Ibrahim et al. Beni-Suef University Journal of Basic and Applied Sciences (2020) 9:49
https://doi.org/10.1186/s43088-020-00077-5

Beni-Suef University Journal of Basic and Applied Sciences

RESEARCH Open Access

Quantitative structure-activity relationship, molecular docking, drug-likeness, and pharmacokinetic studies of some non-small cell lung cancer therapeutic agents

Muhammad Tukur Ibrahim*, Adamu Uzairu, Sani Uba and Gideon Adamu Shallangwa

Abstract

Background: Lung cancer has been reported to be among the leading cancer cases in the world. It was also reported to have caused a lot of death every year and accounted for about one-third of the whole cancer deaths in the globe. The main subset of lung cancers that accounts for about 85% of the problems of lung cancer raised above was non-small cell lung cancer (NSCLC). The most common cause of NSCLCs that mostly affects women and cigarette smokers was recognized to be overexpression of epidermal growth factor receptor tyrosine kinase (EGFR TK).

Results: Five models on thirty five (35) NSCLC therapeutic agents were developed via quantitative structure-activity relationship (QSAR) technique. The best model among them was selected and reported due to its fitness statistically with the following validation parameters: R^2 of 0.8764, R^2_{adj} of 0.8370, Q_{cv}^2 of 0.7655, R^2_{test} of 0.7024, and LOF of 0.3312. Molecular docking was used to elucidate the mode of binding interactions between the thirty five (35) NSCLC therapeutic agents and the binding pose of EGFR tyrosine kinase receptor (3IKA) in this research. Compound 29 was recognized to have the most excellent binding affinity of − 8.8 kcal/mol among others. The drug-likeness and pharmacokinetic properties of all the NSCLC therapeutic agents were predicted using SWISSADME, and none among the molecules under investigation violated more than the permissible limit of the conditions stated by Lipinski's RO5 filters. Five hit compounds were identified using molecular docking virtual screening. The five (5) hit compounds were further screened and identified compound 16 and 27 as excellent among them using their pharmacokinetic profiles and drug-likeness properties.

(Continued on next page)

* Correspondence: muhdtk1988@gmail.com
Department of Chemistry, Ahmadu Bello University, Zaria, Kaduna State, Nigeria

(Continued from previous page)
Conclusion: QSAR technique was used to build five models on thirty five (35) NSCLC therapeutic agents. The best model among them was reported because it is statistically significant with good validation parameters. The molecular docking result has identified five (5) hit compounds. The most common amino acid residues to all hit compounds under investigation were Glu762, Leu718, Lys745, and Val726 which might be responsible for the higher inhibitory activities/binding affinities of the compounds under investigation. Furthermore, these five (5) hit compounds were further subjected to drug-likeness and pharmacokinetic properties prediction to determine which among them have the best pharmacokinetic profile. Compounds 16 and 27 among the hit compounds were observed to have high chance of passive absorption by the gastrointestinal tract while the other three have low tendency of passive absorption. More so, only compounds 16 and 27 have higher bioavailability scores, and none of the two have more than one violation of the RO5 criteria. The cause of efficiency of compounds 16 and 27 might be as a result of good pharmacokinetic profiles and drug-likeness properties possessed by the molecules when compared to other hit compounds.

Keywords: QSAR, NSCLC, In silico, SWISSADME, Applicability domain

1 Background

Among the hurdles faced by medicinal chemists was the discovery of inhibitors for mutant-selective kinase and was among the primary interest for epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors [1]. The treatment of EGFR to control non-small cell lung cancers with the T790M resistance mutation prevails as a vital medical necessity [2].

Lung cancer has been reported to be among the leading cancer cases in the world. It was also reported to have caused a lot of death every year and accounted for about one-third of the whole cancer deaths. The main subset of lung cancers that accounted for about 85% of the problems of lung cancer was NSCLC [3]. The most common cause of NSCLCs that mostly affects women and cigarette smokers was recognized to be overexpression of EGFR tyrosine kinase. It was found in about 10–15% and 30–40% of the population of patients in Caucasia and Asia [3].

NSCLC therapeutic agents manifest a very high response rate in patients with stimulating changes of EGFR. NSCLC therapeutic agents are categorized into two different classes: the first class is reversible NSCLC therapeutic agents (first-generation EGFR inhibitors) which include gefitinib and erlotinib. The second class is referred to as irreversible inhibitors and consists of the second and third generation NSCLC therapeutic agents. The second and third generation NSCLC therapeutic agents include afatinib and osimertinib. All these classes of drugs mentioned were designed purposely for the treatment of NSCLC. Most especially, the first-generation reversible NSCLC therapeutic agents were designed to manage $EGFR^{L858R}$ mutations. The second-generation irreversible NSCLC therapeutic agents were designed for the treatment of $EGFR^{T790M}$ mutations. And the third-generation irreversible NSCLC therapeutic agents were designed for the medication of $EGFR^{T790M/L790M}$ double mutations [1, 4–6].

QSAR is a computer-aided molecular modeling technique which quantitatively relates experimentally determined biological activities (response variable) of a molecule and its physicochemical properties (molecular descriptors) [7]. In addition, QSAR modeling is used to develop a model which could be used to predict the activities of newly designed small molecules [8]. Molecular docking is an in silico virtual screening method applied in computer-aided drug design used to elucidate how ligand and receptor interact with one another using their individual 3D structures [9]. The drug-likeness and pharmacokinetic properties of a drug give an insight on how the body responds to the administration of this drug. Therefore, there is a need to study the drug-likeness and pharmacokinetic properties of this drug before it reaches the final (clinical) stage [10].

The aim of this work is to develop a model with good predictive power using QSAR modeling technique, to screen and identify hit among the compounds under investigation (by elucidating the mode of binding interactions between the NSCLC therapeutic agents (ligands) and the EGFR tyrosine kinase enzyme) using molecular docking simulation, and also to predict their drug-likeness and pharmacokinetic properties.

2 Method

2.1 Sourcing of dataset

A set of thirty five (35) *N*-(5-((5-chloro-4-((2-(isopropylsulfonyl) phenyl) amino) pyrimidin-2-yl) amino)-4-methoxy-2-(4-methyl-1, 4-diazepan-1-yl) phenyl) acrylamide derivatives as potent NSCLC therapeutic agents with their inhibitory activities (GI_{50}) in μM, synthesized under the same condition sharing the same assayed procedure with significant variations in their structure and potency, were downloaded from the literature of Chen et al. for this research [11]. The corresponding inhibitory activities (GI_{50}) of these potent NSCLC therapeutic

agents were then converted to their pGI_{50} using Eq. 1 shown below [12]. Table 1 presents the structural formula, GI_{50}, and pGI_{50} for all the dataset used in this research.

