molecules

MDPI

Article

Quantitative Structure–Activity Relationship Evaluation of MDA-MB-231 Cell Anti-Proliferative Leads

Ajaykumar Gandhi [1,*], Vijay Masand [2], Magdi E. A. Zaki [3,*], Sami A. Al-Hussain [3], Anis Ben Ghorbal [4] and Archana Chapolikar [1]

[1] Department of Chemistry, Government College of Arts and Science, Aurangabad 431 004, Maharashtra, India; dadcguide@gmail.com

[2] Department of Chemistry, Vidya Bharati Mahavidyalaya, Amravati 444 602, Maharashtra, India; vijaymasand@gmail.com

[3] Department of Chemistry, Faculty of Science, Imam Mohammad Ibn Saud Islamic University, Riyadh 13318, Saudi Arabia; sahussain@imamu.edu.sa

[4] Department of Mathematics and Statistics, College of Sciences, Imam Mohammad Ibn Saud Islamic University, Riyadh 13318, Saudi Arabia; assghorbal@imamu.edu.sa

* Correspondence: gascajay18@gmail.com (A.G.); Mezaki@imamu.edu.sa (M.E.A.Z.)

Abstract: In the present endeavor, for the dataset of 219 in vitro MDA-MB-231 TNBC cell antagonists, a (QSAR) quantitative structure–activity relationships model has been carried out. The quantitative and explicative assessments were performed to identify inconspicuous yet pre-eminent structural features that govern the anti-tumor activity of these compounds. GA-MLR (genetic algorithm multi-linear regression) methodology was employed to build statistically robust and highly predictive multiple QSAR models, abiding by the OECD guidelines. Thoroughly validated QSAR models attained values for various statistical parameters well above the threshold values (i.e., $R^2 = 0.79$, $Q^2_{LOO} = 0.77$, $Q^2_{LMO} = 0.76–0.77$, $Q^2\text{-}F^n = 0.72–0.76$). Both de novo QSAR models have a sound balance of descriptive and statistical approaches. Decidedly, these QSAR models are serviceable in the development of MDA-MB-231 TNBC cell antagonists.

Keywords: QSAR; TNBC; MD-MBA-231

check for updates

Citation: Gandhi, A.; Masand, V.; Zaki, M.E.A.; Al-Hussain, S.A.; Ghorbal, A.B.; Chapolikar, A. Quantitative Structure–Activity Relationship Evaluation of MDA-MB-231 Cell Anti-Proliferative Leads. *Molecules* **2021**, *26*, 4795. https://doi.org/10.3390/molecules26164795

Academic Editors: Bono Lučić and Bakhtiyor Rasulev

Received: 15 June 2021
Accepted: 3 August 2021
Published: 7 August 2021

1. Introduction

Cancer is among the most clinically challenging and life-threatening ailments, globally. In 2020, more than 19.29 million new cancer cases and nearly 10 million related deaths worldwide were chronicled, of which 2.3 million are breast cancer cases with 685,000 related deaths [1,2]. In 2021, female breast cancer has overcome lung cancer and become the most common cancer in the world [2]. Oncology experts prognosticate an estimated more than 16 million breast cancer-induced deaths by 2040. Ample research is being done and researchers are persistently improvising by studying novel treatments and drugs, along with new combinations of existing treatments. Based on the response to various methods of treatment, breast cancer is categorized in three clinical subtypes. Two of them, viz. Hormone Receptor (HR)-positive and Human Epidermal growth factor Receptor 2 (HER2)-positive, are reparative through hormone therapy with or without chemotherapy. In these two subtypes, cancer growth is triggered as a response to the hormones, i.e., estrogen or progesterone, or both receptor (ER/PR) and overexpressed HER2 protein [3,4].

Triple-negative breast cancer (TNBC), the third subtype, unlike the first two, does not contain ER, PR or overexpressed HER2 protein, and this makes it hardest to treat. Therefore, chemotherapy is the mainstay for the treatment of TNBC, essentially at all the stages of breast cancer. Breast oncology is based on in vivo and in vitro studies performed against breast cancer cell lines (BCCL). BT–20 is the first BCCL long-established back in 1958. Despite of tireless work in this area, permanent BCCLs obtained have been notably low in number (about 100 only). Most of the available BCCLs are issued from metastatic tumors, mainly from pleural effusions. MCF-7, T-47D, SK-BR-3 and MDA-MB-231 are a few BCCLs well-reported in breast oncology [5,6].

