2-(3,4-Dimethoxy-phenyl)-3-[4-(2-oxo-2H-chromen-4-ylamino)-phenyl]-thiazolidin-4-one

13



200

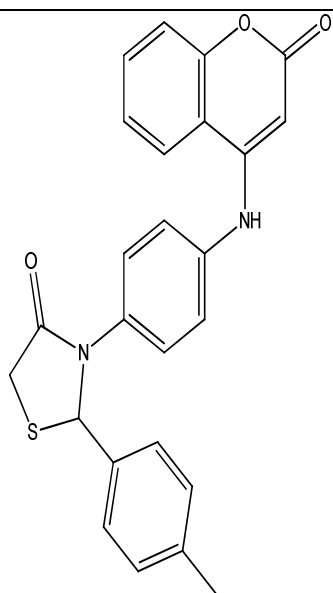
3.35

2-(4-Chloro-phenyl)-3-[4-(2-oxo-2H-chromen-4-ylamino)-phenyl]-thiazolidin-4-one

14

250

3.23

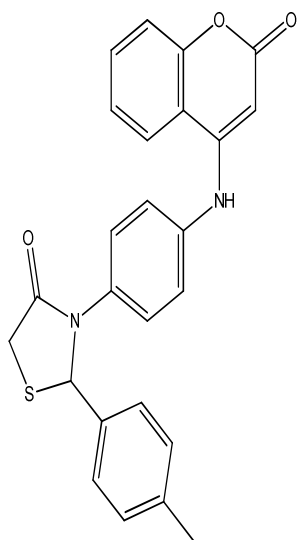


3-[4-(2-Oxo-2H-chromen-4-ylamino)-phenyl]-2-p-tolyl-thiazolidin-4-one

15

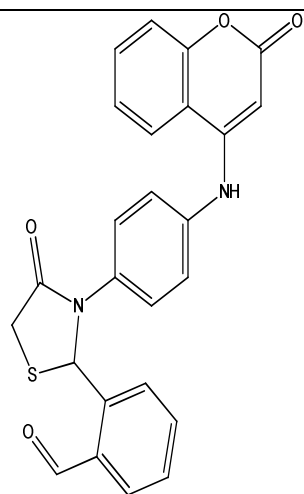
200

3.30



3-[4-(2-Oxo-2H-chromen-4-ylamino)-phenyl]-2-p-tolyl-thiazolidin-4-one

16

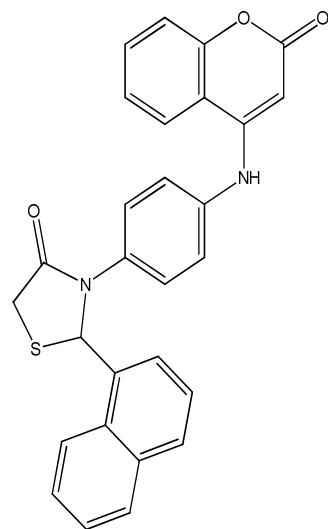


200

3.34

2-{4-oxo-3-[4-(2-oxo-2H-chromen-4-ylamino)-phenyl]-thiazolidin-2-yl}-benzaldehyde

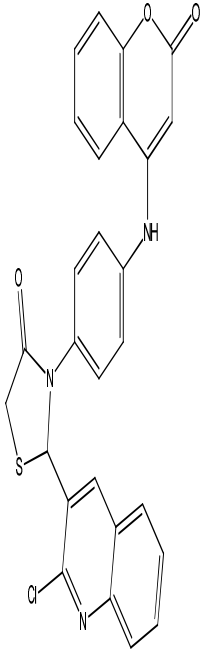
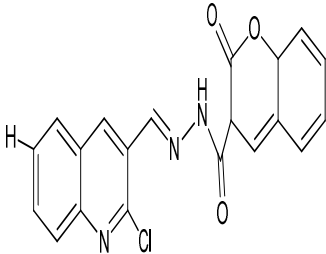
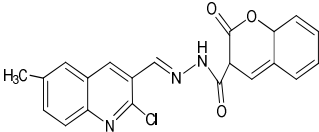
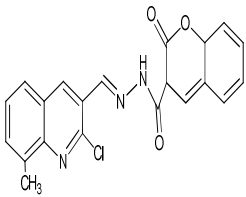
17