$$pGI_{50} = - \log GI_{50} \times 10^{-6} \quad (1)$$

Table 1 Structural formula, GI_{50}, and pGI_{50} of the data set

S/No.	Structural formula	GI_{50} (μM)	pGI_{50} (μM)
1[n]	$C_{33}H_{43}ClN_8O_4S$	0.023	7.6383
2	$C_{28}H_{34}ClN_7O_6S_2$	0.04	7.398
3	$C_{28}H_{34}ClN_7O_4S$	0.043	7.3665
4	$C_{29}H_{36}ClN_7O_4S$	0.015	7.8239
5	$C_{30}H_{38}ClN_7O_4S$	0.025	7.6021
6	$C_{30}H_{38}ClN_7O_5S$	0.041	7.3872
7	$C_{33}H_{42}ClN_7O_4S$	0.077	7.1135
8 [n]	$C_{27}H_{31}ClN_6O_5S$	1.2	5.9208
9	$C_{28}H_{33}ClN_6O_4S$	0.24	6.6198
10	$C_{32}H_{41}ClN_8O_3S$	7.1	5.1487
11	$C_{31}H_{40}ClN_7O_4S$	0.3	6.5229
12	$C_{34}H_{46}N_8O_4S$	0.027	7.5686
13	$C_{30}H_{39}N_7O_4S$	0.043	7.3665
14	$C_{29}H_{37}N_7O_4S$	0.19	6.7212
15	$C_{26}H_{27}ClF_3N_7O_2$	0.26	6.5850
16	$C_{25}H_{26}Cl_2FN_7O_2$	0.1	7.0000
17	$C_{25}H_{26}ClF_2N_7O_2$	0.13	6.8861
18	$C_{24}H_{28}ClN_7O_2S$	0.22	6.6576
19 [n]	$C_{30}H_{37}ClN_8O_3$	0.95	6.0223
20	$C_{27}H_{29}ClF_3N_7O_2$	1.8	5.7447
21	$C_{27}H_{33}ClN_8O_4S$	0.031	7.5086
22	$C_{28}H_{35}ClN_7O_3P$	0.026	7.5850
23	$C_{29}H_{37}ClN_7O_3P$	0.22	6.6576
24	$C_{26}H_{30}ClN_6O_4P$	0.78	6.1079
25	$C_{27}H_{33}ClN_7O_5PS$	0.89	6.0506
26	$C_{32}H_{41}ClN_7O_3P$	0.041	7.3872
27	$C_{27}H_{31}ClN_8O_3$	0.05	7.3010
28 [n]	$C_{29}H_{35}ClN_8O_3$	0.079	7.1024
29	$C_{28}H_{31}ClN_8O_4$	0.25	6.6021
30	$C_{32}H_{40}ClN_9O_3$	0.21	6.6778
31	$C_{27}H_{29}ClN_8O_2$	0.44	6.3566
32	$C_{28}H_{31}ClN_8O_2$	0.43	6.3665
33	$C_{25}H_{24}ClN_7O_3$	1.2	5.9208
34 [n]	$C_{31}H_{36}ClN_9O_2$	0.2	6.6989
35	$C_{29}H_{38}ClN_7O_4S$	0.7	6.1549

[n]Test set

2.2 Stable structure generations and structure sketching

Before stable structure generations, the sketching of the 2D structures of the studied NSCLC therapeutic agents must be done, and this was achieved using the Chemdraw software version 12.0.2 [13]. The Spartan 14 software was used to transform the 2D structures of the sketched NSCLC therapeutic agents to 3D structures before energy minimization (it is achieved by direct importation of the 2D structures to the interface of the software). Also, prior to stable structure generations, there is need to remove constrain from the generated 3D structures, and this was achieved via energy minimization [14]. Stable structure generation is a process of determining the optimum structure of a compound, and this was performed by utilizing the Spartan 14 software. The determination of the optimum structure of all the NSCLC therapeutic agents was achieved adopting density functional theory method at B3LYP/6-311G* level of theory [15].

2.3 Independent variable (descriptors) computation, removal of constant/redundant variables, and data separation

For the computation of the independent variables (descriptors), the most stable structures generated were saved in SDF, a file format that is recognized by the software used in computing descriptors (PaDEL descriptor tool kit) [16]. PaDEL descriptor tool kit was used to compute both fragment count descriptors, topological descriptors, and geometrical descriptors [17]. Pretreatment of data is very vital in QSAR modeling which helps in eliminating constant and redundant descriptors from the data before model development so as to allow GFA select most significant descriptors. In present study, data pre-treatment was performed manually. Another crucial point in QSAR modeling is development of model building (training) and validation (test) sets. As the name implies, model building set is used to develop the model, and the validation set is used in verifying the built model. Data division software retrieved from DTC lab was moreover utilized in splitting the data into model building set which contains 30 molecules and validation set of 5 molecules utilizing Kennard-Stone algorithm in this regard [18].

2.4 Building of the model

In developing the models, the actual pGI_{50} was used as the response parameter while the descriptors were used as independent parameter. Variable selection is very important in building QSAR models. In view of this, the models were built by adopting multi-linear regression (MLR) analysis using genetic function approximation (GFA) method in which it creates an original population of descriptor sets and determines the most suitable set

from it by utilizing evolutionary crossover and mutation speculators which generates a succeeding derivative population of descriptor sets [19]. One of the distinct characteristics of GFA is that it selects highly significant independent variables to generate thousands of models so as to choose the most significant among the generated models [20]. Equation 2 below presents the MLR-GFA equation for the model:

$$\mathrm{pGI}_{50} = A_1 b_1 + A_2 b_2 + \ldots\ldots + C \tag{2}$$

where A's are the descriptors, b's are the coefficient of the corresponding descriptors, and C is the regression constant.

2.5 Validation of the model built

Validation of QSAR model is of utmost importance. This is why a QSAR model is not considered valid unless it undergoes so many assessment, which if it passes then it can be used. The parameters used in evaluating or validating the quality of a QSAR model were the squared coefficient of correlation for the training set (R^2_{training}), adjusted R^2 (R^2_{adj}), cross-validation coefficient (Q_{cv}^2), and squared coefficient of correlation for the test set (R^2_{test}). The equations for the mentioned assessment parameters are given below [21]:

$$\mathrm{R}^2_{\text{training}} = 1 - \frac{\sum \left(\boldsymbol{x}_{obs.} - \boldsymbol{x}_{pred.}\right)^2}{\sum \left(\boldsymbol{x}_{obs.} - \overline{\boldsymbol{x}}_{training}\right)^2} \tag{3}$$

where $\boldsymbol{x}_{\mathbf{obs.}}$, $\boldsymbol{x}_{\mathbf{pred.}}$, and $\overline{\boldsymbol{x}}_{\mathbf{training}}$ represents the actual, estimated, and mean activities of the model building set. The $\boldsymbol{R^2}$ value was established to rely on the number of descriptors in the model.

Therefore, the $\boldsymbol{R^2}$ value must be adjusted. The adjusted $\boldsymbol{R^2}$ is computed utilizing Eq. 4 below:

$$\boldsymbol{R^2_{adj.}} = \mathbf{1} - \left(\mathbf{1} - \boldsymbol{R^2}\right)\frac{\boldsymbol{a} - \mathbf{1}}{\boldsymbol{a} - \boldsymbol{b} - \mathbf{1}} = \frac{(\boldsymbol{a} - \mathbf{1})\boldsymbol{R^2} -}{\boldsymbol{a} - \boldsymbol{b} + \mathbf{1}} \tag{4}$$

where $\boldsymbol{b}$ represents the number of descriptors used in the model and $\boldsymbol{a}$ represents the number of compounds in the model building set.

$$\boldsymbol{Q^2_{CV}} = \mathbf{1} - \frac{\sum_{i=1}^{n}\left(\boldsymbol{Y}_{\mathbf{exp.}} - \boldsymbol{Y}_{\mathbf{pred.}}\right)^2}{\sum_{i=1}^{n}\left(\boldsymbol{Y}_{\mathbf{exp.}} - \overline{\boldsymbol{Y}}\right)^2} \tag{5}$$

where $Y_{\text{exp.}}$, $Y_{\text{pred.}}$, and $\overline{Y}$ are trial, foretold, and the mean inhibition activity values of the training set compounds [22].

