Network Modeling Analysis in Health Informatics and Bioinformatics (2019) 8:22
https://doi.org/10.1007/s13721-019-0202-8

ORIGINAL ARTICLE

QSAR modeling for the prediction of pGI_{50} activity of compounds on LOX IMVI cell line and ligand-based design of potent compounds using in silico virtual screening

Bello Abdullahi Umar[1] · Adamu Uzairu[1] · Gideon Adamu Shallangwa[1] · Uba Sani[1]

Received: 8 May 2019 / Revised: 22 August 2019 / Accepted: 28 August 2019 / Published online: 12 September 2019

Abstract
The anti-melanoma activity (pGI_{50}) values of 71 compounds from the National Cancer Institute (NCI) data bank on LOX IMVI cell line were modeled to illustrate the Quantitative structure–activity relationship (QSAR) of the compounds. The genetic function algorithm (GFA) has been used to select the most relevant descriptors so as to improve the performance of the QSAR model. The statistical significance of the model was verified based on the values of validation parameters such as R^2_{train} (0.867), R^2_{adj} (0.848), Q^2_{cv} (0.809) and R^2_{test} (0.749) needed to evaluate the robustness and strength of the model. The result of the internal and external validation of the model indicates that the model is good and could be used to predict pGI_{50} of anti-melanoma compounds on LOX IMVI cell line for which no experimental data are available. Compound 41 was selected using in-silico screening method as a template due to its good pGI_{50} (9.793) and was utilized to design new potent compounds, thereby enhancing the activity of the parent structure. Ten (10) new potent compounds were deigned and predicted using the proposed model. The predicted pGI_{50} of the majority of the designed analogous were more than the lead compound 41 used for the design and among which compound *N5* showed the best activity (pGI_{50}=13.186). Thus, this study provides a valuable approach and new direction to novel drug discovery.

Keywords LOX IMVI cell line · NCI · Anti-melanoma · GFA-MLR · QSAR · Williams plot

1 Introduction

Melanoma is one of the tumors developed from melanocytes and among the most deadly cancers among young adults (Lee et al. 2015). It has a high capability of invasion and quick metastasis to other organs which are caused by abnormalities in the cells; this may be as a result of genes (inherited) or due to exposure of the body to radiation, chemicals, or even infectious agents (Liotta et al. 1991; Mignatti and Rifkin 1993). Patients with sophisticated melanoma have a median survival time of less than 1 year, and the guessed five-year survival rate is less than fifteen percent (15%) (Anderson et al. 1995; Barth et al. 1995). With the rapid increase of melanoma in the United States (US) and other developed countries, there is an urgent need to identify more effective drugs (Gray-Schopfer et al. 2007; Lee et al. 2015). Several novel drugs were approved by the US Food and Drug Administration (FDA) such as benzylideneoxindoles, ZM336372, sorafenib, isoquinolones, triarylimidazoles, PLX4032, and XL281 for treatment of melanoma (Wu and Ambudkar 2014). Unfortunately, treatment with the use of such drugs can result in the development of drug resistance and the metastases develop again increasing about 6 months the life expectancy of the patient (Saini et al. 2013; Zubrilov et al. 2015). Therefore, identification and prediction of anti-melanoma activity of novel drugs are of great importance for cancer (Melanoma) research (Roskoski 2012).

Optimal anti-cancer drugs would exterminate cancer cells without damaging normal tissues (Al-Suwaidan et al. 2016; Choi et al. 2011; Naik and Pardasani 2018). Regrettably, currently no available drugs meet this condition, and clinical use of drugs involves a weighing of benefits against toxicity in a search of favorable therapeutic index (Chabner 1990; Makrariya and Pardasani 2019). Thus, these limitations have

✉ Bello Abdullahi Umar
abdallahbum@yahoo.com

[1] Department of Chemistry, Faculty of Physical Sciences, Ahmad Bello University, P.M.B.1045, Zaria, Kaduna State, Nigeria

made it necessary to search for novel anti-cancer drugs with diverse chemical structure as potential anti-cancer agents (Al-Suwaidan et al. 2016). Nevertheless, in the field of medicinal chemistry, activity prediction of new compounds is a primary goal for the drug design process (Vaidya et al. 2014). The chemical and molecular computing models are used in designing new drugs which helped in reducing the time and cost involved in designing more potent drugs. Among several computational methods used, quantitative structure–activity relationship (QSAR) has a remarkable role in designing a drug.

QSAR is an attempt to correlate structural descriptors of compounds quantitatively with biological activities. The molecular descriptors include parameters that account for conformational, constitutional, thermodynamic, steric effects and electronic properties of a molecule. Others like fragment constant, hydrophobicity, topology, hydrogen bond acceptor, and hydrogen bond-donor are also determined recently by computational methods (Arthur et al. 2016; Young 2004). QSAR models are mathematical equations which relate the chemical structure of compounds to their biological activity. Therefore, it is necessary to develop a model that could be used for the identification of new potent compounds and prediction of their anti-cancer activity before the synthesis. This will help to reduce the cost and time involved in drug discovery. This study was aimed to develop QSAR model based on the compounds collected from NCI data bank which can be used to predict the anti-melanoma activity of known and new potent compounds on LOX IMVI cell line. Additionally, an in silico screening technique is applied to the proposed QSAR model to predict the structure of new potent anti-melanoma compounds.

2 Materials and methods

2.1 Software and computer specifications

All the molecular modeling studies were carried out on a Dell Intel(R)Core(TM)i7-5500U CPU, 16.00 GB RAM @ 2.400 GHz 2.400 GHz processor, 64-bit Operating system, a 64× based processor on Windows 8.1 Pro. Spartan 14 (Hehre and Huang 1995) was employed to perform density functional theory calculations, Material studios 8.0 was used to develop the model and Microsoft office Excel 2013 was utilized for statistical analysis.