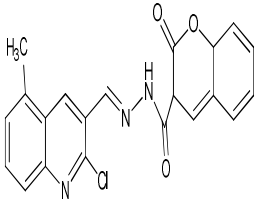
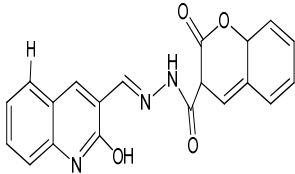
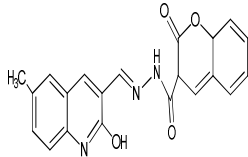
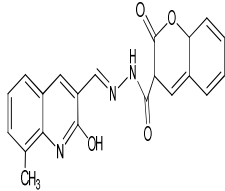
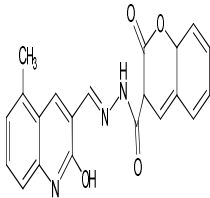
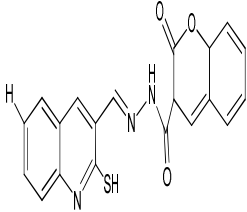


250

3.27

2-Naphthalen-1-yl-3-[4-(2-oxo-2H-chromen-4-ylamino)-phenyl]-thiazolidin-4-one

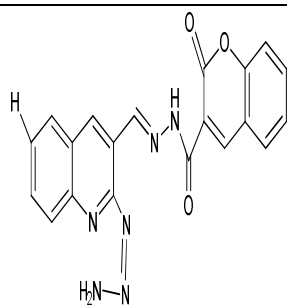
18		200	3.40
19		1000	2.58
20		1000	2.59
21		1000	2.59

22		500	2.89
23		250	3.16
24		250	3.17
25		200	3.00
26		500	2.87
27		1000	2.57

28

1000

2.58



(E)-2-oxo-N'-((2-((Z)-triaz-1-en-1-yl)quinolin-3-yl)methylene)-2H-chromene-3-carbohydrazide

61

62

Computational and statistical details

QSAR studies of Ketone derivatives was carried out on windows7, 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 first step consisted in obtaining the molecular geometry of all the derivatives from the dataset (Table-1) was energy minimization[14] and geometry optimization using Merck Molecular Force Field (MMFF) in B3LYP/6-311+G(d) density functional theoretical level was used to study QSAR[15-18].

Calculations

Molecular modeling software programs perform mathematical calculations to de-termined 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 in-teractions 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 com-plexity 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.

Density functional theory (DFT)

Density function theory is an abinitio estimation strategy for the Schrödinger mathematical statement that depends on the thickness or likelihood of electrons around a molecule as an element of the range from that molecule, rather than wave capacities as different speculations utilize. This hypothesis computes the vitality of this electron thickness by assessing the functional, or numerical capacity, of the electron thickness communicated as another mathematical capacity, implying that it returns a number in view of the electron thickness capacity. The resultant utilitarian is parceled into an active vitality term from the development of electrons, a potential vitality term from the association between the electrons and an emphatically charged core notwithstanding the connection between two cores, and in addition another potential vitality term from the unpleasant communication between an electron and different electrons in its region. The last term portrays all other electron-electron communications that may come about because of different twists of electron combines and is separated into a vitality term for same-turn and blended twist collaborations. DFT is a well known technique for estimate today 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 system [19].

107 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

108

109 The Genetic Function Approximation (GFA) calculation offers another way to deal with the
 110 issue of building quantitative structure activity relationship (QSAR) and quantitative structure-property
 111 relationship (QSPR) models. Supplanting relapse investigation with the GFA calculation empowers the
 112 development of models focused with or better than those delivered by standard procedures and makes
 113 accessible extra data not gave by different methods. Not at all like most different investigation
 114 calculations, GFA gives various models, where the populaces of the models are made by advancing
 115 irregular starting models utilizing a hereditary calculation. GFA can manufacture models utilizing
 116 straight polynomials as well as higher-request polynomials, splines, and other nonlinear capacities The
 117 Genetic Function Approximation calculations are hunt calculations that take motivation from regular
 118 hereditary qualities and development. In this segment, the thoughts hidden Genetic Function
 119 Approximation calculations are quickly depicted, underlining the angles important to the genetic
 120 function approximation (GFA) way to deal with model building. The GFA calculation itself applies
 121 these thoughts to the issue of capacity approximation[20,21], given a substantial number of potential
 122 elements impacting a reaction, including a few forces and different elements of the crude inputs, to

123 discover the subset of terms that corresponds best with the reaction. The focal thoughts of calculations
124 are straightforward. The locale to be looked is coded into one or different strings. In the GFA, these
125 strings are sets of terms – forces and splines of the crude inputs. Every string speaks to an area in the
126 inquiry space. The calculation works with an arrangement of these strings, called a populace. This
127 populace is advanced in a way that leads it toward the goal of the pursuit. This requires that a measure
128 of the wellness of every string, comparing to a model in the GFA, be accessible.