The generated model can then be validated externally to confirm its predictive power and reliability. It is achieved using the validation set compounds. The external predictive power of the model was estimated using the expression shown below [23]:

$$\boldsymbol{R^2_{test}} = \mathbf{1} - \frac{\sum \left(\boldsymbol{x}_{\mathbf{pred.test}} - \boldsymbol{x}_{\mathbf{exp.test}}\right)^2}{\sum \left(\boldsymbol{x}_{\mathbf{pred.test}} - \overline{\boldsymbol{x}}_{\mathbf{Training}}\right)^2} \tag{6}$$

where $\boldsymbol{x}_{\mathbf{pred.test}}$ and $\boldsymbol{x}_{\mathbf{exp.test}}$ are the estimated and actual activities of the validation set, and $\overline{\boldsymbol{x}}_{\mathbf{Training}}$ is the mean of actual activity of the model building set compounds.

Due to some reasons, the values of these parameters are okay and important but not enough to justify the reliability of a model [24]. In view of this, the model has to be subjected to other test such as applicability domain, variation inflation factor, and mean effect.

The multi-collinearity of all the independent variables in the reported model is ascertain by computing the variation inflation factors (VIF) for each. The VIF help in identifying whether these independent variables correlate with one another or not. There is no correlation between the descriptors if their estimated VIF values are equal to 1; there is high possibility of accepting the model if their estimated VIF values are between 1 and 5; and if their estimated VIF values are greater than 10, then the model is therefore rejected not accepted [25]. The VIF value can be calculated using the equation below:

$$\mathrm{VIF} = \frac{1}{1 - R^2} \tag{7}$$

In order to evaluate the individual contribution and participation of each descriptor to the selected model, the mean effect (ME) of each descriptor is therefore calculated. The equation used in calculating the ME is shown below:

$$\mathrm{MEj} = \frac{B_j \sum_{j=1}^{i=n} d_{ij}}{\sum_j^m B_j \sum_i^n d_{ij}} \tag{8}$$

where ME represents the mean effect of a descriptor j in a model, the coefficient of the descriptor J is represented by β_{j} in the model and the value of the independent variables for each compound in the training set is d_{ij}, n is the number of compounds in the training set, and m is the number of descriptor that appear in the model [26].

2.6 Evaluating of applicability domain

The domain of applicability is studied to ensure the reliability of the prediction of the built MLR model. It is also useful in identifying compounds that are distinct to the training set compounds (influential compounds) or response outliers (compounds with standardized residual outside the square area of the model). The method adopted in this research was the leverage approach which is the plot of the standardized residual against the leverages for both the training and test set compounds. The reported model was subjected to AD using the leverage approach [27].

2.7 Docking simulation

For the docking simulation, the virtual screening software used in this research were AutoDock Vina of Pyrex, Discovery studio, and UCSF Chimera on a Dell computer system Latitude E6520 to screen and identify hit compounds by elucidating the binding mode between the binding pose of the target receptor and the NSCLC therapeutic agents.

2.8 EGFR tyrosine kinase enzyme and ligand preparation for the docking simulation

The EGFR tyrosine kinase enzyme with pdb code: **3IKA** in complex with WZ4002 was downloaded from the Protein Data Bank (https://www.rcsb.org) and used as the target receptor for the NSCLC therapeutic agents in this research. Discovery Studio Visualizer version 16.1.0.15350 was adopted in preparing the EGFR tyrosine kinase enzyme for the docking simulation. The preparation process of the target receptor started by adding hydrogen, then followed by the elimination of co-ligands, water molecule, and heteroatoms from the structure of the target receptor and saved in protein data bank file format. The prepared structure of the target receptor is shown in Fig. 1. The NSCLC therapeutic agents were prepared by saving the already determined optimum structures in 2.2 above saved in protein data bank file with the help of the Spartan'14 wave software [14]. The prepared structure of one NSCLC therapeutic agent among the dataset is shown in Fig. 2.

2.9 Execution of the docking simulation

The docking simulation of the NSCLC therapeutic agents into the binding site of the target receptor (Met793, Ala743, Met790, Leu844, Leu844, Leu718, and Val726, these binding sites were determined by visualizing the co-crystalline structure of WZ4002 in the binding site of the enzyme) was carried out using AutoDock Vina of Pyrex software [28]. Re-coupling of the complexes for further investigation was achieved with the help of the UCSF Chimera software [29]. For further investigation of the binding mode interactions of the complexes, a discovery studio visualizer software was used to elucidate the 2D structures of all the reported complexes [30, 31].

2.10 Drug-likeness and pharmacokinetic property prediction

The drug-likeness and pharmacokinetic properties of these NSCLC therapeutic agents were predicted utilizing a free online web tool (SwissADME) (http://www.swissadme.ch/index.php) used in predicting drug-likeness and pharmacokinetic properties of drugs [32]. The input file format for SwissADME is simplified molecular input line entry specification (SMILES) which contains a unit compound by line separated by a space with a title or without a title. The computation can be setup when the molecule is ready by clicking on the "Run" button [32].

Lipinski's rule of five filter is mostly used as the criterion to ascertain whether a molecule is impermeable or badly absorbed. A molecule is considered to be orally bioavailable if it does not violate more than 2 of the RO5 [33].

3 Result

3.1 QSAR study

The results of the QSAR study are given in Tables 2, 3, 4, and 5 and Figs. 3 and 4.

The selected and reported model is given by the equation below with the following validation terms: R^2 of 0.8764, R^2_{adj} of 0.8370, $Q_{cv}{}^2$ of 0.7655, R^2_{test} of 0.7024, and LOF of 0.3312

$$\begin{aligned}\mathbf{pGI_{50}} = {} & 2.797519677*\mathbf{ATSC8c} - 1.977464485*\mathbf{MATS8s} \\ & - 1.229853317*\mathbf{GATS5p} - 0.735278765*\mathbf{VR1\_Dt} \\ & + 1.186969524*\mathbf{minssCH2} + 2.607601502*\mathbf{RDF120m} \\ & + 0.834211273*\mathbf{RDF125m} + 4.685695\end{aligned}$$

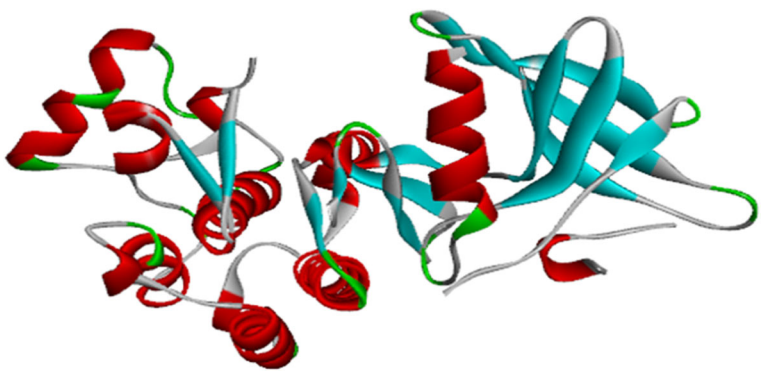

Fig. 1 Prepared structure of EGFR tyrosine kinase enzyme

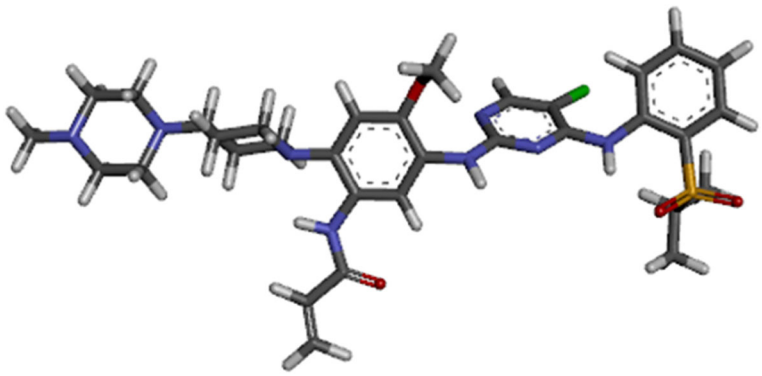

Fig. 2 Prepared structure of a NSCLC therapeutic agent

3.2 Docking simulation

The results of the docking simulation are presented in Table 6 and Fig. 6.