2.2 Data set

In this research, a data set of 71 anti-melanoma compounds and their pGI_{50} activities on LOX IMVI human melanoma cell line were collected from the drug discovery and development section of the National Cancer Institute (NCI) (https://wiki.nci.nih.gov/display/NCIDTPdata/NCI-60+Growth+Inhibition+Data). Their NSC number and anti-melanoma activity results as pGI_{50}, which is the negative log ($-LogGI_{50}$) of the concentration for 50% of cancer cell proliferation, are depicted in Table 2.

2.3 Computation of descriptors

The 2D structure of each of the compounds was converted into the 3D structure using Spartan 14. The structures were cleaned by minimizing and checking using a molecular mechanic force field (MM+) option on Spartan 14, so as to remove all strain from the structure of the molecule. Additionally, this will guarantee a well-defined and stable conformer relationship within the compounds in the study (Viswanadhan et al. 1989). Geometry optimization was set at the ground state utilizing the density functional theory (DFT) at the Becke88 three-parameter hybrid exchange potentials with Lee–Yang–Parr correlation potential (B3LYP) level of theory and for the basis set 6-311G (d) was selected. The fully optimized 3D structure in SD file was then imported into PaDEL descriptor software to compute both thermodynamic, topological, autocorrelation constitutional, electronic, and geometric descriptors (Amin and Gayen 2016) for further studies (Yap 2011).

2.4 Dataset division into modeling and prediction sets

The data set was divided into two sets, the modeling and prediction set. The modeling set is used in developing the model; it contains seventy percent (70%) of the entire data set. While the test set which constitutes the remaining thirty percent (30%) of the whole data set was not used in the construction of the model but to ascertain the predictive ability of the model (Tropsha et al. 2003). This partitioning ensures that a similar principle can be employed for the activity prediction of the test set. Kennard–Stone Algorithm was applied for dividing dataset into a modeling and test set (Kennard and Stone 1969; Rajer-Kanduč et al. 2003).

2.5 Model development

In QSAR studies, the identification and selection of descriptors which provide maximum information in activity variations and have minimum co-linearity are important. Therefore, a genetic function algorithm (GFA) (Leardi 1996) improves the model accuracy in the selection of proper descriptors. Multiple Linear Regression (MLR) was used on the modeling set to show the relationship between the dependent variable Y (pGI_{50}) and independent variable X (molecular descriptors). In regression analysis, the contingent mean of the dependent variable (pGI_{50}) Y relies on (descriptors) X.

2.6 QSAR model validation

In the validation of a QSAR model, the stability and predictive ability of the model is one of the key steps in QSAR modeling. Various statistical parameters have been utilized for the validation of the suitability of the built model for the prediction of the anti-cancer activity of the studied compounds (Asadollahi et al. 2011) this includes correlation coefficient (R^2) which describes the fraction of the total variation attributed to the model. The closer the value of R^2 is to 1.0, the better the regression and equation explain the Y variable. R^2 is the most commonly used internal validation indicator and is expressed as in Eq. (1):

$$R^2 = 1 - \frac{\sum \left(Y_{\text{exp}} - Y_{\text{pred}}\right)^2}{\sum \left(YY_{\text{exp}} - Y_{\text{mtraining}}\right)^2}, \tag{1}$$

where Y_{exp}, Y_{pred}, and $Y_{\text{mtraining}}$ are the experimental property, the predicted property and the mean experimental activity of the compounds in the training set, respectively. The minimum recommended value for this parameter is shown in Table 1 (Wu et al. 2015).

Adjusted R^2 (R^2_{adj}): R^2 value varies directly with the increase in the number of descriptors; thus, R^2 cannot be a useful measure for the goodness of model fit. Therefore, R^2 is adjusted for the number of explanatory variables in the model. The adjusted R^2 is defined as in Eq. (2):

$$R^2_{\text{adj}} = 1 - \left(1 - R^2\right)\frac{N-1}{N-P-1} = \frac{(N-1)R^2 - P}{N-P+1}, \tag{2}$$

where P = number of independent variables in the model and N = sample size (Abdulfatai et al. 2017). The minimum recommended value for this parameter is presented in Table 1.

Cross-validation coefficient parameter (Q^2_{CV}) is the most commonly used internal validation indicator and is expressed as in Eq. (3):

$$Q^2_{CV} = 1 - \frac{\sum (Y_{\text{pred}} - Y_{\text{exp}})^2}{\sum (Y_{\text{exp}} - Y_{\text{mtraining}})^2}, \tag{3}$$

Table 1 Minimum recommended values of validated parameters for generally acceptable QSAR

Symbol	Name	Value
R^2	Coefficient of determination	≥ 0.6
Q^2_{cv}	Cross-validation coefficient	< 0.5
R^2_{test}	The coefficient of determination for external test set	≥ 0.6
$R^2 - Q^2$	Difference between R^2 and Q^2	≤ 0.3
N_{test}	Minimum number of an external test set	≥ 5

where Y_{exp} is the experimental activity, Y_{pred} is the predicted activity, and $Y_{\text{mntraining}}$ is the mean of the experimental activity of the validation set (Tropsha et al. 2003). However, it should be noted that a high Q^2_{CV} does not necessarily mean high predictability of the built model (Asadollahi et al. 2011). In other words, the high value of Q^2_{CV} is a necessary condition, but not sufficient for a developed model to have high predictability.

