Future Journal of
Pharmaceutical Sciences

RESEARCH

Open Access

# Computational evaluation of some compounds as potential anti-breast cancer agents

Momohjimoh Ovaku Idris*, Stephen Eyije Abechi and Gideon Adamu Shallangwa

## Abstract

**Background:** The emergence of high resistance and toxicity of the existing anti-breast cancer drugs have demanded the need to design new drugs with improved activities against breast cancer. A computational technique incorporating quantitative structure–activity relationship and virtual template-based design was carried out to evaluate thirty-four compounds from derivatives of thiophene, pyrimidine, coumarin, pyrazole and pyridine with anti-breast cancer activities. The chemical structures of the compounds were drawn with chem draw v.12.0.2 and they were optimized using Spartan 14 software. The molecular descriptors were calculated with the aid of PaDel descriptor software. The dataset was curated and then divided into training and test set that was used to generate and validate the model.

**Results:** The first out of the four models generated was chosen as the paramount model with statistical validations of $R^2 = 0.9847$, $R^2_{adj} = 0.9814$, $Q^2_{cv} = 0.9763$, min expt. error for non-significant LOF (95%) = 0.0679, an external validation $R^2_{test}$ of 0.8240 and coefficient of Y-randomization ($cR^2_p$) = 0.8200, which confirm the robustness of the model.

**Conclusions:** The high predictive power of the generated model describes the models' reliability and the designed compounds pointed out compound 2 with $pGI_{50} = 4.2504$ as the best designed compound to inhibit breast cancer, compared to its co-designed compounds and the template. The results of this research provide vital information to the pharmaceutical chemists and the pharmacologist in the course of developing new breast cancer drugs.

**Keywords:** Anti-breast cancer, Dataset, Dataset division, Model validations, Template, Design compounds

## Background

Cancer is a word used to describe the unusual growth of the cells leading to one of the most dangerous health problems for humans all over the world [1]. Despite the availability of improved drugs targeting cancer therapies, the worldwide cancer burden is expected to increase to 19.3 million new cancer cases, and nearly 10 million cancer deaths were observed in the year 2020 [2].

Breast cancer is the most common cancer among women all over the world and impermanence from breast cancer is commonly due to tumour metastasis [3]. It constitutes a major public health issue globally, with over 1 million new cases diagnosed annually; resulting in over 400,000 annual deaths and about 4.4 million women living with the disease [4]. The mortality rate of breast cancer among Nigerian women is about 16% [5].

Amino-thiophene derivatives were known to be one of the most important groups of heterocyclic compounds with a wide spectrum of biological activities such as antitumor [6] anti-mitotic [7] and antiviral [8]. Furthermore, thieno [2, 3-d] pyrimidine derivatives show anti-proliferative activity [9] while pyrazole derivatives have a specific effect with favourable antitumor activity [10]. Coumarin scaffold turn out to be an attractive subject due to their broad spectrum of pharmacological activities, its derivative is extensively explored for anticancer activities as it possesses minimum side effect along with multi-drug

*Correspondence: eedrismj@gmail.com
Department of Chemistry, Ahmadu Bello University, P.M.B. 1044, Zaria, Nigeria

reversal activity [11]. Most pyridine derivatives had been synthesized as potentially biologically active compounds and had a multitude of pharmacological characteristics, in particular, anti-cancer activity [12–14].

Quantitative Structure Activity Relationship (QSAR) is one of the commonly used computational method for predicting the activities/properties of molecules in drug design as it saves time and lesser cost [15]. Generating a good QSAR model depends on factors such as: the quality of biological data, the choice of descriptors, variable selection, statistical methods and validations.

The aim of this research is to develop a good QSAR model for predicting the activity of some selected compounds against breast cancer and also design new compounds with better activities against breast cancer.

## Methods

### Data collection

The dataset used in this work was collected from the literature [16] and were reported as fifty percent growth inhibition ($GI_{50}$) concentrations in (mmol $L^{-1}$). These reported inhibitory activities were converted to logarithm scale to have a well-defined range with the help of Eq. (1) shown below.

$$pGI_{50} = -\log_{10}(GI_{50} \times 10^{-3}) \tag{1}$$

### Compounds sketching, optimization and descriptors calculations

The two-dimensional structure (2D) of the compounds were sketched using ChemDraw software version 12.0.2 [17], they were imported into Spartan 14 V.1.1.4 software to obtain the optimized three-dimensional spatial conformer (3D) at Density Functional Theory (DFT) level applying B3LYP 6-31G$^{*}$ basis set [18]. The optimized compounds in Spartan format were converted to SD file format and later imported into the PaDEL software to calculate the models' descriptors.

### Dataset normalization and pre-treatment

To give the descriptors equal chance of occurrence, the compounds were normalized using Eq. (2), [19]. The normalized data were pre-treated using the data pretreatment software obtained from Drug Theoretic and Cheminformatics Laboratory (DTC Lab) to remove all empty columns and some useless descriptors [20].

$$X = \frac{X_i - X_{min}}{X_{max} - X_{min}} \tag{2}$$

where $X_i$ in the equation is the value of each descriptor for a given molecule and $X_{max}$ and $X_{min}$ are the maximum and minimum values for each column of descriptors $X$ respectively.

### Model generation and validation

In other to generate a good QSAR model, the pretreated dataset was divided into training and test set in the ratio 7:3 by the means of data division software of DTC Lab [20]. The model was built using the training set, employing GFA-MLR method from the material studio. The test set was then used to validate the built model [21]. The suitability notch of the generated model was assessed using the lack of fit (LOF) [22], as in Eq. (3).

$$LOF = \frac{SEE}{\left(1 - \frac{C+d*P}{M}\right)^2} \tag{3}$$

SEE being the Standard Error of Estimation, $C$ is the number of terms in the model, $d$ is a user-defined smoothing parameter, $P$ is the total number of descriptors in the model and $M$ is the number of training dataset. SEE can be expressed as:

$$SEE = \sqrt{\frac{\left(Y_{exp} - Y_{pre}\right)^2}{N - P - 1}} \tag{4}$$

where $Y_{exp}$ and $Y_{pre}$ are the experimental activity and the predicted activity in the training set respectively [22].