MDA-MB-231, an epithelial human BCCL, is known to be the most aggressive, invasive and ill differentiate TNBC cell line. Proteolytic degradation of the extracellular matrix brings about invasiveness of the MDA-MB-231 cells [3]. Various genres of compounds such as withangulatin–A derivatives [7], urea-based FGFR1 inhibitors [8], Fam20c inhibitors [9], bradykinin B2 agonists [10], spiroketospirazoles [11], triazole spirodienones [12], calix[4]arene-based carbonyl amide derivatives [13], 6-(2-amino-1H-benzo[d] imidazole-6-yl)quinazolin-4(3H)-one derivatives [14], Purine/purine isostere-based scaffolds as new derivatives of benzamides [15], 5-flurouracil scaffold derivatives [16], pyrimidine-thioindole conjugates [17], thieno[3,2-d]pyrimidine derivatives [18], nitrogen-based chalcogens [19], pyrimidine based benzothiazoles [20], 1,3,4-thiadiazoles [21] and 3-Methylthiazolo[3,2-a]benzimidazole-benzenesulfonamides [22] are tested for their anti-proliferative activities against MDA-MB-231 TNBC cells. Still, the thirst for better anti-TNBC drug candidates is not quenched, and researchers are tantalized for further optimization of the leads.

In silico lead optimization is a feasible, economical, absolutely eco-friendly, quantitatively predictive, relatively faster drug discovery approach, which is the utmost need of time. Sparse animal trials and result accustomization make computer-assisted drug designing (CADD) a pragmatic approach. QSAR is one of the prospering branches of CADD, which persistently and outstandingly contributes towards lead optimization [23,24].

QSAR–a cross-curricular, blended approach, ascertains mathematical correlation between structural traits of the molecule and associated bioactivity on a statistical basis. General steps in a QSAR analysis protocol are (I) selection of a sufficiently large, pertinent molecular dataset with desired bioactivity, (II) 3D structure generation and optimization, (III) molecular descriptor calculation and consequential pruning using an apt statistical method, (IV) QSAR model generation using an algorithm that furnishes propitious molecular descriptors and (V) cromulent validation of developed QSAR model(s) [25]. Descriptive QSAR analysis quantitatively interprets the interrelation of salient but superficially enigmatic molecular structural traits with their reported bioactivity. Statistical QSAR predicts the bioactivity of the molecule prior to its laboratory synthesis and in vivo testing. Coherently balanced descriptive and statistical QSAR improves insight for the pharmacokinetics [25–32]. This underlines the importance of QSAR analysis for further optimization of the leads.

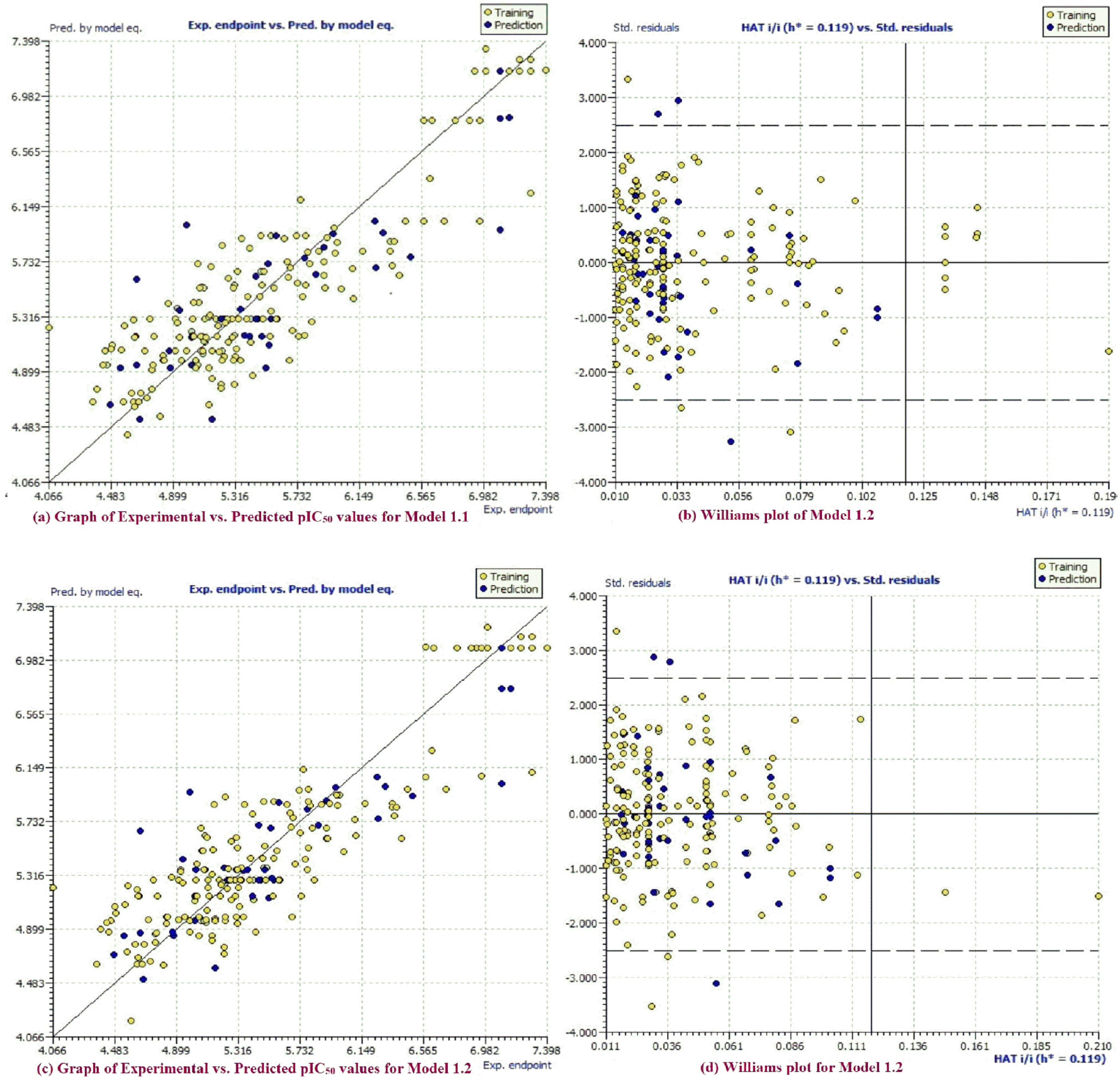
In the present endeavour, a qualitative cum quantitative SAR model for a series of 219 MDA-MB-231 cell anti-proliferative compounds has been performed. The results are useful to optimize compounds for better anti-TNBC activity.