To assess the predictive ability and to check the statistical significance of the developed model, the proposed model was applied to predict the pGI50 values of an external (test) set compounds that were not used in building the model. The predictive powers of the proposed regression model on the training set were evaluated by predicted values of the prediction (test) set. Therefore, validation through an external prediction set (R^2_{test}) is a very important parameter that is used to test the external predictive ability of a QSAR model. The R^2_{test} value is calculated by Eq. (4):

$$R^2_{\text{test}} = 1 - \frac{\sum (Y_{\text{pred}} - Y_{\text{exp}})^2}{\sum (Y_{\text{exp}} - Y_{\text{mntraining}})^2}, \tag{4}$$

where Y_{exp} is the experimental activity, Y_{pred} is the predicted activity, and $Y_{\text{mntraining}}$ is the mean of the experimental activity of the training set (Tropsha et al. 2003).

2.7 Y-randomization test

To assess the robustness of the built model, the Y-randomization test was applied to the training set data as suggested by Tropsha et al. (2003). The dependent variable vector (activity data) was randomly shuffled and a new QSAR model was developed using the original independent variable matrix. For the built QSAR model to be robust and reliable, the model is expected to have low R^2 and Q^2 values for several trials. The coefficient of determination cR^2_p for Y-randomization is another parameter calculated which should be greater than 0.5 for passing this test as in Eq. (5):

$$cR^2_p = R \times \left[R^2 - R^2_r\right]^2 \tag{5}$$

cR^2_p is coefficient of determination for Y-randomization, R is the coefficient of determination for Y-randomization and R_r is average 'R' of random models.

3 Results and discussion

On the basis of Kennard–Stones algorithm, 49 compounds out of 71 were selected as the modeling (training) set and the remaining 22 were selected as the prediction (test) set. GFA regression was used on the modeling data set to select

the significant descriptors and it was found that among 1875 calculated descriptors, the SpMin-Bhv, SpMax4-Bhe, SpMin5-Bhi, SpMin3-Bhs, piPC1, and GGI4 build the best model and a new GFA-MLR QSAR regression equation was developed based on modeling set.

3.1 QSAR model for predicting pGI_{50} on LOX IMVI cell line

$$\begin{aligned} pGI_{50} &= 3.938350117\left(SpMin6_{Bhv}\right) \\ &- 3.500212746\left(SpMax4_{Bhe}\right) \\ &- 2.734990552\left(SpMin5_{Bhi}\right) \\ &- 2.424058833\left(SpMin3_{Bhs}\right) \\ &+ 7.110589756\,(piPC1) \\ &- 0.421160719\,(GGI4) - 4.34594 \end{aligned}$$

$$N_{train} = 49,\ R^2_{train} = 0.867,\ R^2_{adjusted} = 0.848,$$
$$Q^2_{cv} = 0.809,\ N_{test} = 22 \text{ and } R^2_{test} = 0.749,$$

where N is the number of compounds in the training and test sets, R^2_{train} is the squared correlation coefficient, $R^2_{adjusted}$ is the adjusted R-squared, Q^2_{cv} is the cross-validation coefficients of the training set and R^2_{test} is the squared correlation coefficient of the prediction (test) set.

3.2 QSAR model validation

In a further study, the built QSAR model from the modeling data set was used to evaluate its predictive ability by predicting the pGI_{50} values in the prediction set (test set). The results are given in Table 2. The predicted pGI_{50} values for the training and test sets were plotted against the experimental pGI_{50} as shown in Fig. 1. The predicted pGI_{50} results obtained for both the modeling set and prediction set (Table 2) are in good agreement with the experimental pGI_{50} obtained from NCI. The residual values obtained between predicted and experimental pGI_{50} were very low.

The result of the QSAR model is in conformity with the standard shown in Table 1 as seen from the built model. The closeness of coefficient of determination (R^2) to its absolute value of 1.0 is an indication that the model explained a very high percentage of the response variable (descriptor) variation, high enough for a robust QSAR model. Its 0.867 value illustrates that 86.7% of the variation is residing in the residual meaning that the model is very good.

The high adjusted R^2 (R^2_{adj}) value as seen in the model and its closeness in value to the value of R^2 imply that the model has excellent explanatory power to the descriptors in it. It also demonstrates the real influence of applied descriptors on the pGI_{50}. Also, the high and closeness of Q^2_{cv} to R^2_{train} revealed that the model was not over-fitted. The high R^2_{test} as seen in the model is an indication that the model is capable of providing valid predictions for new compounds.

Additionally, to assess the robustness of the model, the Y-randomization test was applied. The dependent variable vector (inhibitory activity) was randomly shuffled and a new QSAR model was developed using the original independent variable matrix. As was expected, the new QSAR models (after several repetitions) have low R^2 and Q^2 values and also, the cR^2_p value was greater than 0.5 as presented in Table 3. This test affirms that the proposed model is powerful and not inferred by chance.

3.3 Contribution and interpretation of descriptors in model

The six-variable QSAR model adequately represents the pGI_{50} data, based on direct statistics as well as validation methods. Each of the variables is a descriptor of an aspect of molecular structure and will be discussed to indicate the specific structural information encoded. By interpreting the descriptors contained in the QSAR model, it is possible to gain some insights into factors, which are related to the anti-cancer activity. For this reason, an acceptable interpretation of the selected descriptors is provided. The brief descriptions of the descriptors are shown in Table 4. The relative importance and contribution of each descriptor in the model were determined by the calculation of the value of the mean effect (MF) (Jalali-Heravi and Konuze 2002) for each descriptor using Eq. (6) and the MF values are presented in Table 4:

$$MF_j = \frac{\beta_j \sum_{i=1}^{i=n} d_{ij}}{\sum_j^m \beta_j \sum_i^n d_{ij}}, \tag{6}$$

where MF_j represents the mean effect for the descriptor j, β_j is the coefficient of the descriptor j, $d_i d$ is the value of the interested descriptors for each molecule and m is the number of descriptors in the model.