The squared correlation coefficient ($R^2$) is a validation test used to match the predicted and experimental activities. The model would be considered robust with an $R^2$ value close to 1. $R^2$ is expressed as:

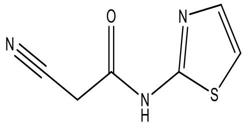
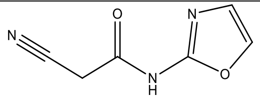
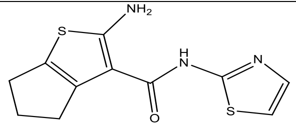
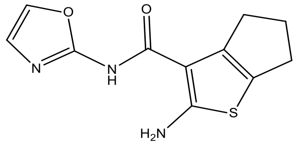
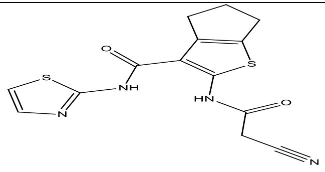
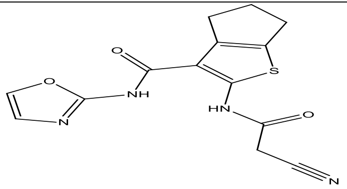
$$R^2 = 1 - \left[\frac{\sum(Y_{exp} - Y_{pred})^2}{\sum(Y_{exp} - \overline{Y}_{training})^2}\right] \tag{5}$$

where $Y_{exp}$, $Y_{pred}$ and $\overline{Y}_{training}$, were respectively the experimental activity, the predicted activity, and the mean experimental activity of the samples in the training set. The validity of the model cannot be based on $R^2$ only, therefore an adjustment in the $R^2$ would give a more reliable model. The adjusted $R^2$ is givens by:

$$R^2_{adj} = \frac{R^2 - d(n-1)}{n - P + 1} \tag{6}$$

where $d$ is the number of descriptors in the model and $n$ is the number of training set compounds.The predictive power of the model is usually determined by the

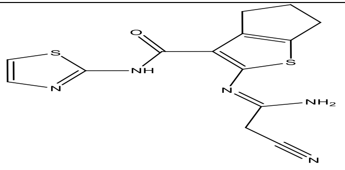
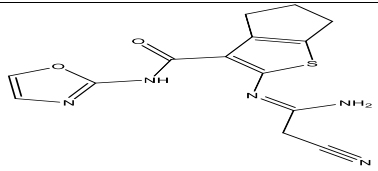
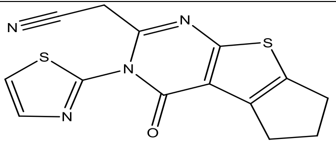
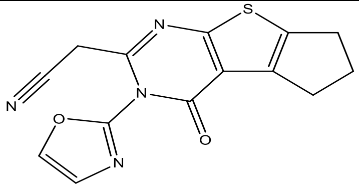
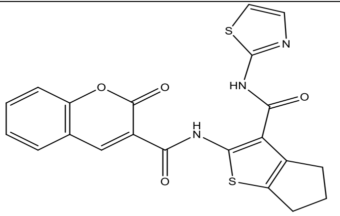
**Table 1** 2D structure and names of the dataset and their 50% growth inhibitory activities in mmol $L^{-1}$

| S/N | Names and structure of the compounds | $GI_{50}$ | $pGI_{50}$ |
|---|---|---|---|
| C1 | <br>2-cyano-*N*-(thiazol-2-yl)acetamide | 23.7 | 1.6253 |
| C2 | <br>2-cyano-*N*-(oxazol-2-yl)acetamide | 27.1 | 1.5670 |
| C3 | <br>2-amino-*N*-(thiazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 44.6 | 1.3507 |
| C4 | <br>2-amino-*N*-(oxazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 41.9 | 1.3778 |
| C5 | <br>2-(2-cyanoacetamido)-N-(thiazol-2-yl)-5,6-dihydro-4H-cyclopenta[b]thiophene-3-carboxamide | 51.3 | 1.2899 |
| C6 | <br>2-(2-cyanoacetamido)-*N*-(oxazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 44.2 | 1.3546 |

Cross-validation ($Q^2_{\text{cv}}$) and the external validation test as expressed in Eqs. (7) and (8) respectively.

$$Q^2_{\text{cv}} = 1 - \left[ \frac{\sum (Y_{\text{exp}} - Y_{\text{pred}})^2}{\sum (Y_{\text{exp}} - \overline{Y}_{\text{training}})^2} \right] \tag{7}$$

**Table 1** (continued)

| | | | |
|---|---|---|---|
| C7 | <br>(*Z*)-2-((1-amino-2-cyanoethylidene)amino)-*N*-(thiazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 48.0 | 1.3188 |
| C8 | <br>(*Z*)-2-((1-amino-2-cyanoethylidene)amino)-*N*-(oxazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 41.8 | 1.3788 |
| C9 | <br>2-(4-oxo-3-(thiazol-2-yl)-4,5,6,7-tetrahydro-3*H*-cyclopenta[4,5]thieno[2,3-*d*]pyrimidin-2-yl)acetonitrile | 0.1 | 4.0000 |
| C10 | <br>2-(3-(oxazol-2-yl)-4-oxo-4,5,6,7-tetrahydro-3*H*-cyclopenta[4,5]thieno[2,3-*d*]pyrimidin-2-yl)acetonitrile | 0.09 | 4.0458 |
| C11 | <br>2-oxo-*N*-(3-(thiazol-2-ylcarbamoyl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophen-2-yl)-2*H*-chromene-3-carboxamide | 37.9 | 1.4214 |

$$R^2_{\text{test}} = 1 - \left[\frac{\sum(Y_{\text{pred}_{\text{test}}} - Y_{\text{exp}_{\text{test}}})^2}{\sum(Y_{\text{pred}_{\text{test}}} - \overline{Y}_{\text{training}})^2}\right] \quad (8)$$

where $Y_{\text{pred}_{\text{test}}}$ is the predicted activity, $Y_{\text{exp}_{\text{test}}}$ is the experimental activity of the test set and $\overline{Y}_{\text{training}}$ is the mean activity of the training set [21].

### Y-randomization

Y-randomization is an external validation test performed to generate a new model from the bogus dataset so as to improve the models' efficacy. For a good model, the randomized squared correlation coefficient ($cR^2_p$) must be greater than 0.5, and is expressed as:

**Table 1** (continued)

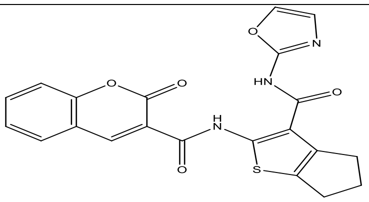
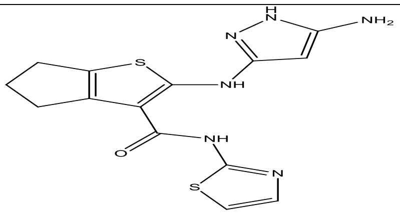
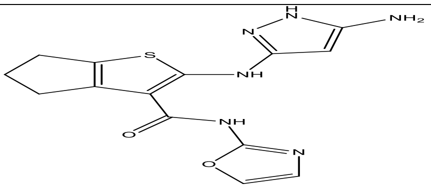
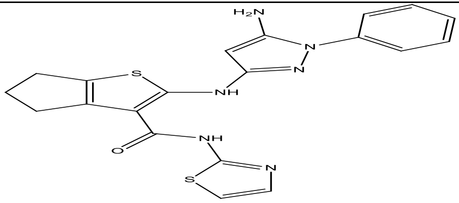
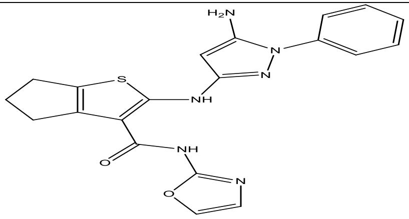
| | | | |
|---|---|---|---|
| C12 | <br>*N*-(3-(oxazol-2-ylcarbamoyl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophen-2-yl)-2-oxo-2*H*-chromene-3-carboxamide | 39.3 | 1.4056 |
| C13 | <br>2-((5-amino-1*H*-pyrazol-3-yl)amino)-*N*-(thiazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 40.3 | 1.3947 |
| C14 | <br>2-((5-amino-1*H*-pyrazol-3-yl)amino)-*N*-(oxazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 33.1 | 1.4802 |
| C15 | <br>2-((5-amino-1-phenyl-1*H*-pyrazol-3-yl)amino)-*N*-(thiazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 29.0 | 1.5376 |
| C16 | <br>2-((5-amino-1-phenyl-1*H*-pyrazol-3-yl)amino)-*N*-(oxazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 35.9 | 1.4449 |

**Table 1** (continued)

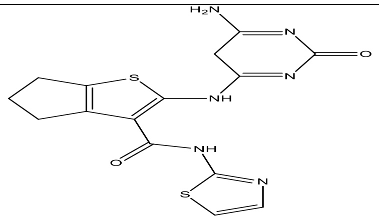
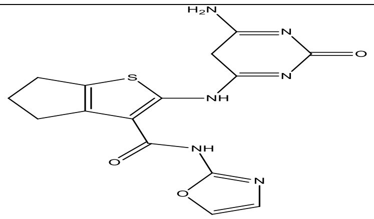
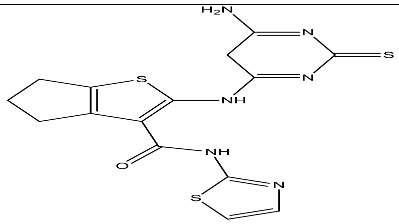
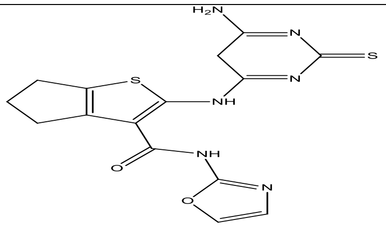
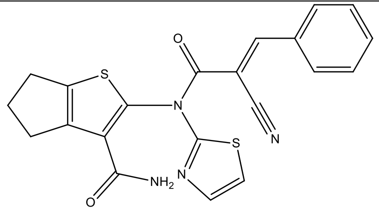
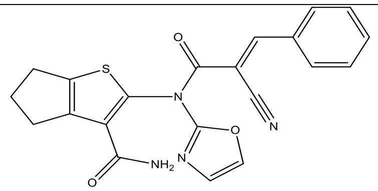
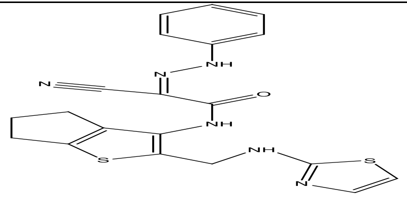
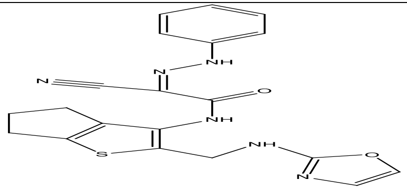
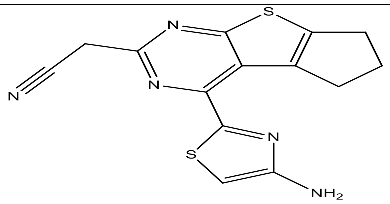
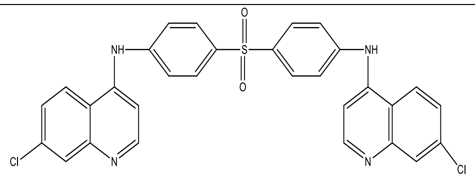
| | | | |
|---|---|---|---|
| C17 | <br>2-((6-amino-2-oxo-2,5-dihydropyrimidin-4-yl)amino)-*N*-(thiazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 44.2 | 1.3546 |
| C18 | <br>2-((6-amino-2-oxo-2,5-dihydropyrimidin-4-yl)amino)-*N*-(oxazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 34.9 | 1.4572 |
| C19 | <br>2-((6-amino-2-thioxo-2,5-dihydropyrimidin-4-yl)amino)-*N*-(thiazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 36.7 | 1.4353 |
| C20 | <br>2-((6-amino-2-thioxo-2,5-dihydropyrimidin-4-yl)amino)-*N*-(oxazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 41.2 | 1.3851 |
| C21 | <br>(*E*)-2-(2-cyano-3-phenyl-*N*-(thiazol-2-yl)acrylamido)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 42.7 | 1.3696 |

**Table 1** (continued)

| | | | |
|---|---|---|---|
| C22 | <br>(*E*)-2-(2-cyano-*N*-(oxazol-2-yl)-3-phenylacrylamido)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 39.2 | 1.4067 |
| C23 | <br>(*Z*)-2-oxo-*N*'-phenyl-2-((2-((thiazol-2-ylamino)methyl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophen-3-yl)amino)acetohydrazonoyl cyanide | 45.2 | 1.3449 |
| C24 | <br>(*Z*)-2-((2-((oxazol-2-ylamino)methyl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophen-3-yl)amino)-2-oxo-*N*'-phenylacetohydrazonoyl cyanide | 38.4 | 1.4157 |
| C25 | <br>2-(4-(4-aminooxazol-2-yl)-6,7-dihydro-5*H*-cyclopenta[4,5]thieno[2,3-*d*]pyrimidin-2-yl)acetonitrile | 21.7 | 1.6635 |
| C26 | <br>2-((4,6-diamino-5-cyanopyridin-2-yl)amino)-*N*-(thiazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 21.2 | 1.6737 |