2. Results

Although the present study is based on the moderate size dataset, the presence of diverse molecular scaffolds, functional groups, substituents, different rings, viz. non-aromatic, homoaromatic, heteroaromatic, fused rings, spiro compounds, etc., have notably covered an enormous chemical space. Hence, both of the QSAR models generated are based on the divided dataset only.

Fitting parameters, such as R^2, R^2_{adj}, CCC_{tr}, etc., have values well above the approved threshold values, which confirms that the QSAR models are statistically acceptable with an adequate number of molecular descriptors in them. Internal validation parameters such as Q^2_{LOO}, $Q^2_{LMO,}$ etc. have values that vouchsafe the statistical robustness of the QSAR models (Figure 1a,c). External predictability of both the models is evident from high values of the external validation parameters R^2_{ext}, Q^2-F^n, etc. Williams plots for models 1.1 and 1.2 (Figure 1b,d) corroborate the model applicability domain (AD). Accomplishment of approved threshold values for many parameters, as well as low correlation among the molecular descriptors, rules out the possibility of accidental development of the QSAR models [31–35] (Tables S1 and S2; Figures S1a and S2a in supplementary information).

These evidences substantiate statistical robustness and good external predictability of these models.



(a) Graph of Experimental vs. Predicted pIC_{50} values for Model 1.1

(b) Williams plot of Model 1.2

(c) Graph of Experimental vs. Predicted pIC_{50} values for Model 1.2

(d) Williams plot for Model 1.2

Figure 1. (**a**) Graph of experimental vs. predicted pIC_{50} values for model 1.1; (**b**) Williams plot for model 1.1; (**c**) graph of experimental vs. predicted pIC_{50} values for model 1.2; (**d**) Williams plot for model 1.2.

2.1. GA-MLR QSAR Models

2.1.1. Model-1.1 (Divided Set: Training Set–80% and Prediction Set–20%)

$$pIC_{50} = 4.876(\pm 0.138) + 0.013(\pm 0.006) * \mathbf{all\_MSA3} - 0.538(\pm 0.198) * \mathbf{com\_Splus\_7A} + 0.379(\pm 0.265) * \mathbf{fHringC4B} + 0.107(\pm 0.027) * \mathbf{fOH5B} - 0.178(\pm 0.077) * \mathbf{fringNC8B} + 0.924(\pm 0.140) * \mathbf{com\_sp2N\_2A}.$$

R^2 = 0.79, R^2_{adj} = 0.78, Q^2_{LOO} = 0.77, Q^2_{LMO} = 0.77, RMSEtr = 0.35, MAE_{tr} = 0.27, RSS_{tr} = 21.23, CCC_{tr} = 0.88, $RMSE_{cv}$ = 0.36 MAEcv = 0.28, $PRESS_{cv}$ = 22.89, CCC_{cv} = 0.87, R^2_{ext} = 0.72, Q^2-F^1 = 0.72, Q^2-F^2 = 0.72, Q^2-F^3 = 0.72.

2.1.2. Model-1.2 (Divided Set: Training Set–80% and Prediction Set–20%)

pIC_{50} = 5.072(±0.193) + 0.012(±0.006) * **all_MSA3 –** 0.480(±0.203) * **com_Splus_7A** – 0.096(±0.068) * **com_don_6A** + 0.101(±0.028) * **fOH5B** – 0.160(±0.078) * **fringNC8B** + 0.969(±0.121) * **com_sp2N_2A**.