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

133 Steadiness and meeting in the same way as other iterative minimization calculations, there are issues
134 with the strength and union of the GFA calculation. A sign of the solidness of the GFA calculation can
135 be got by producing a plot demonstrating the development of variable utilization with time. Such a plot
136 demonstrates the quantity of events of every variable in the populace for every era of the development.
137 For pragmatic reasons, to decrease the measure of information that would be gathered, such a plot is
138 produced just for those variables that happen most regularly in the last populace and the information
139 are not ordinarily gathered for each era. The GFA calculation is accepted to have met when no change
140 is found in the score of the populace over a noteworthy time allotment, either that of the best model in
141 every populace or the normal of the considerable number of models in every populace. At the point
142 when this rule has been fulfilled, no further eras are figure
143

144 Advantages of GFA

145 The GFA algorithm approach has several important advantages over other techniques[20,21].

146 (1) It builds multiple models rather than a single model.

147 (2) It automatically selects which features are to be used in the models.

148 (3) It is better at discovering combinations of features that take advantage of correlations between
149 multiple features.

150 (4) It incorporates Friedman's lack-of-fit (LOF)[22,23], error measure, which estimates the most
151 appropriate number of features, resists overfitting, and allows control over the smoothness of fit.

152 (5) It can use a large variety of equation term types in construction of its models, e.g., splines, step
153 functions, high order polynomials.

154 (6) It provides, through investigation of the advancing models, extra data not accessible from standard
155 relapse examination, for example, the favored model length and helpful allotments of
156 the information set. The strategy proceeds for a client indicated number of eras, unless union happens
157 meanwhile. Joining is activated by absence of advancement in the most noteworthy and normal scores
158 of the populace.

159

160

161 3: **RESULT AND DISCUSSION**

162 QSAR addresses two in a general sense imperative inquiries in logical examination; what auxiliary and
163 electronic properties of an atom decide its movement and what can be modified to enhance this action?
164 Computational devices permit analysts to recognize chemicals with ideal physico-compound properties
165 in silico, before costly experimentation. This spares both time and cash, and permits the tossing of sub-
166 par competitors well ahead of time of coming to the research facility [24].

167 Table 3. List of descriptors used in this study.

Descriptors	Type	Significance	VIF
SHBint5	Electrotopological state	Strength for potential hydrogen bonds of path length5	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

168

169

170 **QSAR Study**

171 To examine the Observed data, the conveyance of the information must be initially explored. Most
172 relapse calculation depends on the information that is in effect typically examined, on the off chance
173 that the information are not ordinarily disseminated, we ought to think about applying as a numerical
174 change to accomplish an ordinary circulation. Observed data in Table 4, show acceptable normal
175 distribution, so no need to perform a numerical transformation. Table 4, shows a univariate analysis for

176 the actual data. Table 4 ,contains several statistical measures that describe the actual data. The most
177 important parameters in Table 4, are the skewness and kurtosis. Skewness is the third moment of the
178 distribution, which indicates the symmetry of the distribution.

179

180 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
	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

181

182

183 Developing QSAR model is a procedure that takes an arrangement of inputs and gives an arrangement
184 of yields. For instance, a vitality minimization is a model which takes a structure as data and gives an
185 enhanced structure as yield. As of right now in a normal QSAR study, estimation of descriptors
186 happens. These are models which take a solitary structure as a data and give a solitary number or
187 gathering of firmly related numbers as yields. Table 5: Shows the experimental pMIC and the predicted

188 pMIC using the GFA approach of the training set. This shows how the GFA method predicted the
189 pMIC.

190

191 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

193 Table 6. demonstrates the GFA guess investigation which gives a synopsis of the information
 194 parameters utilized for the count. Additionally, it reports whether the GFA calculation united in the
 195 predetermined number of eras. Merging is accomplished when there has been no change in the scoring
 196 capacity for various eras. It can be seen from Table 6. that the accuracy of the model, indicated by the
 197 R^2 value, is reasonably high therefore the predictive power of the model, as indicated by the adjusted
 198 R^2 and cross validated R^2 values, is also, high, even though the regression is significant according to F-
 199 test. In Table 6, the Friedman's lack-of-fit (LOF) score[22,23], which evaluates the QSAR model by
 200 considering the number of descriptors as well as the quality of fitness, is chosen: the lower the LOF, the
 201 less likely it is that GFA model will fit the data.

202 Table 6. Validation of the genetic function approximation

Equation 1:	
1	Friedman LOF 0.05023100
2	R^2 0.90643100
3	R^2_{adj} 0.88303900
4	$R^2(cv)$ 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

203

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

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

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.

Table 7. best model:

$$Y = 0.04703a + 3.017924b - 0.05263c + 0.01986d - 1.13568$$

a = SHBint5, b = WD.unify, c = wlamba2.eneg and d = wlamba3.polar

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

$$\text{VIF} = \frac{1}{1-R^2} \quad (3)$$

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 re-check is necessary[27]. The corresponding VIF values of the five descriptors are presented in

229 Table 3. As can be seen from this table, all the variables have VIF values of less than five, indicating
230 that the obtained model has statistical significance and the descriptors were found to be reasonably
231 orthogonal [27].