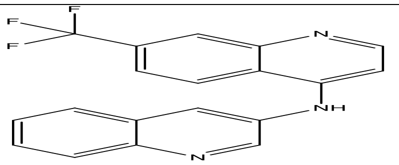
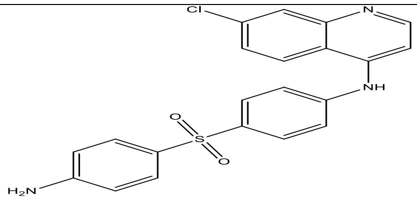
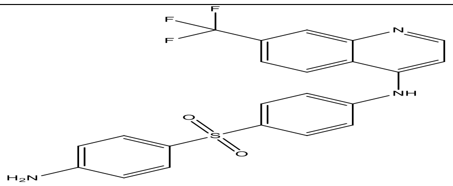
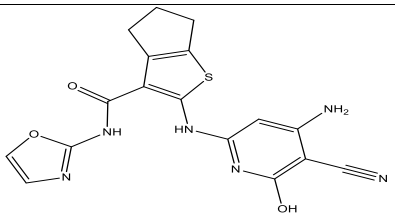
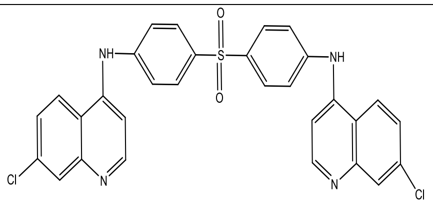
$$\mathrm{cR_p^2} = R[R^2 - (R_\mathrm{r})^2]^2 \tag{9}$$

where $\mathrm{cR_p^2}$ is the Y-randomization coefficient and $R_\mathrm{r}$ is the average '*R*' of random models [19].

## Applicability domain (AD)

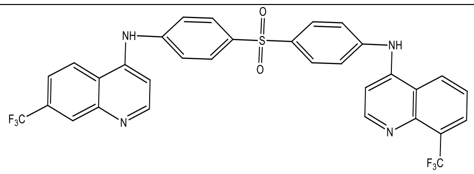
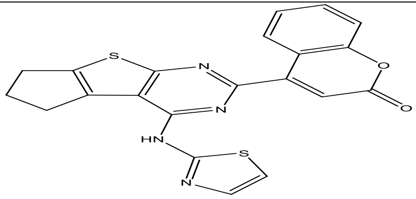
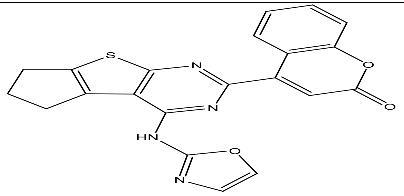
Applicability domain is a theoretical region of the chemical space that is defined by the model descriptors, model response and nature of the training set. The leverage

**Table 1** (continued)

| | | | |
|---|---|---|---|
| C27 | <br>2-((4,6-diamino-5-cyanopyridin-2-yl)amino)-*N*-(thiazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 16.2 | 1.7905 |
| C28 | <br>2-((4,6-diamino-5-cyanopyridin-2-yl)amino)-*N*-(oxazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 25.2 | 1.5986 |
| C29 | <br>2-((4-amino-5-cyano-6-hydroxypyridin-2-yl)amino)-*N*-(thiazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 19.9 | 1.7011 |
| C30 | <br>2-((4-amino-5-cyano-6-hydroxypyridin-2-yl)amino)-*N*-(oxazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 20.2 | 1.6946 |
| C31 | <br>2-((4-amino-5-cyano-6-hydroxypyridin-2-yl)amino)-*N*-(oxazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 41.1 | 1.3862 |

approach was employed to measure the data within the AD [23], any dataset that lies outside the AD would be treated as an outlier. Equation (10) is normally used to calculate the AD.

**Table 1** (continued)

| | | | |
|---|---|---|---|
| C32 |  NH<br>O<br>S<br>O<br>NH<br>F<sub>3</sub>C<br>N<br>N<br>F<sub>3</sub>C<br>(*Z*)-2-((amino(2-oxo-2*H*-chromen-3-yl)methylene)amino)-*N*-(oxazol-2-yl)-5,6-dihydro-4*H*-cyclopenta[*b*]thiophene-3-carboxamide | 39.7 | 1.4012 |
| C33 |  <br>4-(4-(thiazol-2-ylamino)-6,7-dihydro-5*H*-cyclopenta[4,5]thieno[2,3-*d*]pyrimidin-2-yl)-2*H*-chromen-2-one | 3.1 | 2.5086 |
| C34 |  <br>4-(4-(oxazol-2-ylamino)-6,7-dihydro-5*H*-cyclopenta[4,5]thieno[2,3-*d*]pyrimidin-2-yl)-2*H*-chromen-2-one | 6.2 | 2.2076 |

**C1–C34**: compounds ranging from 1 to 34

$$l_i = X_i(X^{\mathrm{T}}X)^{-1}X_i^{\mathrm{T}} \tag{10}$$

where $l_i$ is the leverage of each compound, $X_i$ is the descriptor row-vector of the query compound $i$, and $X$ is the $(m \times n)$ descriptor matrix of the training set compounds used in building the model. The critical value ($l^*$) is defined by Eq. (11).

$$l^* = 3\frac{p+1}{n} \tag{11}$$

where $p$ is the number of descriptors in the model and $n$ is the number of objects used to develop the model.

**Table 2** Standard validation parameters for a good QSAR model

| Validation parameters | Meaning | Values |
|---|---|---|
| $R^2$ | Coefficient of determination | $\geq 0.6$ |
| $P_{95\%}$ | Confidence interval at 95% confidence level | $< 0.06$ |
| $Q^2_{\mathrm{cv}}$ | Cross-validation coefficient | $> 0.5$ |
| $R^2 - Q^2_{\mathrm{cv}}$ | Difference between $R^2$ and $Q^2_{cv}$ | $\leq 0.3$ |
| $N_{\text{ext. test set.}}$ | Minimum number of external test sets | $\geq 5$ |
| $R^2_{\mathrm{test}}$ | Coefficient of determination for external test set | $\geq 0.6$ |
| $\mathrm{c}R^2_{\mathrm{p}}$ | Coefficient of determination for *Y*-randomization | $> 0.5$ |

## Mean effect (ME) and variance inflation factor (VIF)

The mean effect is used to elucidate the comparative importance of each descriptor in the model while the VIF is used to determine the linearity between the descriptors in the model. VIF value of 1 show no linearity among the descriptors and value above 10 indicates a bad model. The ME and VIF are respectively calculated using Eqs. (12) and (13).

$$\mathrm{ME} = \frac{B_j \sum_i^n D_j}{\sum_j^m \left(B_j \sum_i^n D_j\right)} \tag{12}$$

where $B_j$ is the coefficient of the descriptor $j$ in the model, $D_j$ is the value of each descriptor in the data matrix for each of the training set data, $m$ and $n$ are respectively the