R^2 = 0.79, R^2_{adj} = 0.78, Q^2_{LOO} = 0.77, Q^2_{LMO} = 0.76, RMSEtr = 0.35, MAE_{tr} = 0.27, RSS_{tr} = 21.24, CCC_{tr} = 0.88, $RMSE_{cv}$ = 0.36 MAEcv = 0.28, $PRESS_{cv}$ = 22.94, CCC_{cv} = 0.87, R^2_{ext} = 0.76, Q^2-F^1 = 0.76, Q^2-F^2 = 0.75, Q^2-F^3 = 0.75.

Although for both these models, values of almost all the statistical parameters related to fitting criteria and internal validation are essentially the same, the differences in values of statistical parameters related to external validation, i.e., R^2_{ext} and Q^2-F^n are noteworthy. This highlights the importance of both of the models. Moreover, these two QSAR models differ in one variable (molecular descriptor) only, viz. **fHringC4B** in model 1.1 (with positive sign of the coefficient) and **com_don_6A** (with negative sign of the coefficient) in model 1.2. In the discussion section, we have explained the importance of these two molecular descriptors, and consequently, the importance of these two QSAR models in terms of their usability and applicability.

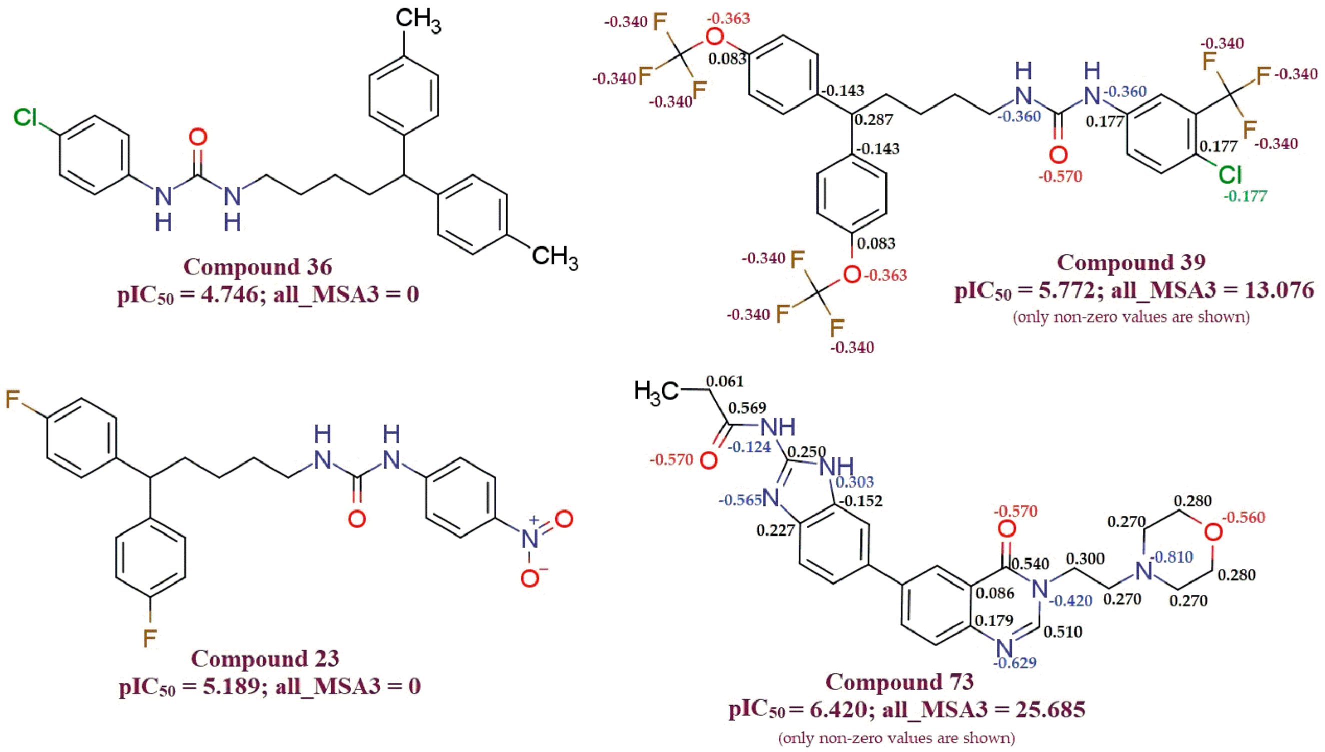
3. Discussion

Among the reported QSAR studies of MDA-MB-231 TNBC cells, the particularly recent work has been done using the dataset of 61 parthenolide derivatives (R^2 = 0.67, Q^2 = 0.55 and R^2_{pred} = 0.53) [36], and yet another using the dataset of 18β –glycyrrhetinic acid derivatives (R^2 = 0.84, Q^2_{LOO} = 0.83 and R^2_{pred} = 0.75) [37]. These QSAR models are developed on the dataset of compounds with a single scaffold, e.g., parthenolide scaffold, 18β–glycyrrhetinic acid etc., and limited pharmacophoric features. This limits the applicability of these QSAR models. The QSAR models developed in the present work are based on the dataset of relatively large number of compounds with various different scaffolds and large number of pharmacophoric features that have increased the scope of applicability of these models. Subjective feature selection provided some simple molecular descriptors; those are reflected in the QSAR models. Values of these molecular descriptors can be easily modified by introducing some simple constitutional and structural alterations to bring about optimization.