232 **QSAR model Validation**

233 The real usefulness of QSAR models is not just their ability to reproduce known data, verification by
234 their fitting power (R^2), but mainly is their potential for predictive application. For this reason, the
235 internal consistency of the training set was confirmed by using leave-one-out (LOO) cross-validation
236 method to ensure the robustness of the model. The high calculated Q^2_{LOO} value, 0.7321 suggests a
237 good internal validation. A second validation method was also developed on the basis of a leave-group-
238 out (LGO) internal cross-validation method. In this case, a group of compounds in. Leave- N -out (LNO)
239 crossvalidation,[28,30,31] known also as leave-many-out, is highly recommended to test the robustness
240 of a model. The training set of M samples is divided into consecutive blocks of N samples, where the
241 first N define the first block, the following N samples is the second block, and so on. This way, the
242 number of blocks is the integer of the ratio M/N if M is a multiple of N ; otherwise the left out samples
243 usually make the last block. This test is based on the same basic principles as LOO: each block is
244 excluded once, a new model is built without it, and the values of the dependent variable are predicted
245 for the block in question. LNO is performed for $N = 2, 3$, etc., and the leave- N -out crossvalidated
246 correlation coefficients Q^2_{LNO} are calculated in the same way as for LOO . LNO can be performed in
247 two modes: keeping the same number of factors for each value of N (determined by LOO for the real
248 model) or with the optimum number of factors determined by each model. LNO is sensitive to the order
249 of samples in the data set. For example, leave-two-out crossvalidation for even M means that $M/2$
250 models are obtained, but this is only a small fraction $(0.5 \cdot (M - 1) - 1)$ of all possible combinations of
251 two samples $M!/(M - 2)! = M(M - 1)$. To avoid any systematic variation of descriptors through a data
252 set or some subset what would affect LNO, the samples should be randomly ordered (in $\mathbf{X}$ and $\mathbf{y}$

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 $SDEP = 0.1978$.

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. The y-randomization test. 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 deal with judge whether the genuine model is described by 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

Q^2_{yrand} and $R^2_{yrand} > 0.4 \rightarrow$ 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

278 R^2_{yrand} plot.

279

280 **Model applicability domain and determination of outlier towards the model**

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

$$288 \quad H_{tr} (n \times n) = X_{tr} \cdot (X_{tr}^T X_{tr})^{-1} \cdot X_{tr}^T \quad (4)$$

$$289 \quad H_{tr} = X_{tr} (X_{tr}^T X_{tr})^{-1} X_{tr}^T \rightarrow \hat{y}_{tr} = H_{tr} y_{tr} \leftrightarrow \hat{y}_{tr} = X_{tr} (X_{tr}^T X_{tr})^{-1} X_{tr}^T y_{tr} \quad (5)$$

$$290 \quad H_{ext} = X_{ext} (X_{tr}^T X_{tr})^{-1} X_{tr}^T \rightarrow \hat{y}_{ext} = H_{ext} y_{ext} \leftrightarrow \hat{y}_{ext} = X_{ext} (X_{tr}^T X_{tr})^{-1} X_{tr}^T y_{ext} \quad (6)$$

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

$$297 \quad h^* = \frac{3(k+1)}{n} \quad (7)$$

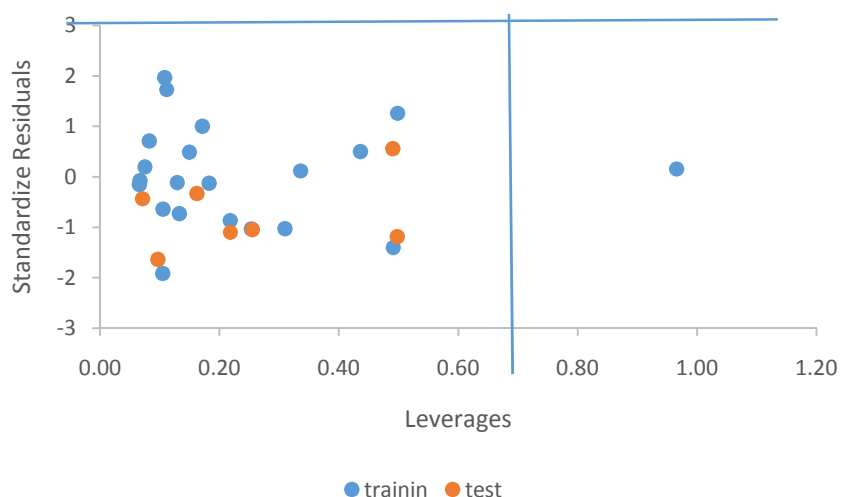
298 Where h^* the warning leverage, k is the number of descriptor in the model and n is the number of
299 observation that make up the training set. When h_{ii} value of any molecule is lower than h^* the