**Table 3** QSAR model validations values

| Validation parameters | Model 1 | Model 2 | Model 3 | Model 4 |
|---|---|---|---|---|
| Friedman LOF | 0.0330700 | 0.034214 | 0.035402 | 0.036218 |
| $R^2$ | 0.984748 | 0.984220 | 0.983673 | 0.983296 |
| $R^2_{\mathrm{adj}}$ | 0.981359 | 0.980714 | 0.980044 | 0.979584 |
| $Q^2_{\mathrm{cv}}$ | 0.976276 | 0.973887 | 0.945063 | 0.970229 |
| $R^2 - Q^2_{\mathrm{cv}}$ | 0.008472 | 0.010333 | 0.03861 | 0.013067 |
| $R^2_{\mathrm{ext}}$ | 0.8240 | | | |
| $cR^2_p$ | 0.8200 | | | |
| Significance-of-regression *F*-value | 290.545886 | 280.675457 | 271.112245 | 264.901430 |
| Min expt. error for non-significant LOF (95%) | 0.06792300 | 0.06908900 | 0.07027700 | 0.07108300 |

**Table 4** Y-randomization

| Model | *R* | $R^2$ | $Q^2$ |
|---|---|---|---|
| Original | 0.9436 | 0.8904 | 0.7631 |
| Model 1 | 0.3428 | 0.1175 | − 0.2877 |
| Model 2 | 0.2754 | 0.0758 | − 0.5373 |
| Model 3 | 0.1695 | 0.0287 | − 0.4698 |
| Model 4 | 0.1420 | 0.0202 | − 0.2828 |
| Model 5 | 0.5947 | 0.3536 | − 0.1640 |
| Model 6 | 0.2385 | 0.0569 | − 0.1526 |
| Model 7 | 0.3906 | 0.1525 | − 0.1264 |
| Model 8 | 0.1609 | 0.0259 | − 0.2854 |
| Model 9 | 0.5167 | 0.2670 | − 0.8873 |
| Model 10 | 0.8452 | 0.7144 | 0.3324 |
| Average randomized model | | | |
| Average *R*: | 0.3676 | | |
| Average $R^2$: | 0.1813 | | |
| Average $Q^2$: | − 0.2861 | | |
| $cR^2_p$: | 0.8200 | | |

number of descriptors that appears in the model and the number of molecules in the training set

$$\mathrm{VIF} = \frac{1}{1 - R^2} \tag{13}$$

where $R^2$ is the multiple regression correlation coefficient between the variables in the model [24].

## Molecular design

An In-silico approach of template-based design was employed to design new compounds with enhance activity against breast cancer. This method has been hired frequently to screen and modelled compounds with better-quality activity by relating the experimental activities of the compounds with their structures [25]. Henceforth, compound with the highest activity would be defined as the template to design new compounds with enhanced activities.

# Results

All the tables and figures that describes the outcome of the built model and the designed compounds are presented in this section.

# Discussion

All the thirty-four compounds used in this study were first sketched by ChemDraw to obtain the 2D structures, they were imported to the spartan 14 software to obtain their 3D optimised structures. The optimized dataset was normalized, pre-treated and the molecular descriptors were calculated with the help of PaDEL descriptor software. A large number of 1874 of molecular descriptors

**Table 5** Correlation matrix, VIF and mean effect (ME) for the QSAR model descriptors

| Descriptors | Inter-correlation | | | | VIF | Mean effect |
|---|---|---|---|---|---|---|
| | GATS8c | maxHBd | TDB10p | RNCS | | |
| GATS8c | 1 | | | | 1.2216 | − 0.2514 |
| MaxHBd | 0.2208 | 1 | | | 1.1071 | 0.8382 |
| DB10p | − 0.2053 | 0.0884 | 1 | | 1.373 | 0.2425 |
| RNCS | 0.3934 | 0.1704 | − 0.4878 | 1 | 1.5357 | 0.1706 |

**Table 6** Descriptive analysis

| Statistical analysis | Activity | |
|---|---|---|
| | Training dataset | Test dataset |
| Number of compounds | 23 | 11 |
| Confidence level (95%) | 0.2727 | 0.4963 |
| Mean | 1.6531 | 1.8125 |
| Median | 1.4067 | 1.6635 |
| Maximum | 4.0458 | 4 |
| Minimum | 1.2899 | 1.3546 |
| Kurtosis | 9.4055 | 9.9844 |
| Range | 2.7559 | 2.6454 |
| Skewness | 2.9392 | 3.0973 |
| Standard deviation | 0.6307 | 0.7388 |
| Sample variance | 0.3978 | 0.5458 |

**Table 7** Details of the descriptors used in the model

| S/N | Descriptors | Descriptor type | Number | Class |
|---|---|---|---|---|
| 1 | GATS8c | Autocorrelation | 346 | 2D |
| 2 | maxHBd | Atom type electro-topological state | 489 | 2D |
| 3 | TDB10p | 3D autocorrelation | 80 | 3D |
| 4 | RNCS | Charged partial surface area | 29 | 3D |

that are responsible for encrypting the important features of the structures were calculated.

The 2D structures and activities of the studied compounds were presented in Table 1. The Genetic Function Approximation (GFA), was used to generate four models, the first model out of the four models was selected as the optimum model since it best agrees with the minimum criteria for generating good QSAR model, reported in Table 2. Table 3 display the validation parameters for the generated models. Table 4 present the Y-randomization test used to affirm the strength of a model. This test was carried out on the training set by keeping the independent variable constant and randomizing the dependent variables. The low values of $R$, $R^2$ and $Q^2$ indicate the robustness of the generated model and the coefficient of Y-randomization ($cR_p^2 = 0.8200$) confirmed the generated model was not gotten by chance.

Table 5 displays the correlation matrix, VIF and the ME of the four descriptors used to build the models. The low value of the Pearson's correlation indicates that there is no significant connection between the descriptors, this means that each descriptor gives different information that influenced the model. The relative importance of each of the descriptor in the model was measured with the low value of the Variance