3.3 Drug-likeness and pharmacokinetic property prediction

The results of the drug-likeness and pharmacokinetic property prediction are shown in Tables 7 and 8 and Figs. 7 and 8 respectively.

4 Discussion

4.1 QSAR studies

Using the model building set, five (5) different models were built using MLR-GFA method. Among these five models, the best model was selected and reported since it has passed the minimum requirements for the evaluation of a valid QSAR model as reported by Veerasamy et al. as presented in Table 2 [23].

The descriptions of the descriptors contained in the reported model are shown in Table 3. The negative coefficients of **MATS8s**, **GATS5p**, and **VR1_Dt** descriptors clearly indicated their negative contribution to the inhibitory activities of the NSCLC therapeutic agents. It means that when the amount of these independent descriptors is reduced in the structures of these NSCLC therapeutic agents under investigation, there might be an improvement in the potency of these NSCLC therapeutic agents toward their target receptor (EGFR tyrosine kinase enzyme) and reverse is the case. On the other side, the positive coefficient of **ATSC8c**, **minssCH2**, **RDF120m**, and **RDF125m** descriptors in the model gave the positive contributions of these independent descriptors to the inhibitory activities of the NSCLC therapeutic agents under investigation. It means when the amount of these descriptors in the compositions/structures of these NSCLC therapeutic agents are increased, there might be an improvement in the potency of these NSCLC therapeutic agents toward their target receptor and vice versa.

4.1.1 Description of the descriptors that appear in the reported model

ATSC8c is an average centered Broto-Moreau autocorrelation; the ATS descriptor is a graph invariant describing how the property considered is distributed along the topological structure and can be seen as a special case in which other types of descriptors can also be derived from [34]. The recognized spatial autocorrelation on a molecular graph G is given as

$$ATS_k = \frac{1}{2} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} w_i \cdot w_j \cdot \delta(d_{ij}; k) = \frac{1}{2} \cdot \left(\mathrm{w}^{\mathrm{T}} \cdot {}^{k}\mathrm{B} \cdot \mathrm{w}\right)$$

MATS8s is a Moran autocorrelation which is applied to a molecular graph. Moran coefficient usually takes value in the interval [− 1, + 1]. Positive autocorrelation corresponds to positive values of the coefficient whereas negative autocorrelation produces negative values [34]. It can be defined as

Table 2 General limit required for the QSAR model assessment

Parameter	Details	Accepted value	Selected model
R^2_{trng}	Squared correlation coefficient of training set	≥ 0.6	0.8764
Q_{cv}^2	Cross-validation coefficient	≥ 0.5	0.7655
$R^2 - Q^2$	Difference between R^2 and Q^2	≤ 0.3	0.1109
$N_{(test\ set)}$	Minimum number of external test set	≥ 5	5
$R^2_{ext.}$	Squared correlation coefficient of test set	≥ 0.5	0.7024

Table 3 Descriptions, full name, and categories of descriptors contained in the reported model

S/no	Description	Full name	Category
1	**ATSC8c**	Average centered Broto-Moreau autocorrelation—lag 8/weighted by charges	2D
2	**MATS8s**	Moran autocorrelation—lag 8/ weighted by I-state	2D
3	**GATS5p**	Geary autocorrelation—lag 5/ weighted by polarizabilities	2D
4	**VR1_Dt**	Randic-like eigenvector-based index from detour matrix	2D
5	**minssCH2**	Minimum atom-type E-State: –CH2–	2D
6	**RDF120m**	Radial distribution function—120/ weighted by relative mass	3D
7	**RDF125m**	Radial distribution function—125/ weighted by relative mass	3D

$$I_k = \frac{\frac{1}{\Delta_k} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} (w_i - \overline{w}) \cdot \delta(d_{ij}; k)}{\frac{1}{A} \cdot \sum_{i=1}^{A} (w_i - \overline{w})^2}$$

Radial distribution function descriptors (**RDF** descriptors) were suggested based on a radial arrangement function distinct from that generally used to determine molecular changes I (s) (Hemmer, Steinhauer et al., 1999). The radial distribution function chosen here is that one frequently utilized for the description of the diffraction patterns gotten in powder X-ray diffraction experiments.

Ideally, the radial distribution function of a collection of atoms B may be described as the possible occurrence to obtain an atom in a spherical volume of radius *R*. The common mode of the radial distribution function is expressed by the equation below

$$g(R) = f \cdot \sum_{i=1}^{A-1} \sum_{j=i+1}^{A} w_i \cdot w_j \cdot e^{-\beta \cdot (R - r_{ij})^2}$$

Table 4 shows the pGI_{50}, predicted pGI_{50}, and the residual values for all the molecules under investigation. The high predicted power of the reported model was confirmed by the low residual values observed between the experimental and predicted pGI_{50} in the table (meaning that the reported model was reliable with high predicted power). Furthermore, Fig. 3 presents the plot of the predicted pGI_{50} versus actual pGI_{50} for the test and training sets compounds, the distribution of the predicted pGI_{50} and the actual pGI_{50} of the test and training set compounds throughout the line reaffirmed the reliability of the model. More so, the R^2 values of both the internal validation (0.8175) and that of the plot (0.8764) agreed with one another which further confirmed the stability and reliability of the reported model. On the other hand, Fig. 4 presents the scatter plot of the residuals against actual pGI_{50} in which the unusual occurrence of these residuals of both sets on the upper and lower sides of zero on the plot confirm that the reported model was free from methodological error (systematic deviations).

Table 4 The pGI_{50}, predicted pGI_{50}, and the residual values for the studied molecules

S/No	pIG_{50} (nM)	Predicted pIG_{50}	Residual values
1[n]	7.6383	8.0744	0.4362
2	7.398	7.2607	0.1373
3	7.3665	7.1608	0.2058
4	7.8239	7.9458	− 0.1219
5	7.6021	7.5017	0.1003
6	7.3872	7.4848	− 0.0976
7	7.1135	7.2791	− 0.1656
8 [n]	5.9208	6.0323	0.1115
9	6.6198	6.5679	0.0518
10	5.1487	5.2937	− 0.1449
11	6.5229	6.5563	− 0.0334
12	7.5686	6.9651	0.6035
13	7.3665	7.6696	− 0.3031
14	6.7212	6.9231	− 0.2018
15	6.5850	6.5074	0.0776
16	7.0000	6.8746	0.1254
17	6.8861	6.6254	0.2606
18	6.6576	6.9125	− 0.2549
19 [n]	6.0223	6.2074	0.1851
20	5.7447	6.1388	− 0.3941
21	7.5086	7.4315	0.0772
22	7.5850	7.2457	0.3393
23	6.6576	6.5749	0.0826
24	6.1079	5.8741	0.2338
25	6.0506	6.3517	− 0.3011
26	7.3872	7.2227	0.1645
27	7.3010	7.4012	− 0.1001
28 [n]	7.1024	7.2748	0.1724
29	6.6021	6.7544	− 0.1523
30	6.6778	7.0444	− 0.3666
31	6.3566	6.3405	0.0161
32	6.3665	6.2983	0.0682
33	5.9208	6.0691	− 0.1482
34 [n]	6.6989	6.0783	− 0.6207
35	6.1549	5.9133	0.2416

[n]Test set

Table 5 VIF, ME, and correlation statistical analysis of descriptors of the reported model

	ATSC8c	*MATS8s*	*GATS5p*	*VR1_Dt*	*minssCH2*	*RDF120m*	*RDF125m*	*VIF*	*ME*
ATSC8c	1							2.32682	0.796318
MATS8s	0.617583	1						1.794148	− 0.43128
GATS5p	0.102951	0.078452	1					1.536439	− 0.35878
VR1_Dt	0.272343	0.192376	0.194143	1				1.64444	− 0.05016
minssCH2	− 0.41896	− 0.21208	0.33861	0.11543	1			1.555337	0.402876
RDF120m	− 0.03966	0.134895	0.453081	0.366743	0.266443	1		1.535671	0.438001
RDF125m	0.147709	0.257319	0.286406	0.506069	0.098872	0.320932	1	1.498465	0.203025

The correlation statistical analysis on the independent descriptors in the reported model shown in Table 5 indicated that no relationship exists between the descriptors contained in the reported model. This clearly portrayed the high performance of the descriptors used in developing the reported model.