The MF value shows the relative importance of each descriptor compared to the other descriptors. The MF of the descriptors SpMin-Bhv, SpMax4-Bhe, SpMin5-Bhi, SpMin3-Bhs, piPC1 and GGI4 are also shown in Table 4 and indicate that among the selected descriptors, the most important one is piPC1 (Conventional bond order ID number of order 1 ($\ln(1+x)$) as it has the highest mean effect value and has the largest effect on the pGI_{50} of the compound. On the basis of MF values, the associated descriptors are arranged in a sequence pertaining to their contribution towards overall pGI_{50} of the compounds, in the following increasing order of pGI_{50} of compounds.

Table 2 NSC numbers, chemical names, experimental and predicted pGI_{50} of the dataset with residuals

S/N	NSC	Name	Experimental pGI_{50}	Predicted pGI_{50}	Residuals
1t[a]	267,469	Deoxydoxorubicin	7.531	7.147	0.384
2	269,148	MENOGARIL	6.293	7.251	−0.958
3	268,242	*N*,*N*-Dibenzyldaunorubicin hydrochloride	8.000	8.176	−0.176
4	126,771	Dichloroallyl lawsone	5.572	5.615	−0.043
5	136,044	RHODOMYCIN A	7.681	6.988	0.693
6	140,377	Arnebin 1	6.583	6.045	0.538
7	196,524	epsilon-Rhodomycinone	5.626	6.620	−0.994
8	212,509	4beta-Hydroxywithanolide	6.876	6.759	0.117
9t	215,139	Bikaverin	6.272	7.975	−1.703
10	236,613	Plumbagin	5.742	5.666	0.076
11	252,844	SHIKALKIN	5.915	6.018	−0.103
12t	257,450	Dermocybin	4.618	5.359	−0.741
13	143,095	Pyrozofurin	6.298	6.091	0.207
14	629,971	9-Aminocamptothecin (R,S)	8.000	7.497	0.503
15t	606,173	11-Hydroxymethyl-20(RS)-camptothecin Camptothecin,*N*-diethyl	5.738	7.535	−1.797
16	364,830	Glycinate	7.934	7.413	0.521
17	94,600	Camptothecin	7.596	6.865	0.731
18t	606,985	Campothecin analog	8.050	7.92795	0.122
19	606,499	Camptothecin butylglycinate ester hydrochloride	7.142	7.613	−0.471
20	606,497	Camptothecinethylglycinate esterhydrochloride	7.049	7.675	−0.626
21	176,323	9-Methoxycamptothecin	8.353	7.939	0.414
22t	3088	Chlorambucil	5.113	4.306	0.807
23	338,947	Clomesone	4.379	4.726	−0.347
24	95,678	Picolinaldehyde	5.276	4.542	0.734
25	264,880	Dihydro-5-azacytidine	5.646	5.493	0.153
26	163,501	Acivicin	5.484	5.525	−0.041
27t	71,851	alpha-Thiodeoxyguanosine	4.71	5.826	−1.116
28t	132,483	L-Aspartic acid	7.927	8.041	−0.114
29	308,847	Amonafide	5.604	5.889	−0.285
30t	355,644	Anthra[1,9-cd]pyrazol-6(2H)-one der	9.000	7.865	1.135
31t	63,878	Cytosine, monohydrochloride	7.145	5.69	1.455
32	182,986	Diaziquone	5.614	6.405	−0.791
33t	139,105	Triazinate	7.311	7.597	−0.286
34	409,962	Carmustine	4.428	4.043	0.385
35	337,766	Bisantrene hydrchloride	8.000	8.633	−0.633
36	750	Busulfan	3.650	4.076	−0.426
37t	95,382	Camptothecin, acetate	5.995	8.009	−2.014
38t	107,124	10-Hydroxycamptothecin	7.603	7.329	0.274
39	79,037	Lomustine	4.848	4.159	0.688
40	132,313	Dianhydrodulcitol	4.670	4.705	−0.035
41	376,128	AC1L2OAS	9.793	8.987	0.806
42	73,754	Fluorodopan	3.690	3.402	0.288
43	148,958	Uracil	3.185	3.521	−0.336
44[a]	1895	Guanazole	2.449	3.203	−0.754
45	329,680	Hepsulfam	3.793	3.638	0.155
46	142,982	Hycanthone mesylate	5.427	6.638	−1.211
47[a]	32,065	Hydroxyurea	3.205	3.032	0.173
48	153,353	Alanosine monosodium salt	6.546	6.008	0.538
49	249,992	Amsacrine	6.809	6.813	−0.004

Table 2 (continued)

S/N	NSC	Name	Experimental pGI_{50}	Predicted pGI_{50}	Residuals
50t	740	Methotrexate	7.573	7.826	−0.253
51t[a]	95,441	Semustine	4.834	3.809	1.025
52	26,980	Mitomycin C	6.489	6.559	−0.07
53	353,451	Mitozolomide	4.663	5.887	−1.224
54t[a]	268,242	*N*,*N*-Dibenzyldaunorubicin hydrochloride	6.312	8.176	−1.864
55t[a]	95,466	Urea	4.327	5.561	−1.234
56	25,154	Pipobroman	4.312	3.759	0.553
57t	56,410	Profiromycin	5.94	6.5751	−0.635
58t	366,140	Pyrazoloacridine mesylate	6.769	6.975	−0.206
59	51,143	Pyrazoloimidazole	3.000	2.576	0.424
60	172,112	Spiromustine	4.024	5.496	−1.472
61	125,973	Paclitaxel	7.992	7.781	0.211
62	296,934	Teroxirone	4.885	5.229	−0.344
63t	363,812	5-((4-Chlorobenzyl)thio)-3-(trifluoromethyl)-1H-1,2,4-triazole	6.063	6.016	0.047
64	361,792	3-Demethylthiocolchicine	8.000	6.881	1.119
65	6396	Thiotepa	5.146	5.056	0.09
66	9706	Triethylenemelamine	5.507	5.474	0.033
67t[a]	83,265	Tritylcysteine	6.066	5.885	0.181
68	49,842	Vinblastine sulfate	9.154	8.078	1.076
69	67,574	Vincristine sulfate	6.955	8.164	−1.209
70	757	Colchicine	8.268	7.362	0.906
71	33,410	*N*-Benzoyl-deacetylcolchicine	7.849	7.474	0.375