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

$$309 \quad SDR = \frac{\hat{y}_i - y_i}{RMSE} \quad (8)$$

310 And all $(h_{ii,te \text{ or } ext}, h_{ii,tr})$ the leverages will be extracted and a plot of the standardized residual (SDR)
311 against the leverages will be made. On that plot along the abscissa the cut-off leverage value will be
312 made the boundary. And along the ordinate ± 2.5 or at most ± 3 standard deviation unit will be used as
313 boundary because point that lie within ± 3 standardised residual from the mean covers ninety nine
314 percent of the normally distributed data [37-39]. Any compounds with cross-validated standard residual
315 larger than 2.5 standard deviation unit are considered as outlier.

316



317

318 Figure 1. The Williams plot, the plot of the standardized residuals versus the leverage
319 values.

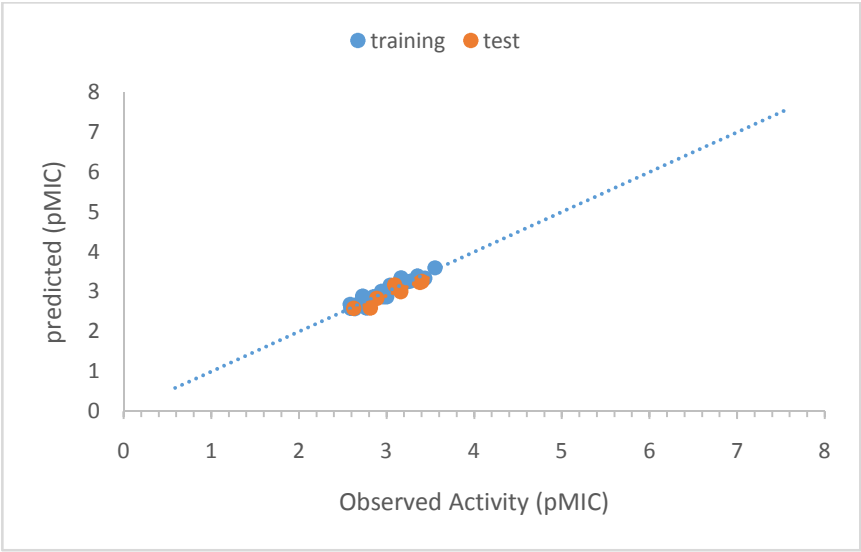
320 The Figure 1 above, shows that all the compounds of the training set and test set are inside of this
321 square area with the exception of a compound, i.e with the exception of a compound, all the chemicals
322 have a leverage lower than the warning h^* value of 0.714.

323

324 Table 8 : 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

325

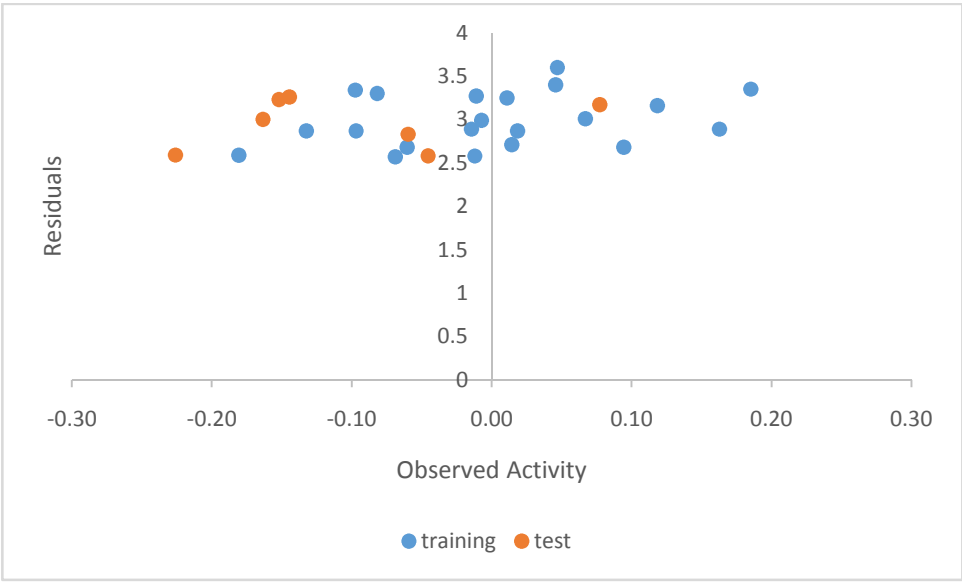


326

327 Figure 2. The predicted Activity against the Observed values (pMIC) for the training and test sets.

328 Figure 2: shows a relation between the predicted values against Observed activity for training and tests
329 set using the equation in Table 7.