The variation inflation factor (VIF) values were further used to confirm if there is multi-collinearity problem or not in the descriptors of the training set used in building the model. The VIF of all descriptors in the training set were estimated and realized to be within the acceptable range presented in Table 5 (meaning the values are less than 10 for all the descriptors). This confirms the absence of multi-collinearity problem in the descriptors used in building the reported model.

The mean effect (ME) values for all the descriptors were computed to ascertain the participation and individual contribution of a descriptor in opposition to other ones in the selected model and presented in Table 5. The indicator for either increase or decrease in potency of the molecules is the sign of the coefficient of each descriptor in the model. If a descriptor in the model has a positive coefficient it means that an increase in such descriptor may increase the potency of the molecules. But when a descriptor has negative coefficient, it indicates that an increase in such descriptor may decrease the potency of the molecules. Whereas the coefficient of the descriptors indicate the degree of contribution of each descriptor in the model. It is observed that from the model and ME values (Table 5), **ATSC8c** descriptor gave the highest positive contribution both in the model and ME analysis with + 2.797519677 and + 0.796318. **MATS8s** gave the lowest negative contribution in both the model and ME analysis with − 1.977464485 and − 0.43128.

The applicability domain (AD) of the reported model was achieved by the plot of the standardized residuals against leverages of both the test and training sets (Williams' plot) as shown in Fig. 5. The AD is carried out to identify compounds with standardized residuals greater than three standard deviation unit (outliers) and compounds with leverage values greater than the warning leverage h* (influential) in the data used in building the model. Apart from that, it is also used to ascertain the

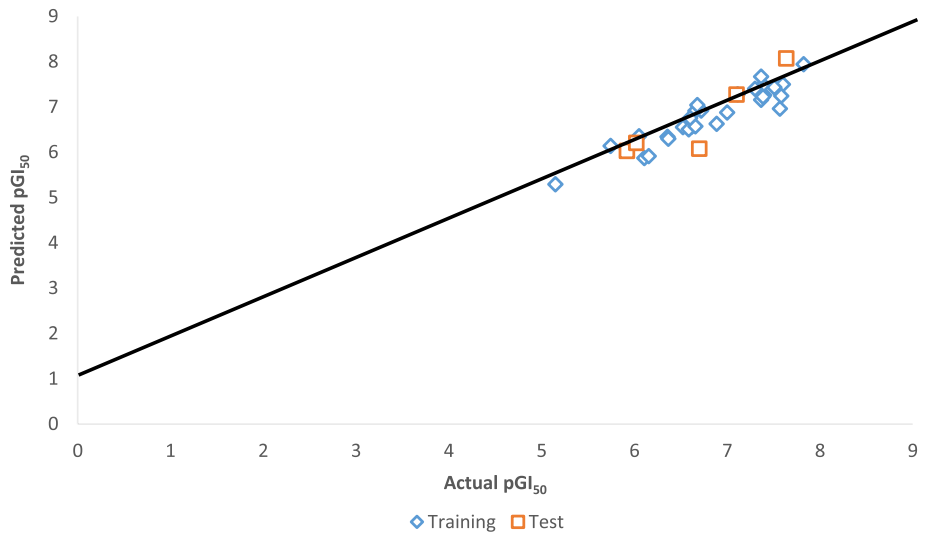

Fig. 3 Scatter plot of predicted pGI_{50} versus the actual pGI_{50} for the reported model

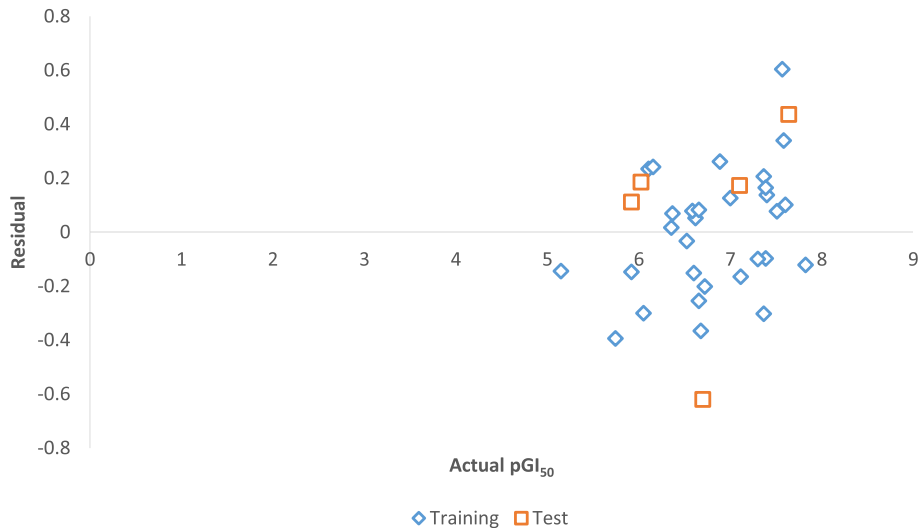

Fig. 4 Plot of residuals versus the actual pGI_{50} for the reported model

quality of prediction of a model and prevent the misuse of the results obtain by the model. From the plot, four (4) compounds (influential compounds) of the test set with their leverage value greater than the threshold/warning leverage (h*) of 0.83 were identified. No influential or response outlier was identified from the training set which means the model was valid and void. It is very paramount to decipher that these molecules with leverage value greater than the threshold are not put into consideration when designing new NSCLC therapeutic agents. These molecules might be structurally different from those used to generate the reported model and thus may have different mechanism of action

4.2 Docking simulation

Molecular docking as in silico virtual screening tool was used to elucidate the nature of binding interactions between the NSCLC therapeutic agents and the binding pose of EGFR tyrosine kinase receptor (pdb code: 3IKA which is selected based on published literature) in this research (Supplementary Table 1). Five hit compounds were identified by the virtual screening technique. Compound 29 was identified to have the highest binding affinity of − 8.8 kcal/mol (Table 6). The compound was seen to have interacted with the binding pose of EGFR tyrosine kinase receptor through hydrophobic bond interaction with Phe795, Gly796, Ala743, Leu844, Leu718, Val726, Ala743, and Lys745 amino acid residues of the EGFR tyrosine kinase receptor. More so, it also interacted with EGFR tyrosine kinase receptor via electrostatic bond interaction with lys745 amino acid residue. Next compound in the trend with higher binding affinity (− 8.7 kcal/mol) was compound 12 as shown in Table 6. The interactions of the compound in the binding pose of the EGFR receptor were through hydrogen bond with UNK1 Arg841, Asp855, Glu762, and Glu762 amino acid residues with bond distances of 2.49587 (Å), 3.7987 (Å), 3.25863 (Å), 3.72606 (Å), and 3.72287 (Å). It also interacted with the active site of the EGFR receptor via hydrophobic interactions with Leu718, Leu718,