't' represents test sets

[a]Identified compounds found outside the applicability domain of the QSAR model

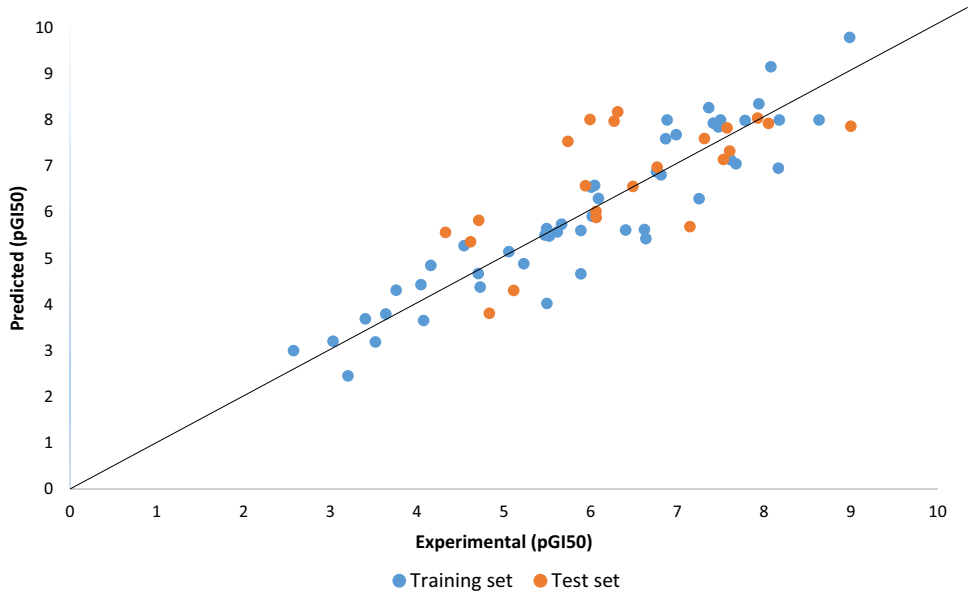

Fig. 1 The predicted pGI_{50} against the experimental values for the training and test sets

Table 3 R^2 and Q^2 values after several Y-randomization tests

Model	R	R^2	Q^2
Original	0.879700251	0.77387253	0.561628025
Random 1	0.482831891	0.23312663	0.052213103
Random 2	0.25891164	0.06703524	−0.30980934
Random 3	0.568132869	0.32277496	0.129408597
Random 4	0.409536891	0.16772047	−0.07251305
Random 5	0.362492706	0.13140096	−0.12337785
Random 6	0.436926035	0.19090436	−0.00890762
Random 7	0.336000891	0.1128966	−0.28251916
Random 8	0.465496064	0.21668659	−0.02450686
Random 9	0.34931455	0.12202066	−0.06406767
Random 10	0.26429312	0.06985085	−0.20286786
Random models parameters			
Average r	0.393393666		
Average r^2	0.163441731		
Average Q^2	−0.090694771		
cR_p^2	0.69218154		

piPC1 > SpMin6_Bhv > GGI4 > SpMin5_Bhi > SpMin3_Bhs > SpMax4_Bhe

The SpMin6_Bhv descriptors have been proposed as the chemical structure descriptors derived from a new representation of the molecular structure. SpMin6_Bhv is the smallest absolute eigenvalue of Burden modified matrix-*n* 6/weighted by relative van der Waals volumes. The SpMin6_Bhv mean effect has a positive sign as presented in Table 4. This sign suggests that the anti-melanoma activity is directly related to this descriptor.

SpMax4_Bhe is defined as the largest absolute eigenvalue of Burden modified matrix-*n* 4/weighted by relative Sanderson electro-negativities. The SpMax4_Bhe mean effect has a negative sign as shown in Table 4. This sign suggests that the decrease of value for this descriptor will increase the anti-cancer activity of a molecule and vice versa. SpMin5_Bhi is the smallest absolute eigenvalue of Burden modified matrix-*n* 5/weighted by relative first ionization potential. The negative sign of the mean effect (Table 4) of SpMin5_Bhi suggests that its decrease may increase the anti-cancer activity. SpMin3_Bhs is the smallest absolute eigenvalue of Burden modified matrix-*n* 3/weighted by relative I-state. SpMin3_Bhs also has a negative mean effect value which suggests that the decrease of value for this descriptor will increase the anti-melanoma activity of a molecule. The SpMin5_Bhi has a negative mean effect (Table 4) and its decrease may improve the anti-melanoma activity.

piPC1 is a 2D descriptor defined as the conventional bond order ID number of order 1 ($\ln(1+x)$; it is also describe as the molecular multiple path counts of order 01; and the mean effect of piPC1 was found to positively influence the anti-melanoma activity of the compounds when increased as shown in Table 4. GGI4 is defined as topological charge index of order 4. The mean effect value for this descriptor has a negative sign (Table 4). This sign suggests that the anti-melanoma activity will increase with the decrease in its value. The descriptors used for building the QSAR model in this work encoded topological, electronic and geometrical aspects of molecules. Appearances of these descriptors in the model reveal the role of electronic and steric interactions in inducing anti-melanoma pGI_{50} activity on LOX IMVI cell line.