Both models 1.1 and 1.2 have been constructed using the divided dataset only. These models differ in only one descriptor out of a total of six descriptors. Although, the change in the activity of each molecule, to a large extent, is a combined effect of all the six molecular descriptors, in the ensuing section effect of variation in each molecular descriptor on biological activity of the irrespective molecule, and is illustrated with examples (see supplementary information Table S3).

1. **all_MSA3, fHringC4B, fOH5B and com_sp2N_2A:** All of these molecular descriptors have positive values of the coefficient and increase in the values of these molecular descriptors, which increases MDA-MB-231 anti-proliferative activity.

all_MSA3 (Molecular Surface Area of all atoms having partial charge in the range of 0.099 to 0.000): The observation is supported by comparing compound 36 (IC_{50} = 4.746) with compound 39 (pIC_{50} = 5.772), for which an increase in the value of **all_MSA3** from 0 for compound 36 to 13.07 for compound 39 results in an increase in pIC_{50} value by about 1 unit (about ten-fold increase in MDA-MB-231 cell anti-proliferative activity). Compound 23 (**all_MSA3** = 0; pIC_{50} = 5.189) and compound 73 (**all_MSA3** = 25.69; pIC_{50} = 6.420) is another pair used as an example to support this observation. Partial charges on each atom of the compound with the non-zero value of the **all_MSA3** descriptor are shown explicitly in Figure 2. The descriptor **all_MSA3** to calculate molecular surface area takes into consideration atoms with partial charges in the range −0.099 to 0.000 only.



Compound 36
pIC_{50} = 4.746; all_MSA3 = 0

Compound 39
pIC_{50} = 5.772; all_MSA3 = 13.076
(only non-zero values are shown)

Compound 23
pIC_{50} = 5.189; all_MSA3 = 0

Compound 73
pIC_{50} = 6.420; all_MSA3 = 25.685
(only non-zero values are shown)

Figure 2. Illustration for molecular descriptor **all_MSA3**.