Table 6 The ligand-receptor, binding affinity, hydrogen bond, bond distance, and other interaction of some selected ligands

Ligand-receptor (3IKA)	Binding affinity (Kcal/mol)	Hydrogen bond	Bond distance (Å)	Halogen, hydrophobic, and other amino acid residues
Complex 4	− 8.4	Thr854 and Asp855	2.31432 and 2.5532	Lys745, Cys797, Ala743, Leu844, Leu718, Val726, Ala743, and Lys745
Complex 12	− 8.7	Unk1, Arg841, Asp855, Glu762, and Glu762	2.4959, 3.7987, 3.2586, 3.7260, and 3.7228	Leu718, Leu718, Lys745, Val726, Leu844, Lys728, and Leu792
Complex 16	− 8.6	Met793, Lys728, Glu762, and Asp855	2.8649, 2.2705, 3.5998, and 3.6036	Leu718, Leu718, Lys745, Val726, Leu844, Lys728, and Leu792
Complex 27	− 8.5	Glu762	3.68933	Leu718, Leu718, Lys745, Val726, Leu844, Lys728, and Leu792
Complex 29	− 8.8			LYS745, PHE795, GLY796, ALA743, LEU844, LEU718, VAL726, ALA743, and LYS745

Table 7 Drug-likeness properties

Molecule	MW	No. of H-bond acceptors	No. of H-bond donors	TPSA	WLOGP	Lipinski's RO5 violations
Molecule 4	614.16	7	3	137.17	5.2	1
Molecule 12	662.85	8	3	140.41	4.55	2
Molecule 16	546.42	6	3	94.65	4.76	1
Molecule 27	551.04	6	4	123.75	2.91	1
Molecule 29	579.05	6	4	140.82	2.83	2

Lys745, Val726, Leu844, Lys728, and Leu792 amino acid residues. Also, among the compounds with good binding affinity was compound 16. The interaction of compound 16 in the binding pose of the EGFR receptor was through hydrogen bond with Met793, Lys728, Glu762, and Asp855 residues with bond distances of 2.86496 (Å), 2.27045 (Å), 2.54982 (Å), 3.59983 (Å), and 3.60362 (Å) respectively. It also interacted with the binding pose of the EGFR receptor via hydrophobic interaction with Leu718, Leu718, Lys745, Val726, Leu844, Lys728, and Leu792. The rest other two complexes among the reported ones interacted in the binding pose of the EGFR receptor through electrostatic interactions, hydrogen bond interactions, and hydrophobic bond interactions as shown in Table 6. Figure 6 showed the 2D structures of compounds 29, 12, and 16 in complex with the receptor (3IKA). Based on the molecular docking results, the most common amino acid residues to all of the hit compounds under investigation were Glu762, Leu718, Lys745, and Val726 (Table 6), and these important amino acid residues might be responsible for the higher binding affinity of the reported compounds.

4.3 Drug-likeness and pharmacokinetic property prediction

The drug-likeness and pharmacokinetic properties of all the NSCLC therapeutic agents were predicted and presented in Supplementary Table 2 and 3. Based on the results of the molecular docking, the drug-likeness properties of the hit compounds ware reported and presented in Tables 7 and 8. From the table, no molecule among these reported ones violated more than the permissible limit of the conditions stated by Lipinski's rule of five filters (MW˂ 500, HBD ≤ 5, HBA ≤ 10, Log p ≤ 5, and PSA ˂ 140 Å2). As such, these molecules are expected to be very active pharmacologically. The bioavailability radar plot of lipophilicity, size, polarity, solubility, saturation, and flexibility further reaffirmed the drug-likeness properties of all the reported molecules (Fig. 7). The painted pink area shows the range for each property (XLOGP3 between − 0.7 and + 5.0, MW between 150 and 500 g/mol, TPSA between 20 and 130 A^2, log S not higher than 6, fraction of carbons in the sp3 hybridization not less than 0.25, and no more than 9 rotatable bonds). Based on the condition mentioned, all the molecules might be orally bioavailable even though they were all too flexible and lipophilic.

Table 8 presents the pharmacokinetic properties of the reported molecules. From the table, only molecules 16 and 27 have high probability of passive absorption by the gastrointestinal tract while others have low tendency of passive absorption. None of the reported molecules was found to have high probability of brain penetration. Molecules 12, 27, and 29 were predicted to be actively effluxed by P-gp and the other two were predicted as non-substrate of P-gp. Also, only molecules 16 and 27 have higher bioavailability scores (this confirmed the oral bioavailability and permeability of these molecules among the reported molecules and also they have low toxicity level and good absorption properties). The boiled-egg plot (Fig. 8) of TPSA against WLOGP was used to portray the graphical presentation of the brain penetration and gastrointestinal absorption of the reported molecules. From Fig. 8, it can be clearly seen that all the NSCLC therapeutic agents were outside BBB region (yellow) but some were within the GI absorption region (white color) and some were predicted to be actively effluxed by P-gp (blue in color) and then some were predicted as non-substrate of P-gp (red color).

5 Conclusion

In conclusion, QSAR technique was used to build a model with a very high predictive power on some thirty five (35) NSCLC therapeutic agents. The reported model was

Table 8 Pharmacokinetic properties

Molecule	Gastrointestinal absorption	Brain penetration	Pgp substrate	Bioavailability score
Molecule 4	Low	No	No	0.17
Molecule 12	Low	No	Yes	0.17
Molecule 16	High	No	No	0.55
Molecule 27	High	No	Yes	0.55
Molecule 29	Low	No	Yes	0.17

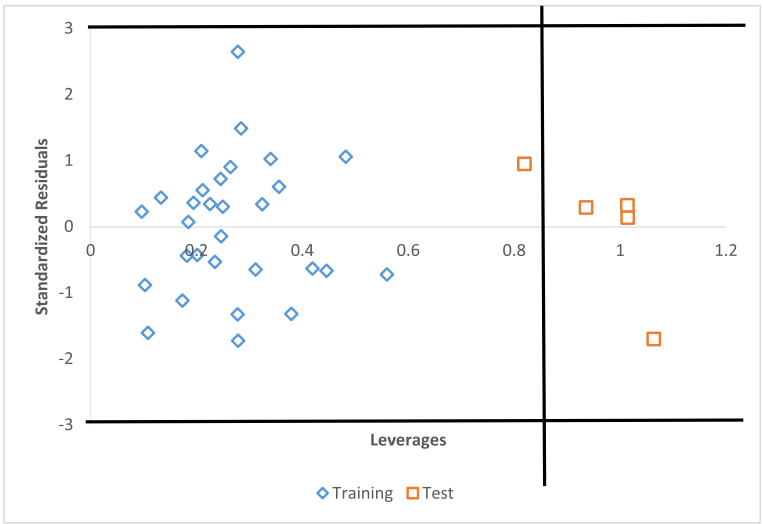

Fig. 5 Williams' plot of the selected model

found to be statistically fit by passing validation techniques employed on it with the validation parameters: R^2 of 0.8764, R^2_{adj} of 0.8370, $Q_{cv}{}^2$ of 0.7655, R^2_{test} of 0.7024 and LOF of 0.3312 such as internal and external validations and AD. The molecular docking results showed that the most common amino acid residues to all of the reported complexes were Glu762, Leu718, Lys745, and Val726, and these important amino acid residues might be responsible for the higher inhibitory activities/binding affinity of the