3.4 In silico screening

An in silico screening method is a very powerful tool used for identifying new biologically potent compounds with improved characteristics and predicting their activities before their actual synthesis (Muegge and Oloff 2006; Tropsha et al. 2003). Therefore, the in silico technique reduces the time and cost involved in identifying potent compounds. Virtual screening was performed by deletion, insertion, and substitution of different substitutes on the original template (molecule) (Melagraki et al. 2007, 2009) and the effects of the structural alterations on the biological activity were

Table 4 Specification of entered descriptors and their mean effect

Descriptors	Definition	Descriptor type	MF
SpMin6_Bhv	Smallest absolute eigenvalue of Burden f-*n* 6/weighted by relative van der Waals volumes	2D	0.427624
SpMax4_Bhe	Largest absolute eigenvalue of Burden modified matrix-*n* 4/weighted by relative Sanderson electronegativities	2D	−1.08867
SpMin5_Bhi	Smallest absolute eigenvalue of Burden modified matrix-*n* 5/weighted by relative first ionization potential	2D	−0.25575
SpMin3_Bhs	Smallest absolute eigenvalue of Burden modified matrix-*n* 3/weighted by relative I-state	2D	−0.31453
piPC1	Conventional bond order ID number of order 1 (ln(1 + x)	2D	2.322979
GGI4	Topological charge index of order 4	2D	−0.09165

evaluated. Then, the applicability domain (AD) of the QSAR model was defined to use the model for predicting and screening new leads. Defining the domain of application of the QSAR model is essential in establishing the model capability to make predictions within the space (chemical) for which it was developed (Tropsha et al. 2003).

Various methods have been utilized to define the AD of the QSAR models (Eriksson et al. 2003). The most usual one was described by Gramatica et al. (2007) which used the leverage values for each compound. The leverage approach allows the determination of the position of new chemical in the QSAR model (Gramatica et al. 2007), In this regard, Leverage approach is used and is represented as h_i in Eq. (7):

$$h_i = x_i(X^TX)^{-1}x_i^T, \tag{7}$$

where x refers to the descriptor vector of the considered compound and X represents the descriptor matrix derived from the training set descriptor values. The warning leverage (h^*) was determined as in Eq. (8):

$$h^* = \frac{3(p+1)}{N}, \tag{8}$$

where N is the number of training compounds and p is the number of descriptors in the model.

The defined applicability domain (AD) was then viewed via a Williams plot, the plot of the standardized residuals against the leverage values (h). A compound with $h_i > h^*$ seriously influences the model performance and may be eliminated from the AD applicability, but it does not appear to be an outlier since its standardized residual could be small. Furthermore, a value range of ±3 standardized residuals is often used as a cutoff value for accepting predictions of a molecule, because points which lie within ±3 standardized residual from the mean cover ninety-nine percent (99%) of the normally distributed data (Jaworska et al. 2005). Thus, the leverage and the standardized residuals were used jointly for the characterization and determination of the applicability domain. The Williams plot for the built QSAR is shown in Fig. 2. The warning leverage (h^*) was found to be 0.430 for the developed QSAR model. Based on the leverages ($h_i > 0.430$), the two compounds among the training set (44 and 47) and five test set compounds (1, 51, 54, 55 and 67) were found to be outside of the defined AD (Fig. 2) of the QSAR model; so, they were identified as structurally influential chemical based on their large leverage values ($h_i > h^*$).

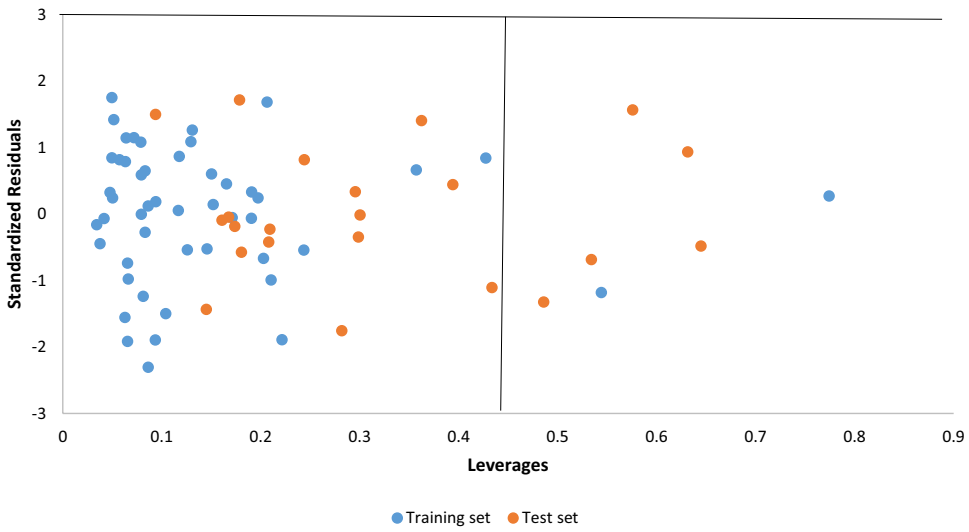

Fig. 2 The Williams plot, the plot of the standardized residuals versus the leverage value