330



331

332 Figure 3. The Residuals versus Observed Activity values for the training and test sets.

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

336

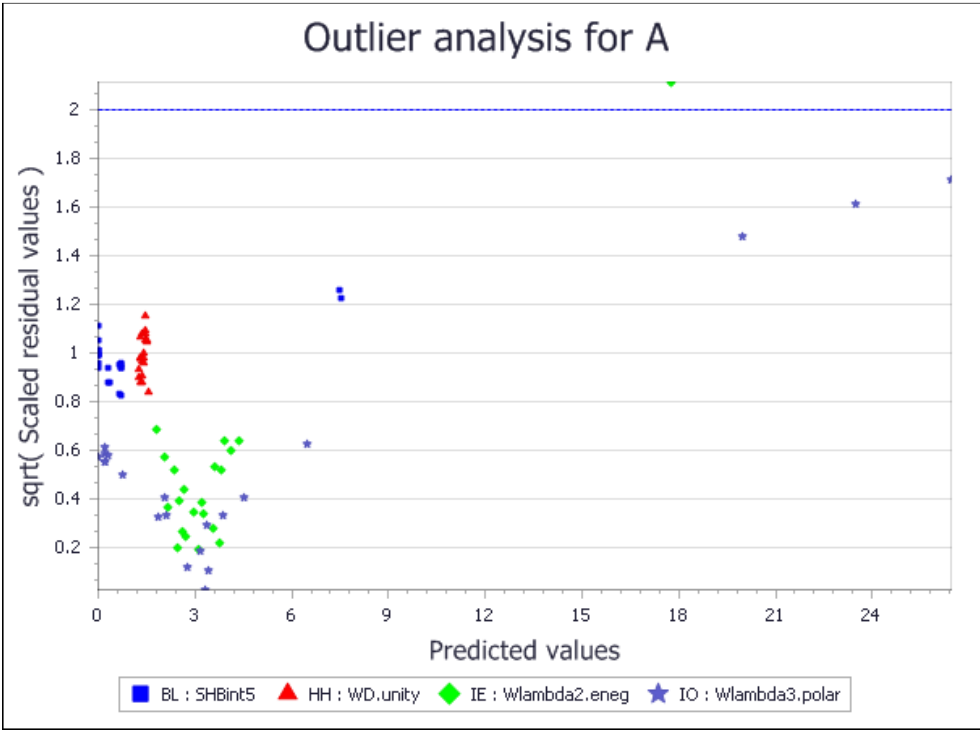
337

338

339

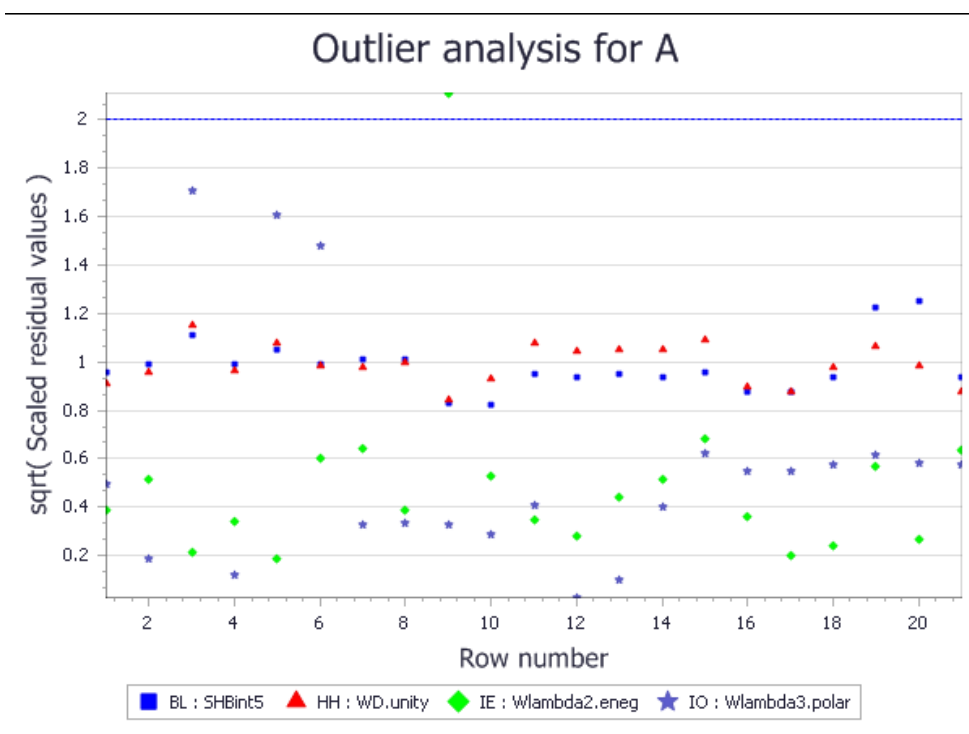
340

341



342

343



344

345 Figure 4a-b :outlier analysis

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

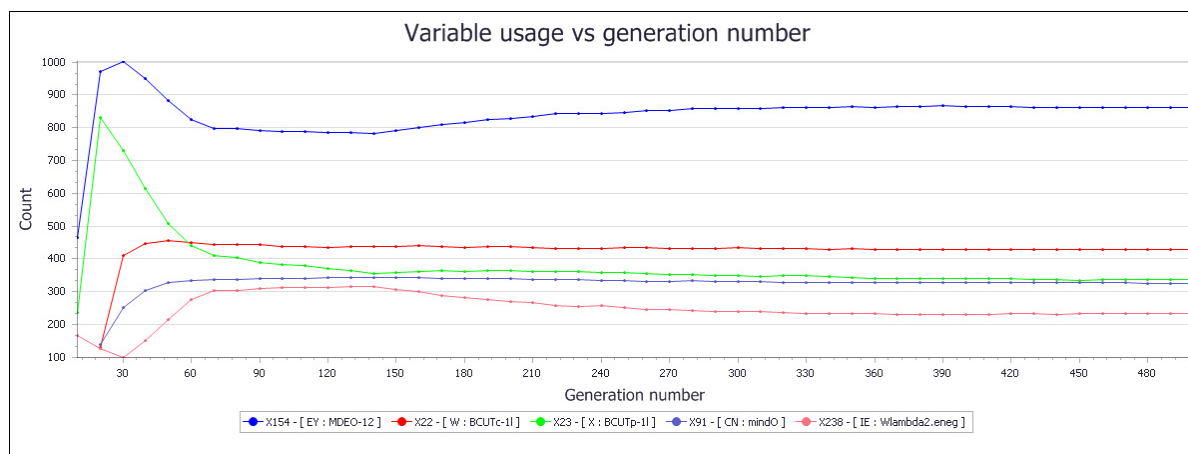


Figure 5:

The graph of the variable usage against generation number

In figure 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 figure 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

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 Ketones derivatives. These models were validated by means of leave-one-out crossvalidation, leave-*N*-out crossvalidation, external validation, y-randomization and applicability domain.

The built model was assessed comprehensively (internal and external validations), and all the validation indicate that the QSAR model we built is robust and satisfactory. Selection of four variables showed that the SHBint5, WD.Unity, wlamba2.eneg and wlamba3.polar of the molecule play a main role in the antifungal activity of the compounds.

References

- [1] A. Groll , P. Shah, C. Mentzel, M. Schneider, G. Just-Neubling, K. Huebner, Trends in the postmortem epidemiology of invasive fungal infections at a university hospital. *Journal of Infection*, **1996**, 3, 23–32.