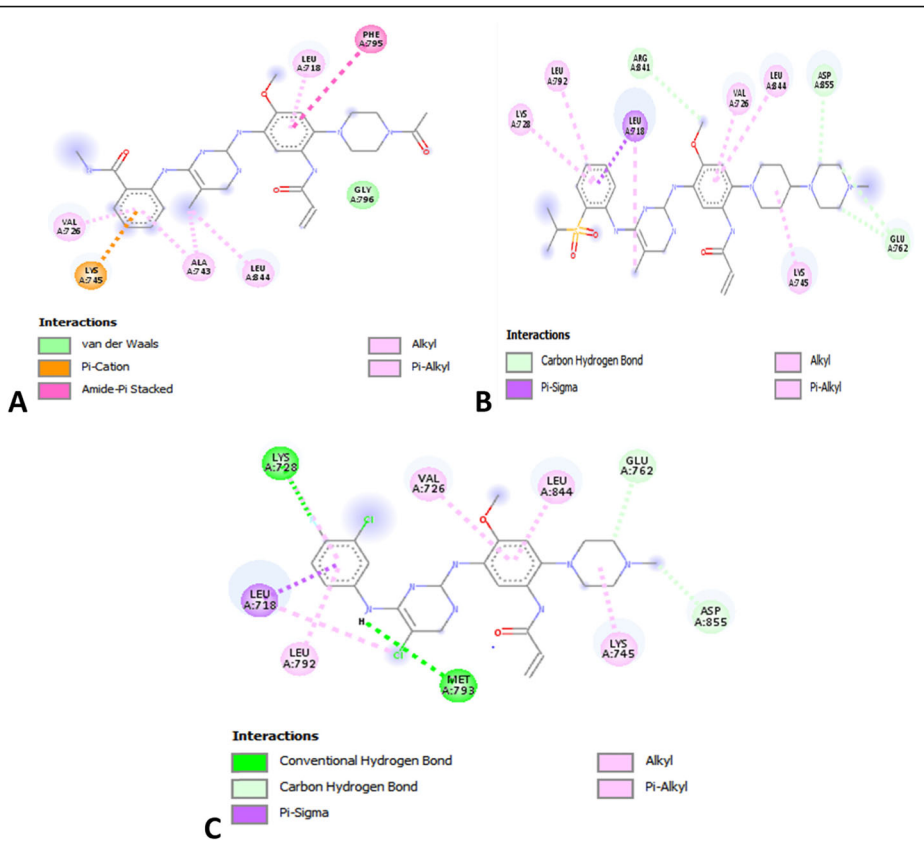

Fig. 6 2D view of **a** complex 29, **b** complex 12, and **c** complex 16 using Discovery studio visualizer

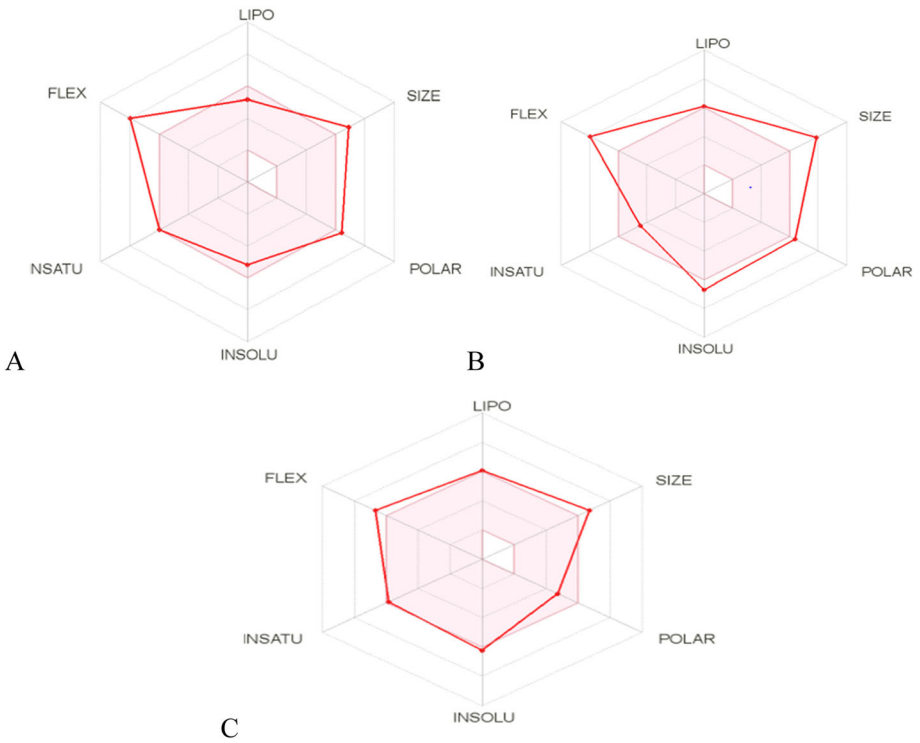

Fig. 7 The plot of lipophilicity, size, polarity, solubility, saturation, and flexibility of **a** molecule 29, **b** molecule 12, and **c** molecule 16

reported compounds. More so, the drug-likeness and pharmacokinetic properties of all the NSCLC therapeutic agents were predicted using SwissADME and indicated that molecules 16 and 27 among the hit have high probability of passive absorption by the gastrointestinal tract while the other three have low tendency of passive absorption and also none of the reported molecules was found have high probability of brain penetration. Also, only molecules 16 and 27 have higher bioavailability scores. Based on this finding, it is suggested that when designing new NSCLC therapeutic agents these hit compounds with good binding affinity and pharmacokinetic profile should be considered for structural modifications. And also, in vivo and in vitro assay for the ADME properties should be validated experimentally.

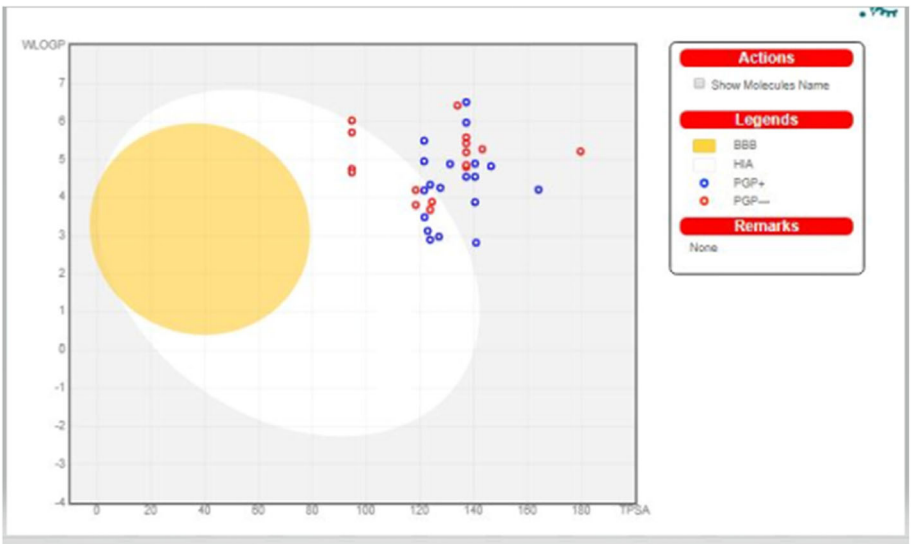

Fig. 8 The plot of WLOGP against TPSA for all the molecules

6 Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s43088-020-00077-5.