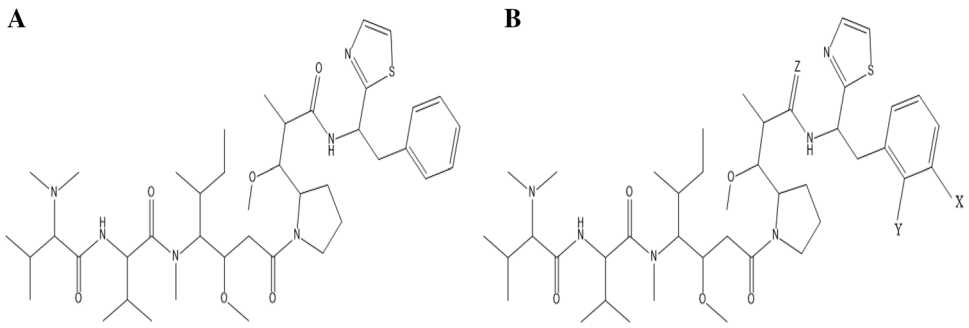

Fig. 3 **a** Structure of compound 41. **b** Structure of the template used for design

Table 5 Structural modification of compound 41 and predicted pGI_{50} with leverage limit

ID	*X*	*Y*	*Z*	Predicted pGI_{50}	Leverage-limit
N1	H	NO_2	NH	9.8360	0.329
N2	H	NO_2	NMe	12.876	0.323
N3	H	NO_2	NOH	10.901	0.412
N4	H	NO_2	S	9.2630	0.218
N5	Br	NO_2	NH	13.186	0.229
N6	Br	NO_2	NMe	9.3420	0.513
N7	Br	NO_2	NOH	10.159	0.387
N8	Br	NO_2	S	9.2950	0.488
N9	Br	OMe	NH	10.845	0.406
N10	Br	OMe	NMe	9.2530	0.585
N11	Br	OMe	NOH	11.842	0.314
N12	Br	OMe	S	9.1192	0.333

Furthermore, the in silico screening method was used for the design of new potent structures with pGI_{50} activity on LOX IMVI cell line according to the developed QSAR model. For this purpose, compound 41 (AC1L2OAS, NSC-376,128) listed in Table 2 with pGI_{50} of 9.793 was chosen as a template due to its high pGI_{50} activity, low residual value and was found to be within the defined AD (Fig. 2). The structure of compound 41 and the template used for modifications are shown in Fig. 3. The compound was altered in a way that will make its synthesis experimentally possible. Then, the in silico screening was applied by the insertion and substitution of different groups in the *X*, *Y* and *Z* positions as presented in Fig. 3; the results of this are presented in Table 5. The model endures various AC1L2OAS substituents since the majority of the newly designed analogous was within the applicability domain. The predicted pGI_{50} of the majority of the designed analogous were more than the lead compound (41) used for the design and among which compound N5 showed the best activity ($pGI_{50} = 13.186$). Thus, it is clear that using a simple QSAR model, there is a possibility to simultaneously predict and identify compounds with better activity and to determine which of the structural modifications do not fall within the AD. Lastly, the result in this research confirms the robustness and reliability of the developed QSAR model and it illustrates that with the modeling of the QSAR model and use of an in silico screening technique, it is possible to identify new potent synthetic targets for drug development.

4 Conclusions

In this research, GFA-MLR modeling tool was used in the construction of a QSAR model for predicting pGI_{50} of anti-melanoma compounds on LOX IMVI cell line. The accuracy and predictability of the proposed model was illustrated by various criteria, the model is statistically fit both internally ($R^2_{train} = 0.867$, $R^2_{adj} = 0.848$ and $Q^2_{cv} = 0.809$), externally ($R^2_{test} = 0.749$), and Y-randomization. This satisfies the criteria of acceptable QSAR model proposed by different groups. Moreover, in silico screening method was applied to the developed QSAR model which enables the design and prediction of pGI_{50} of new potentially active compounds on LOX IMVI cell line. The predicted pGI_{50} of the majority of the designed analogous were more than the lead compound 41 used for the design. The proposed model was found to be useful for the prediction of pGI_{50} of anti-melanoma compounds for which no experimental data are available and it also helps in the reduction of time and cost involved in the synthesis and anti-melanoma activity prediction of compounds on LOX IMVI cell line.

Acknowledgements The authors sincerely acknowledge Ahmadu Bello University, Zaria for providing the softwares used and all the members of the group for their advice and encouragement in the cause of this research.

Funding The authors received no direct funding for this research.

Compliance with ethical standards

Conflict of interest The authors have declared they have no conflict of interest.

Human and animal rights statement This article does not contain any studies with human or animal subjects.

References

Abdulfatai U, Uzairu A, Uba S (2017) Quantitative structure-activity relationship and molecular docking studies of a series of quinazolinonyl analogues as inhibitors of gamma amino butyric acid aminotransferase. J Adv Res 8(1):33–43

Al-Suwaidan IA, Abdel-Aziz AA-M, Shawer TZ, Ayyad RR, Alanazi AM, El-Morsy AM, Mohamed MA, Abdel-Aziz NI, El-Sayed MA-A, El-Azab AS (2016) Synthesis, antitumor activity and molecular docking study of some novel 3-benzyl-4 (3H) quinazolinone analogues. J Enzyme Inhib Med Chem 31(1):78–89

Amin SA, Gayen S (2016) Modelling the cytotoxic activity of pyrazolo-triazole hybrids using descriptors calculated from the open source tool "PaDEL-descriptor". J Taibah Univ Sci 10(6):896–905

Anderson CM, Buzaid AC, Legha SS (1995) Systemic treatments for advanced cutaneous melanoma. Oncology 9:4–5

Arthur DE, Uzairu A, Mamza P, Abechi S (2016) Quantitative structure–activity relationship study on potent anticancer compounds against MOLT-4 and P388 leukemia cell lines. J Adv Res 7(5):823–837

Asadollahi T, Dadfarnia S, Shabani AMH, Ghasemi JB, Sarkhosh M (2011) QSAR models for CXCR6 receptor antagonists based on the genetic algorithm for data preprocessing prior to application of the PLS linear regression method and design of the new compounds using in silico virtual screening. Molecules 16(3):1928–1955