- [2] D. Denning, E. Evans, C. Kibbler, M. D. Richardson, M .M. Roberts, T.R. Rogers, D. W. Warnock, R. E. Warren, Guideliness 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
- [3] J. D. Meyers, Fungal infections in bone marrow transplant patients. *In Seminars in oncology* , **1990**, 17, No. 3 Suppl 6, pp. 10-13
- [4] B. DiDomenico, Novel antifungal drugs. *Current opinion in microbiology*, **1999**, 2(5), 509-515.

- [5] S. Y. Ablordeppey, P. Fan, J.H. Ablordeppey, L. Mardenborough,
Systemic antifungal agents against AIDS-related opportunistic infections: current
status and emerging drugs in development. *Current medicinal chemistry*, **1999**,
6(12), 1151-1196.
- [6] T. C. White, K.A. Marr, R.A. Bowden, Clinical, cellular, and
molecular factors that contribute to antifungal drug resistance. *Clinical
microbiology reviews*, **1998**, 11(2), 382-402.
- [7] C. J. Manly, S. Louise-May, J. D. Hammer, The impact of informatics
and computational chemistry on synthesis and screening. *Drug discovery today*,
2001, 6(21), 1101-1110.
- [8] E. Pourbasheer, S. Riahi, M. R. Ganjali, P. Norouzi, 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.
- [9] E. Pourbasheer, S. Riahi, M. R. Ganjali, P. Norouzi, QSAR study of
C allosteric binding site of HCV NS5B polymerase inhibitors by support vector
machine. *Molecular diversity*, **2011**, 15(3), 645-653.
- [10] B. Delley, An all-electron numerical method for solving the local density
functional for polyatomic molecules. *The Journal of chemical physics*, **1990**, 92(1),
508-517.