**Table 8** Residual values for both training and test dataset

| Datasets | $pGI_{50}$ | Predicted $pGI_{50}$ | Residuals |
|---|---|---|---|
| *C1 | 1.6253 | 1.2498 | 0.3755 |
| C2 | 1.5670 | 1.5673 | −0.0003 |
| C3 | 2.5229 | 2.4293 | 0.0936 |
| C4 | 1.3778 | 1.4520 | −0.0742 |
| C5 | 1.2899 | 1.3575 | −0.0676 |
| *C6 | 1.3546 | 1.5116 | −0.1570 |
| C7 | 1.3188 | 1.3985 | −0.0797 |
| *C8 | 1.3788 | 1.3171 | 0.0617 |
| *C9 | 4.0000 | 3.3794 | 0.6206 |
| C10 | 4.0458 | 4.0494 | −0.0037 |
| C11 | 1.4214 | 1.5261 | −0.1047 |
| C12 | 1.4056 | 1.3025 | 0.1031 |
| C13 | 1.3947 | 1.3116 | 0.0831 |
| C14 | 1.4802 | 1.4158 | 0.0644 |
| C15 | 1.5376 | 1.4644 | 0.0732 |
| C16 | 1.4450 | 1.4122 | 0.0328 |
| C17 | 1.3546 | 1.3360 | 0.0186 |
| *C18 | 1.4572 | 1.4106 | 0.0466 |
| C19 | 1.4353 | 1.3687 | 0.0666 |
| C20 | 1.3851 | 1.4706 | −0.0855 |
| C21 | 1.3696 | 1.2988 | 0.0708 |
| C22 | 1.4067 | 1.5193 | −0.1126 |
| C23 | 1.3449 | 1.2428 | 0.1021 |
| C24 | 1.4157 | 1.4835 | 0.06783 |
| *C25 | 1.6635 | 1.8581 | −0.1946 |
| *C26 | 1.6737 | 1.5049 | 0.1688 |
| *C27 | 1.7904 | 1.2847 | 0.5057 |
| *C28 | 1.5988 | 1.3909 | 0.2079 |
| *C29 | 1.7011 | 1.0303 | 0.6708 |
| *C30 | 1.6946 | 1.1253 | 0.5693 |
| C31 | 1.3862 | 1.4618 | −0.0756 |
| C32 | 1.4012 | 1.4220 | −0.0208 |
| C33 | 2.5086 | 2.4212 | 0.0874 |
| C34 | 2.2076 | 2.3106 | −0.1030 |

**NB**: $^*$**C** → Test set compounds

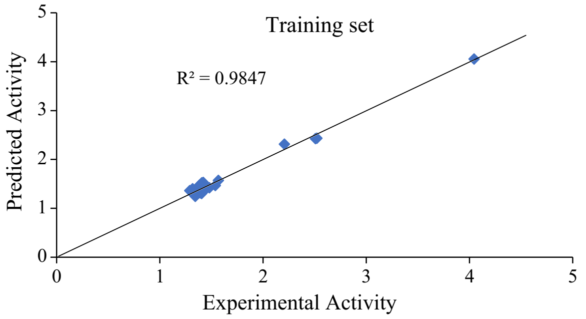


**Fig. 1** Plot of experimental activity against predicted activity of the training set

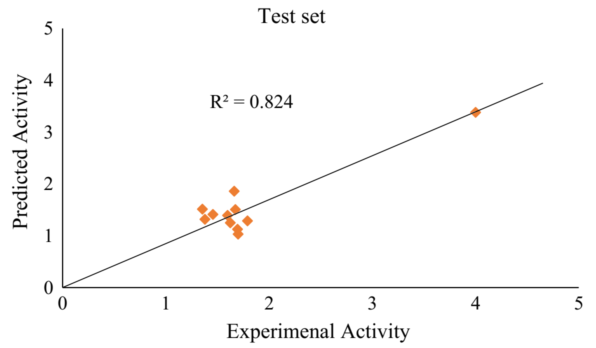


**Fig. 2** Plot of experimental activity against predicted activity of the test set

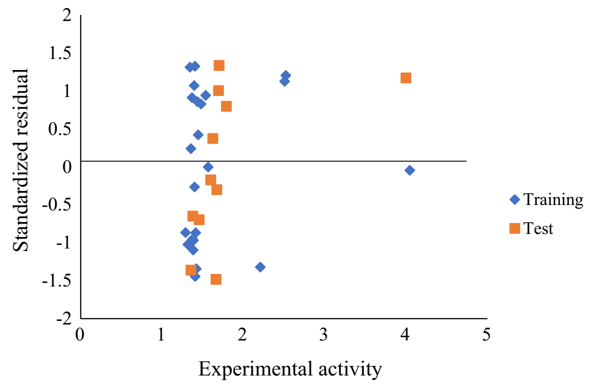


**Fig. 3** The plot of standardized residual against experimental activity

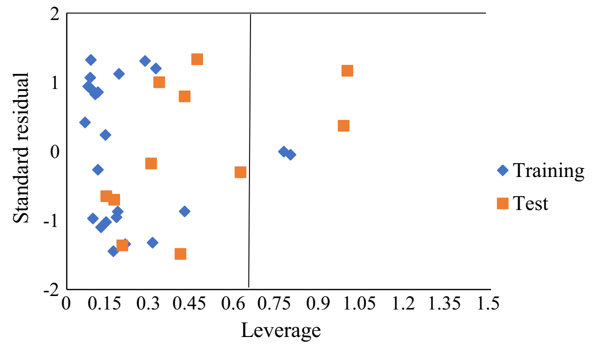


**Fig. 4** The Williams plot

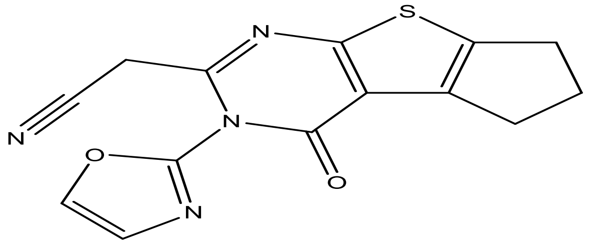


**Figure5** Template compound (C10), 2-(3-(oxazol-2-yl)-4-oxo-4, 5, 6, 7-tetrahydro-3H-cyclopent [4, 5] thieno[2,3-d] pyrimidin-2-yl) acetonitrile for the designed compounds

Inflation Factor (VIF) and since the VIF value were all less than 2, henceforth, the descriptors in the model were rightfully selected and the model is therefore said to be statistically satisfactory [24]. Meanwhile descriptor **MaxHBd** with highest positive ME value indicates its prominence in the models' activity, as such, the descriptor was made the focal point when designing new enhanced compounds. The descriptor (**MaxHBd**), means Maximum E-States for (strong) Hydrogen Bond donors.

Descriptive analysis was carried out to back up the evidence that the dataset was well divided into a new set (training set and test set). Table 6 present the maximum, minimum and standard deviation values for both training and test sets were very close suggesting no significant difference in them, as a result, we deduce that the training set is extrapolative within the test set, this confirm the fit of the Kennard and stone method employed in the data division.

Table 7 present the details of the descriptors used to build the model. The first two descriptors were 2D and the last two being 3D. The equations generated from the material studio software displayed below, indicates Eq. (1) as the best model when compared to the standard validation parameters for generating a good QASR model in Table 2.