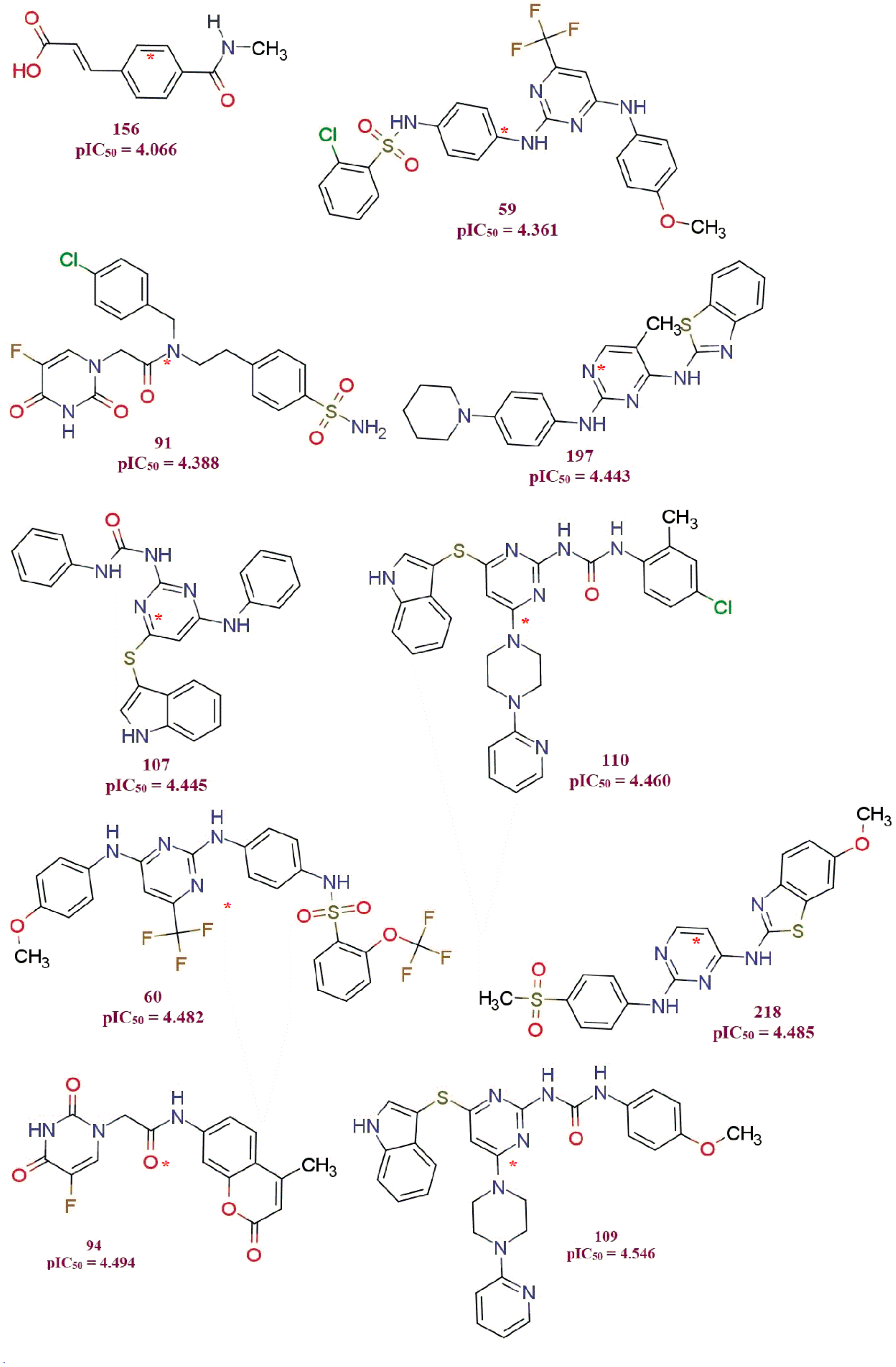
fHringC4B (Frequency of occurrence of ring carbons which are present exactly at four bonds from the hydrogen atom): this molecular feature is absent in all the ten least active compounds but present in all the ten most active compounds (see Figure 3a,b and Figure 4), which highlights the importance of the presence of this molecular feature for better MDA-B-231 anti-proliferative activity in the compound.

fOH5B (Frequency of occurrence of number of hydrogen atoms, which are present exactly at 5 bonds from Oxygen atom): In compound 42 (**fOH5B** = 4; pIC_{50} = 4.914) there are four hydrogen atoms which are five bonds away from oxygen. Whereas, in ten-fold more potent MDA-MB-231 anti-proliferative compound 41 (**fOH5B** = 8, pIC_{50} = 5.910), eight such hydrogens are present. This illustrates the significance of the higher value of the **fOH5B** molecular descriptor in order for leads to be better MBA-MD-231 anti-proliferative agents (Figure 5).

com_sp2N_2A (Number of sp2-nitrogen atoms within 2Å from center of mass of molecule): Significance of this molecular descriptor can be rationalized with the fact that in all the ten least active compounds (Figure 3a), this descriptor gets a value of zero, and in all the ten most active compounds, the value of this descriptor is two (except for compound 183, which has one sp2-N atom within 2Å from the center of the mass of the molecule). (Center of mass in each molecule from Figure 3a shown with red asterisk '*' mark)