- [11] A. Burger, D. J. Abraham, Burger's Medicinal Chemistry and Drug Discovery. Willey, hoboken, New jersey, **2003**
- [12] D. Rogers, Approximation with Comparison to Evolutionary Techniques. *Genetic algorithms in molecular modeling*. **1996**, 87
- [13] B. Seema, S. Ramar, M. S. Degani, Synthesis and biology evaluation of α , β -unsaturated ketone as potential antifungal agents. *Medicinal Chemistry Research*, **2009**, 18, 309-316
- [14] D. A. Williams, T. L. F. Lemke, Principles of Medicinal Chemistry (Ed). Lippincott Williams Wilkins, Baltimore. **2002**, 81
- [15] T. A. Haigren, Merck molecular force field-11. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions, *Journal of Computational Chemistry*, **1996**, 17, 520 – 552
- [16] T. A. Haigren, Merck molecular force field-11. Molecular Geometries and vibrational frequencies for MMFF94, *Journal of Computational Chemistry*, **1996**, 17, 553 – 586.
- [17] T. A. Haigren, Merck molecular force field-1V. Conformational energies and geometries for MMFF94, *Journal of Computational Chemistry*, **1996**, 17, 587 – 615
- [18] T. A. Halgren, Merck molecular force field-1. Basis, form, scope, parametrization and performance of MMFF94, *Journal of Computational Chemistry*, **1996**, 17, 490 – 519
- [19] D. C. Young, Computational Chemistry: A Practical Guide for Applying Techniques to Real-World Problems. John Wiley & Sons, Inc., 605 Third Avenue, New York, NY, **2001**

- [20] D. Rogers, 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**
- [21] D. Rogers, 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**
- [22] J. H. Friedman, Multivariate adaptive regression splines. *The annals of statistics*, **1991**, 1-67.
- [23] J. H. Friedman, Multivariate Adaptive Regression Splines, *Technical Report Laboratory* **1991**, 102
- [24] K. F. Khaled, 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] M. Jaiswal, P. V. Khadikar, A. Scozzafava, C. T. Supuran, 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] S. Shapiro, B. Guggenheim, Inhibition of oral bacteria by phenolic compounds. Part 1. QSAR analysis using molecular connectivity. *Quantitative Structure-Activity Relationships*, **1998**, 17(04), 327-337.
- [28] K. Baumann, N. J. Stiefl, *Journal of Computer-Aided Molecular Design*, **2004**, 18, 549

- [29] L. Eriksson, J. Jaworska, A. P. Worth, M. T. Cronin, R. M. McDowell, P. Gramatica, Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. *Environmental health perspectives*, **2003**, *111*(10), 1361.
- [30] R. D. Clark, R. D., P. C. Fox, Statistical variation in progressive scrambling. *Journal of computer-aided molecular design*, **2004**, *18*(7-9), 563-576.
- [31] K. Baumann, Cross-validation as the objective function for variable-selection techniques. *TrAC Trends in Analytical Chemistry*, 2003, *22*(6), 395-406.
- [32] R. F. Teófilo, J.P.A. Martins, M. M. C. Ferreira, *Journal of Chemometrics*. **2009**, *23*, 32
- [33] S. Wold, L. Eriksson, In *Chemometric Methods in Molecular Design*; H. van Waterbeemd, VCH: Weinheim, **1995**, 309
- [34] C. Rücker, G. Rücker, M. Meringer, *Journal of Chemical Information Model.* **2007**, *47*, 2345
- [35] P. Gramatica, Principle of QSAR model validation: internal and external. *QSAR Combinatorial science*, **2007**, *26*(5)694-701
- [36] A. Tropsha, A. Golbraith, 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] J.S. Jaworska, T.N. Nikola, T. Aldenberg, QSAR applicability domain estimation by projection of the training set in descriptors space: A review. *ALTA Alteran .Lab Anim*, **2005**, *33*, 445-459

- 485 [38] A. Tahereh, D. Shayesteh, H. S. Ali-Muhammad, B.G.Jahan, M. Sarkosh,
486 QSAR models for CXCR2 receptor antagonist based on genetic algorithm for data
487 processing prior to application of the PLS linear regression method and design of
488 new compound using insilico virtual screening. *Molecules*, **2011**, 16, 1928-1955
- 489 [39] N. Minovski, S. Zuperl, V. Drgan, M. Novic, Assessment of applicability domain
490 for multivariate counter-propagation artificial neural network predictive models by
491 minimum Euclidean distance space analysis: A case study. *Analytical Chimica Acta*,
492 **2013**, 759, 28– 42.