## Model 1

$$pGI_{50} = 0.709363893 * \mathbf{GATS8c} - 4.252846824 * \mathbf{maxHBd} - 0.063150018 * \mathbf{TDB10p} - 0.153565552 * \mathbf{RNCS} + 4.211504042;$$

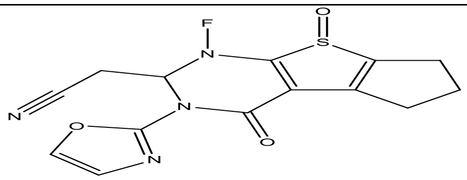
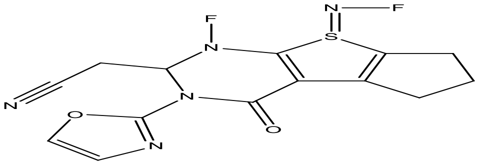
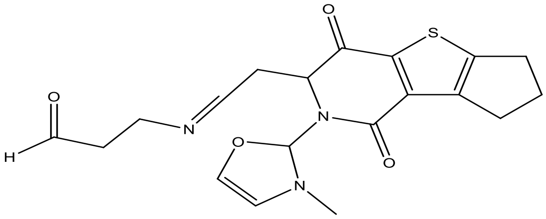
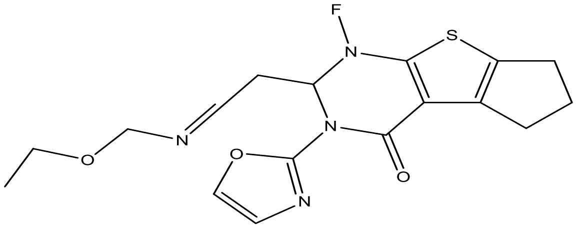
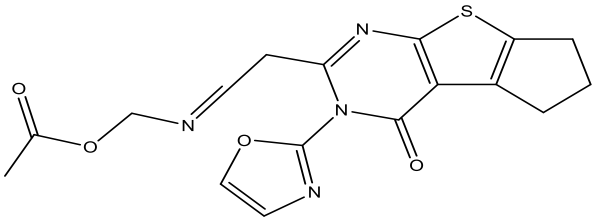
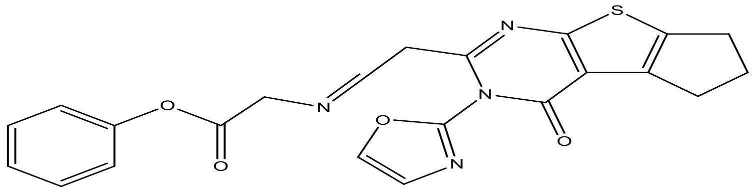
## Model 2

$$pGI_{50} = 0.772397640 * \mathbf{GATS8c} - 4.191643628 * \mathbf{maxHBd} - 0.000418849 * \mathbf{TDB10v} - 0.153328698 * \mathbf{RNCS} + 4.006452472;$$

## Model 3

$$pGI_{50} = 1.245271529 * \mathbf{MATS7c} - 0.826297651 * \mathbf{SpMax5\_Bhi} - 3.598436568 * \mathbf{maxHBd} - 0.135122003 * \mathbf{RNCS} + 6.680738880;$$

**Table 9** 2D Structures and 50% Growth Inhibitory activities ($pGI_{50}$) of the design compounds

| S/N | Structure of the design compounds | $pGI_{50}$ ($mmolL^{-1}$) |
|---|---|---|
| D1 |  | 4.2118 |
| *D2 |  | 4.2504 |
| D3 |  | 3.9945 |
| D4 |  | 3.8080 |
| D5 |  | 3.8154 |
| D6 |  | 4.1688 |

**NB:** $^{*}$**D2** = Best design compound

## Model 4

$pGI_{50} = 0.723219873 * \mathbf{GATS8c} - 4.266084558 * \mathbf{maxHBd} - 0.061174372 * \mathbf{Kier1} - 0.155935615 * \mathbf{RNCS} + 4.726811020$.

The difference between the predicted activity and the reported activity is the residual activity, which is presented in Table 8. The low residual values indicate that the predicted activities lie within the experimental activities, accounting for the high predicting power of the model. Figure 1 and 2 below shows the graphical plot of experimental activity against the predicted activity for both training and test set respectively, the $R^2$ value of the two plots are satisfactory when compared to the recommended $R^2$ value of a good QSAR model reported in Table 2. The plot of standardize residual versus experimental activity in Fig. 3, was used to check for any systematic error in the built model, it was found that the built model was free of systematic error since all it standardizes value lies within ± 2 unit. Figure 4 shows the Williams plot, the plot help to determine compounds that are either influential or outliers. Four compounds were found to be outliers because their leverage values were greater that the critical leverage ($l^* = 0.6$) and those compounds shall not be considered while designing a new anti-breast cancer agent.

In other to design more potent anti-breast cancer compounds, compound 10 (Fig. 5) with the highest reported activity (4.0458) was endorsed as the template. The most influential descriptor **maxHBd** (maximum E-state for Hydrogen bond donor), with mean effect of 0.8382 was investigated. To raise the hydrogen bond donor, H-bond acceptor and strong electronegative atoms (F, O and N) were attached to the appropriate positions, which lead to the design of six new compounds with enhanced 50% growth inhibitory activity as displayed in Table 9.

## Conclusion

This research has effectively built a good QSAR model with high predictive power, using the descriptors **maxHBd**, **GATS8c**, **TDB10p** and **RNCS**. The Williams plot, outlined four compounds (outliers) that should not be considered for further computational study. The validation parameters used to generate the model as discussed above all passed the minimum recommendation for building a valid QSAR model. Descriptor **maxHBd** with positive mean effect value of 0.8382 was found to mostly influence the optimum model, and was chosen as the template that was then used to design six new compounds with better inhibitory activities. Three out of the six designed compounds were found to have $pIC_{50}$ value (4.2118, 4.1688 and 4.2504) greater than the template and the rest of the design compounds. Conclusively, the research aim was achieved and the results of this work would serve as first-hand information to the pharmaceutical chemist, pharmacist and pharmacologist in the course of producing new drug against breast cancer.

### Abbreviations

QSAR: Quantitative structure–activity relationships; VIF: Variance inflation factor; ME: Mean effect; DTC Lab: Drug Theoretic and Cheminformatics Laboratory; $GI_{50}$: 50% Growth inhibition; DFT: Density functional theory.

### Acknowledgements

The authors really appreciate the scholarly guidance of Mr. Adeniji Elijah Shola of Ahmadu Bello University (ABU) Zaria-Nigeria and Mr. Abdullahi Mustapha of Kaduna state University (KASU) Nigeria.

### Authors' contributions

GAS collected the Data and sketched all through to optimizations. SEA preprosed the data, built and validates the model. MOI analysed and discussed the result and also designed the new compounds with improved activity. All authors have read and approved the manuscript.