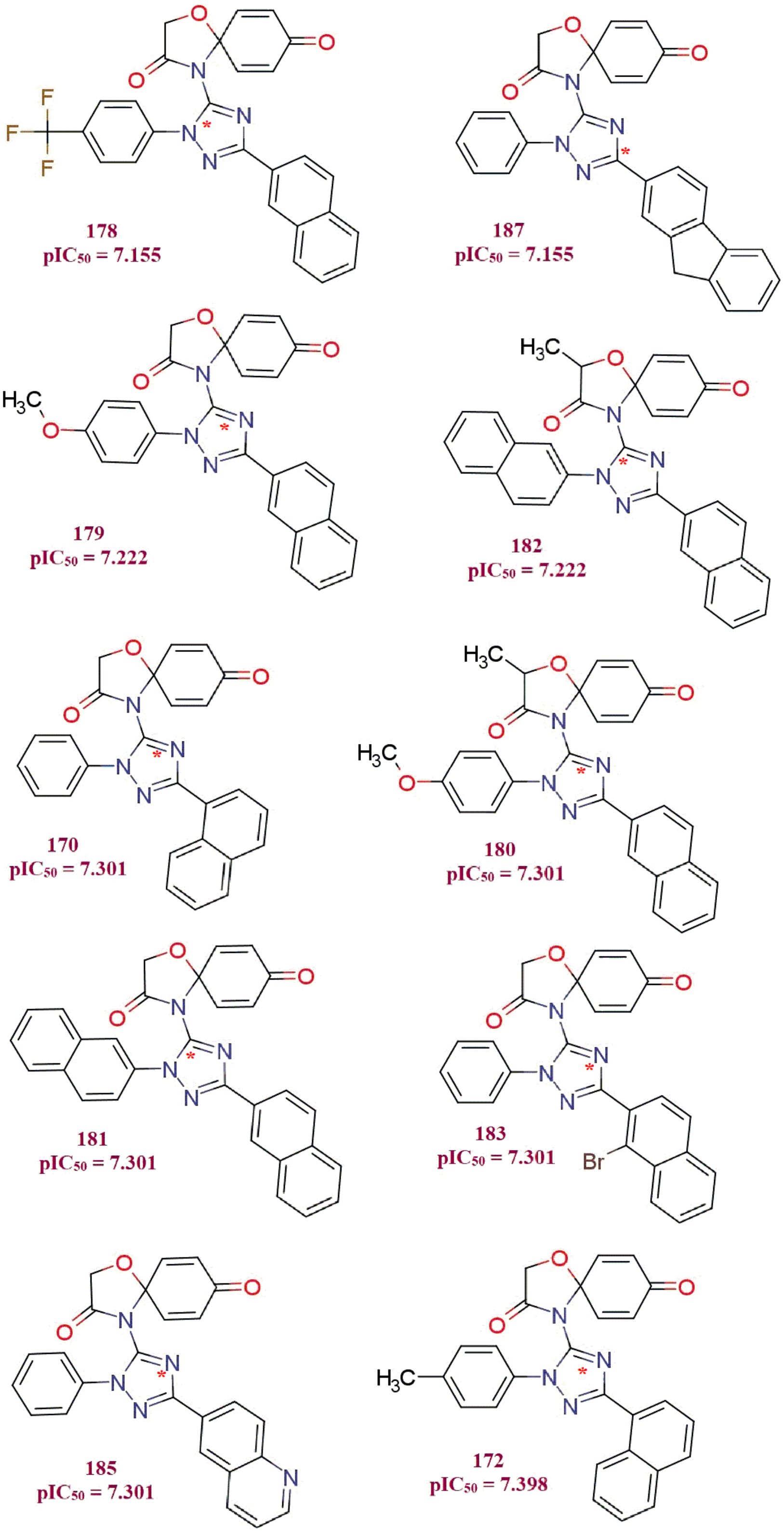
2. **com_Splus_7A, com_don_6A and fringNC8B**: These three molecular descriptors have negative coefficients and hence decrease in their values possibly will increase MDA-MB-231 anti-proliferative activity.

com_Splus_7A (Number of positively charged Sulfur atoms within 7Å from the center of the mass of the molecule) There is no positively charged sulfur atom within 7Å from the center of the mass of molecule (**com_Splus_7A** = 0), rather the sulfur atom itself is absent in these ten most active compounds (pIC_{50} = 7.155–7.398) (Figure 3b). All the compounds having a unit of pIC_{50} > 5.223 have the **com_Splus_7A** descriptor value as zero (Table S3 in supplementary information). (Center of mass in each molecule in Figure 3b shown with red asterisk '*' mark)



(a)

Figure 3. *Cont.*



(b)

Figure 3. Variations in activity and chemical structures in the present dataset of MDA-MB-231 anti-prolifertives: (**a**) ten least active compounds; (**b**) ten most active compounds from the present series. Asterisks * mark denotes the position of the center of the mass of the molecules.

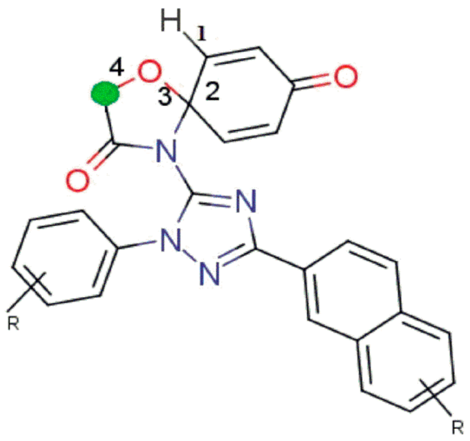


Figure 4. Illustration for molecular descriptor **fHringC4B**.