Additional file 1: Supplementary Table 1. The Ligand-Receptor, Binding affinity, Hydrogen bond, Bond distance, and other interaction for the remaining compounds. **Supplementary Table 2.** Drug-likeness properties for the compounds under investigation. **Supplementary Table 3.** ADME properties for the compounds under investigation.

Abbreviations

QSAR: Quantitative structure-activity relationship; MLR: Multi-linear regression; GFA: Genetic function algorithm; DFT: Density function theory; B3LYP: Becke's three-parameter read-Yang-Parr hybrid; PDB: Protein Data Bank; NSCLC: Non-small cell lung cancer agents; EGFR: Epidermal growth factor receptor; VIF: Variation inflation factor; MF: Mean effect; ADME: Absorption, distribution, metabolism, and excretion

Acknowledgements

The authors acknowledge the technical effort of Ahmadu Bello University, Zaria–Nigeria.

Authors' contributions

1. MTI: contributed throughout the research work. 2. AU: gives directives and technical advices. 3. GAS: partakes in technical activities. 4. SU: also partakes in technical activities. All authors have read and approved the manuscript.

Funding

The authors declare no funding has been received.

Availability of data and materials

Not applicable

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare no competing interest.

Received: 1 February 2020 Accepted: 7 September 2020
Published online: 10 December 2020

References

1. Song J, Jang S, Lee JW, Jung D, Lee S, Min KH (2019) Click chemistry for improvement in selectivity of quinazoline-based kinase inhibitors for mutant epidermal growth factor receptors. Bioorg Med Chem Lett 29:477–480
2. Hanan EJ et al (2016) 4-Aminoindazolyl-dihydrofuro [3, 4-d] pyrimidines as non-covalent inhibitors of mutant epidermal growth factor receptor tyrosine kinase. Bioorg Med Chem Lett 26:534–539
3. Kong L-L, Ma R, Yao M-Y, Yan X-E, Zhu S-J, Zhao P, Yun C-H (2017) Structural pharmacological studies on EGFR T790M/C797S. Biochem Biophys Res Commun 488:266–272
4. Cross DA et al (2014) AZD9291, an irreversible EGFR TKI, overcomes T790M-mediated resistance to EGFR inhibitors in lung cancer. Cancer Disc 4:1046–1061
5. Solca F et al (2012) Target binding properties and cellular activity of afatinib (BIBW 2992), an irreversible ErbB family blocker. J Pharmacol Exp Ther 343: 342–350
6. Tsao M-S et al (2005) Erlotinib in lung cancer—molecular and clinical predictors of outcome. N Engl J Med 353:133–144
7. Ojha Lokendra K, Rachana S, Rani BM (2013) Modern drug design with advancement in QSAR: a review. Int J Res Biosci 2:1–12
8. Abdulfatai U, Uba S, Umar BA, Ibrahim MT (2019) Molecular design and docking analysis of the inhibitory activities of some α_substituted acetamido-N-benzylacetamide as anticonvulsant agents SN. Appl Sci 1:499
9. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov *3*:935
10. Khan MF, Verma G, Akhtar W, Shaquiquzzaman M, Akhter M, Rizvi MA, Alam MM (2016) Pharmacophore modeling, 3D-QSAR, docking study and ADME prediction of acyl 1, 3, 4-thiadiazole amides and sulfonamides as antitubulin agents. Arab J Chem
11. Chen Y et al (2017) Discovery of N-(5-((5-chloro-4-((2-(isopropylsulfonyl) phenyl) amino) pyrimidin-2-yl) amino)-4-methoxy-2-(4-methyl-1, 4-diazepan-1-yl) phenyl) acrylamide (CHMFL-ALK/EGFR-050) as a potent ALK/EGFR dual kinase inhibitor capable of overcoming a variety of ALK/EGFR associated drug resistant mutants in NSCLC. Eur J Med Chem 139:674–697
12. Abdullahia, M., Shallangwaa, G. A., Ibrahima, M. T., Bello, A. U., Arthura, D. E., Uzairua, A., Mamzaa, P. (2018) QSAR studies on some C14-urea tetrandrine compounds as potent anti-cancer agents against leukemia cell line (K562) JKS-S
13. Mills, N. (2006). ChemDraw Ultra 10.0 CambridgeSoft, 100 CambridgePark Drive, Cambridge, MA 02140. www. cambridgesoft. com. Commercial Price: 1910fordownload, 2150 for CD-ROM; Academic Price: 710fordownload, 800 for CD-ROM: ACS Publications.
14. Ibrahim MT, Uzairu A, Shallangwa GA, Uba S (2019a) QSAR modelling and docking analysis of some thiazole analogues asα-glucosidase inhibitors. J Eng Exact Sci 5:0257–0270
15. Kohn W, Becke AD, Parr RG (1996) Density functional theory of electronic structure. J Phys Chem 100:12974–12980
16. Ibrahim MT, Uzairu A, Shallangwa GA, Uba S (2020a) Computer-aided molecular modeling studies of some 2, 3-dihydro-[1, 4] dioxino [2, 3-f] quinazoline derivatives as EGFR WT inhibitors. Beni-Suef Univ J Basic Appl Sci 9:1–10
17. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474
18. Kennard RW, Stone LA (1969) Computer aided design of experiments. Technometrics 11:137–148
19. Ibrahim MT, Uzairu A, Uba S, Shallangwa GA (2020b) Computational modeling of novel quinazoline derivatives as potent epidermal growth factor receptor inhibitors. Heliyon 6:e03289
20. Adedirin O, Uzairu A, Shallangwa GA, Abechi SE (2018a) Computational studies on α-aminoacetamide derivatives with anticonvulsant activities. Beni-Suef Univ J Basic Appl Sci 7:709–718
21. Grisoni F, Ballabio D, Todeschini R, Consonni V (2018) Molecular descriptors for structure–activity applications: a hands-on approach. Comput Toxicol:3–53 Springer
22. Mustapha A, Shallangwa G, Ibrahim MT, Bello AU, Ebuka DA, Uzairu A, Mamza P (2018) QSAR studies on some C14-urea tetrandrine compounds as potent anti-cancer against leukemia cell line (K562). J Turkish Chem Soc Section A Chem 5:1387–1398
23. Veerasamy R, Rajak H, Jain A, Sivadasan S, Varghese CP, Agrawal RK (2011) Validation of QSAR models-strategies and importance. Int J Drug Design Disc 3:511–519
24. Tropsha A, Bajorath JR (2015) Computational methods for drug discovery and design. ACS Publications
25. Beheshti A, Pourbasheer E, Nekoei M, Vahdani S (2016) QSAR modeling of antimalarial activity of urea derivatives using genetic algorithm–multiple linear regressions. J Saudi Chem Soc 20:282–290
26. Adedirin O, Uzairu A, Shallangwa GA, Abechi SE (2018b) QSAR and molecular docking based design of some n-benzylacetamide as γ-aminobutyrate-aminotransferase inhibitors. J Eng Exact Sci 4:0065–0084
27. 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 22:69–77
28. Trott O, Olson AJ (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31:455–461
29. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera—a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612
30. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol 5:e1000585
31. Rizvi SMD, Shakil S, Haneef M (2013) A simple click by click protocol to perform docking: AutoDock 4.2 made easy for non-bioinformaticians. EXCLI J 12:831
32. Daina A, Michielin O, Zoete V (2017) SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. Sci Rep 7:42717

33. Ismail SY, Uzairu A, Sagagi B, Sabiu M (2018) In silico molecular docking and pharmacokinetic study of selected phytochemicals with estrogen and progesterone receptors as anticancer agent for breast cancer. *5*:1337–1350
34. Todeschini R, Consonni V (2009) Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references (Vol. 41). Wiley

Publisher's Note