### Funding

No any form of funding applicable to this work.

### Availability of data and materials

The data used in this work was collected from the work of Albratty, M, El-Sharkawy, K. A., and Alam, S., (2017), Synthesis and Antitumor Activity of Some Novel Thiophene, Pyrimidine, Coumarine, Pyrazole and Pyridine Derivatives. Acta Pharm. 67, 15–33, https://doi.org/10.1515/acph-2017-0004. All the data and materials generated during the current study are included in this published article and are available upon request.

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interest

The authors declare that there was no competing interest regarding this article.

Received: 21 May 2021 Accepted: 9 August 2021
Published online: 24 August 2021

### References

1. Idris MO, AbechiShallagwa SEGA, Uzairu A (2020) QSAR and molecular docking studies of novel thiophene, pyrimidine, coumarin, pyrazole and pyridine derivatives as potential anti-breast cancer agent. Turk Comput Theo Chem (TC&TC) 4(1):12–23
2. Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M et al (2021) Global cancer observatory: cancer today. International Agency for Research on Cancer, Lyon. https://gco.iarc.fr/today. Accessed Feb 2021
3. Jemal A, Siegel R, Ward E, Murray T, Xu J, Thun MJ (2007) Cancer statistics. CA Cancer J Clin 57:43–66
4. GLOBOCAN (2018) Latest Global Cancer data: cancer Burden Rises to 18.1 million new cases and 9.6 million cancer deaths in 2020. IARC Global Cancer Observatory
5. Okobia MN, Bunker CH, Okonofua FE, Osime U (2006) Knowledge, attitude and practice of Nigerian women towards breast cancer: a cross-sectional study. World J Surg Oncol 4:11–15
6. Romagnoli R, Baraldi PG, Lopez-Cara C, Salvador MK, Preti D, Tabrizi MA, Balzarini V, Nussbaumer P, Bassetto M, Brancale A, Fu XH, Gao Y, Li J, Zhang SZ, Hamel E, Bortolozzi R, Basso G, Viola G (2014) Design synthesis and biological evaluation of 3, 5-disubstituted 2-aminothiophene derivatives as a novel class of antitumor agents. Bioorg Med Chem 22:5097–5109. https://doi.org/10.1016/j.bmc.2013.12.030
7. Romagnoli R, Baraldi PG, Carrion MD, Cara CL, Perti D, Fruttarolo F, Pavani MG, Tabrizi MA, Tolomio M, Grimaudo S, Di-Cristina A, Balzarini J, Hadfield JA, Bracale A, Hamel E (2007) Synthesis and biological evaluation of 2- and 3-aminobenzo[b] thiophene derivatives as antimitotic agents and inhibitors of tubulin polymerization. J Med Chem 50:2273–2277. https://doi.org/10.1021/jm070050f
8. Stephens CE, Felder TM, Sowell JW, Andrei G, Balzarini J, Snoeck R, De-Clerq E (2001) Synthesis and antiviral/antitumor evaluation of 2-amino-2-carboxamido-3-aryl-sulfonyl thiophenes and related compounds as a

new class of diarylsulfones. Bio Org Med Chem 9:1123–1132. https://doi.org/10.1016/S0968-0896(00)00333-3

9. Hansch C (1990). In: Ramsden CA (ed) Comprehensive medicinal chemistry, vol 4. Pergamon Press, New York, pp 5–8
10. Li Z, Wan H, Shi Y, Ouyang P (2004) Personal experience with four kinds of chemical structure drawing software: review on ChemDraw, ChemWindow, ISIS/Draw, and ChemSketch. J Chem Inf Comput Sci 44:1886–1890
11. Dandriyal J, Singla R, Kumar M, Jaitak V (2016) Recent developments of C-4 substituted coumarin derivatives as anticancer agents. Eur J Med Chem 119:141–168. https://doi.org/10.1016/j.ejmech.2016.03.087
12. El-Borai MA, Rizk HF, Beltagy DM et al (2013) Microwave-assisted synthesis of some new pyrazolopyridines and their antioxidant, anti-tumor and antimicrobial activities. Eur J Med Chem 66:415–422
13. Salem MS, Ali MAM (2016) Novel Pyrazolo[3,4-b] pyridine derivatives: synthesis, characterization, antimicrobial and antiproliferative profile. Biol Pharm Bull 39:473–483
14. Chavva K, Pillalamarri S, Banda V et al (2013) Synthesis and biological evaluation of novel alkyl amide functionalized trifluoromethyl substituted pyrazolo[3,4-b] pyridine derivatives as potential anti-cancer agents. Bioorg Med Chem Lett 23:5893–5895
15. Becke AD (1993) Becke's three parameter hybrid method using the LYP correlation functional. J Chem Phys 98:5648–5652
16. Albratty M, El-Sharkawy KA, Alam S (2017) Synthesis and antitumor activity of some novel thiophene, pyrimidine, coumarine, pyrazole and pyridine derivatives. Acta Pharm 67:15–33. https://doi.org/10.1515/acph-2017-0004
17. Singh P (2013) Quantitative structure-activity relationship study of substituted-[1, 2, 4] oxadiazoles as S1P1 agonists. J Curr Chem Pharm Sci 3:334–345
18. Kennard RW, Stone LA (1969) Computer aided design of experiments: technometrics. J Sci Res 11:137–148. https://doi.org/10.1080/00401706.1969.10490666
19. 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. https://doi.org/10.1002/qsar.200390007
20. Khaled KF (2011) Modelling corrosion inhibition of iron in acid medium by genetic function approximation method: a QSAR model. Corros Sci 53:3457–3465
21. Friedman JH (1991) Multivariate adaptive regression splines. Ann Stat 19:1–67
22. Adeniji ES, Uba S, Uzairu A (2018) Theoretical modelling for predicting the activities of some active compounds as potent inhibitors against Mycobacterium tuberculosis using GFA-MLR approach. J King Saud Univ Sci 32:575–586. https://doi.org/10.1016/j.jksus.2018.08.010
23. Veerasamy R, Rajak H, Jain A, Sivadasan VCP, Agrawal RK (2011) Validation of QSAR models-strategies and importance. Int J Drug Des Discov 3:511–519
24. Idris MO, Abechi SE, Shallangwa GA, Uzairu A (2020) Insilico elucidation of some quinoline derivative with potent anti-breast cancer activities. J Eng Exact Sci 6(1):265–274
25. Zakari YI, Uzairu A, Shallangwa GA, Abechi SE (2021) Molecular modelling and design of some β-amino alcohol grafted 1,4,5-trisubstituted 1,2,3-triazoles derivatives against chloroquine sensitive, 3D7 strain of Plasmodium falciparum. Heliyon 7:e05924

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.