Barth A, Wanek LA, Morton DL (1995) Prognostic factors in 1521 melanoma patients with distant metastases. J Am Coll Surg 181(3):193–201

Chabner BA (1990) Cancer chemotherapy: principles and practice. Lippincott Williams and Wilkins, United States, pp 341–355

Choi W-K, El-Gamal MI, Choi HS, Baek D, Oh C-H (2011) New diarylureas and diarylamides containing 1, 3, 4-triarylpyrazole scaffold: synthesis, antiproliferative evaluation against melanoma cell lines, ERK kinase inhibition, and molecular docking studies. Eur J Med Chem 46(12):5754–5762

Eriksson L, Jaworska J, Worth AP, Cronin MT, McDowell RM, Gramatica P (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. Environ Health Perspect 111(10):1361–1375

Gramatica P, Giani E, Papa E (2007) Statistical external validation and consensus modeling: a QSPR case study for Koc prediction. J Mol Graph Model 25(6):755–766

Gray-Schopfer V, Wellbrock C, Marais R (2007) Melanoma biology and new targeted therapy. Nature 445(7130):851

Hehre W, Huang W (1995) Chemistry with computation: an introduction to SPARTAN. Wavefunction. Inc, Irvine, CA

Jalali-Heravi M, Konuze E (2002) Use of quantitative structure property relationships in predicting the kraft point of anionic surfactants. Electron J Mol Des 1:410–417

Jaworska J, Nikolova-Jeliazkova N, Aldenberg T (2005) QSAR applicability domain estimation by projection of the training set descriptor space: a review. Atla-Nottingham 33:445

Kennard RW, Stone LA (1969) Computer aided design of experiments. Technometrics 11(1):137–148

Leardi R (1996) Genetic algorithms in molecular modeling. Elsevier, pp 67–86

Lee JA, Roh EJ, Oh C-H, Lee SH, Sim T, Kim JS, Yoo KH (2015) Synthesis of quinolinylaminopyrimidines and quinazolinylmethylaminopyrimidines with antiproliferative activity against melanoma cell line. J Enzyme Inhib Med Chem 30(4):607–614

Liotta LA, Steeg PS, Stetler-Stevenson WG (1991) Cancer metastasis and angiogenesis: an imbalance of positive and negative regulation. Cell 64(2):327–336

Makrariya A, Pardasani K (2019) Numerical study of the effect of non-uniformly perfused tumor on heat transfer in women's breast during menstrual cycle under cold environment. Netw Modeling Anal Health Inform Bioinform 8(1):9

Melagraki G, Afantitis A, Sarimveis H, Koutentis PA, Markopoulos J, Igglessi-Markopoulou O (2007) Optimization of biaryl piperidine and 4-amino-2-biarylurea MCH1 receptor antagonists using QSAR modeling, classification techniques and virtual screening. J Comput Aided Mol Des 21(5):251–267

Melagraki G, Afantitis A, Sarimveis H, Koutentis PA, Kollias G, Igglessi-Markopoulou O (2009) Predictive QSAR workflow for the in silico identification and screening of novel HDAC inhibitors. Mol Divers 13(3):301–311

Mignatti P, Rifkin DB (1993) Biology and biochemistry of proteinases in tumor invasion. Physiol Rev 73(1):161–195

Muegge I, Oloff S (2006) Advances in virtual screening. Drug Discov Today Technol 3(4):405–411

Naik PA, Pardasani KR (2018) 2D finite-element analysis of calcium distribution in oocytes. Netw Model Anal Health Inform Bioinform 7(1):10

Rajer-Kanduč K, Zupan J, Majcen N (2003) Separation of data on the training and test set for modelling: a case study for modelling of five colour properties of a white pigment. Chemom Intell Lab Syst 65(2):221–229

Roskoski R (2012) MEK1/2 dual-specificity protein kinases: structure and regulation. Biochem Biophys Res Commun 417(1):5–10

Saini KS, Loi S, de Azambuja E, Metzger-Filho O, Saini ML, Ignatiadis M, Dancey JE, Piccart-Gebhart MJ (2013) Targeting the PI3K/AKT/mTOR and Raf/MEK/ERK pathways in the treatment of breast cancer. Cancer Treat Rev 39(8):935–946

Tropsha A, Gramatica P, Gombar VK (2003) The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. Mol Inform 22(1):69–77

Vaidya A, Jain S, Jain S, Jain AK, Agrawal RK (2014) Quantitative structure-activity relationships: a novel approach of drug design and discovery. J Pharm Sci Pharmacol 1(3):219–232

Viswanadhan VN, Ghose AK, Revankar GR, Robins RK (1989) Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. J Chem Inf Comput Sci 29(3):163–172

Wu C-P, Ambudkar SV (2014) The pharmacological impact of ATP-binding cassette drug transporters on vemurafenib-based therapy. Acta Pharm Sin B 4(2):105–111

Wu W, Zhang C, Lin W, Chen Q, Guo X, Qian Y, Zhang L (2015) Quantitative structure-property relationship (QSPR) modeling of drug-loaded polymeric micelles via genetic function approximation. PLoS One 10(3):e0119575

Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32(7):1466–1474

Young D (2004) Computational chemistry: a practical guide for applying techniques to real world problems. Wiley, Hoboken, New Jersey, United States

Zubrilov I, Sagi-Assif O, Izraely S, Meshel T, Ben-Menahem S, Ginat R, Pasmanik-Chor M, Nahmias C, Couraud P-O, Hoon DS (2015) Vemurafenib resistance selects for highly malignant brain and lung-metastasizing melanoma cells. Cancer Lett 361(1):86–96