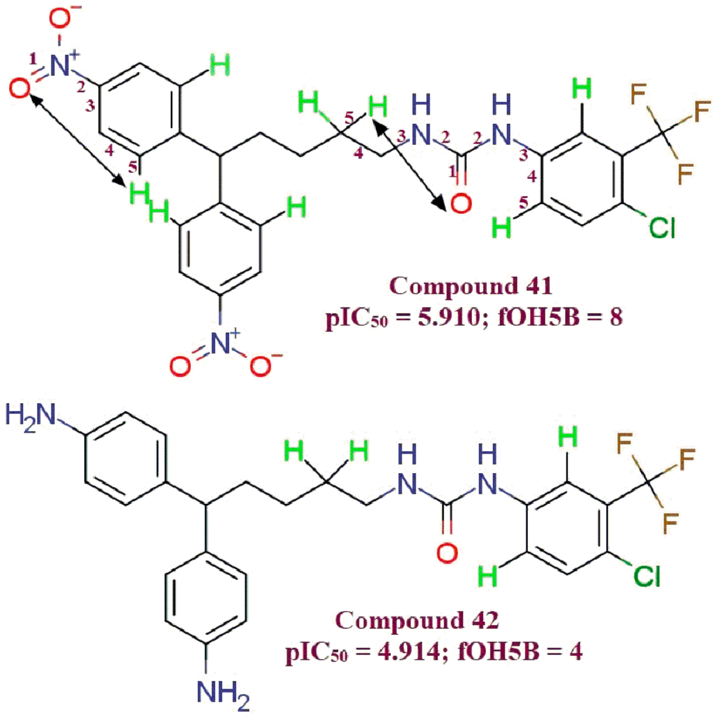
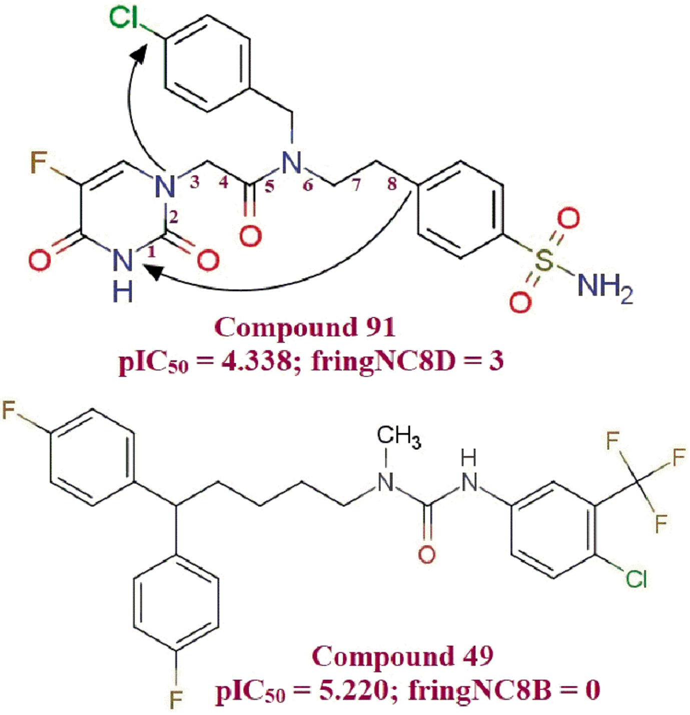


Figure 5. Illustration for molecular descriptor **fOH5B**. (Distinguished hydrogens are shown in green color).

com_don_6A (Number of donor atoms within 6Å from the center of the mass of the molecule): There is no donor atom within 6Å from the center of the mass of the molecule (**com_don_6A**=0); in fact, there is no H-Donor functionality such as N-H, O-H present in all the ten most active compounds (pIC_{50} = 7.155–7.398). Hence, there is great scope to say that the zero value for the **com_Splus_7A** and **com_don_6A** molecular descriptors have gained better MDA-MB-231 cell anti-proliferation potency to them. With two H-donor atoms (in functionality –COOH and –NH-), Compound 156 (**com_don_6A** = 2; pIC_{50} = 4.066) is the least active compound of the series that highlights the importance of absence of **com_don_6A** molecular descriptor for molecule to be better MDA-MB-231 cell anti-proliferator. Comparison of compound 97 (**com_don_6A** = 3; pIC_{50} = 5.159) with 148 (**com_don_6A** = 0; pIC_{50} = 5.212) and 95 (**com_don_6A** = 3; pIC_{50} = 4.916) with 107(**com_don_6A** = 4; pIC_{50} = 4.445) also support the observation.

fringNC8B (Frequency of occurrence of number of carbon atoms which are present exactly at 8 bonds from ring nitrogen atoms): Compound 49 (IC_{50} = 6.02 µM) with no such carbon atoms which are present exactly at eight bonds from the ring nitrogen atom found to be about eight-fold more potent than compound 91 (IC_{50} = 40.92 μM), which contain three such carbons (Figure 6). Moreover, all the ten most active compounds of the series show an absence of carbon atoms, which are present eight bonds away from ring nitrogen (**fringNC8B** = 0). These observations mark the importance of the **fringNC8B** molecular descriptor.



Compound 91
pIC_{50} = 4.338; fringNC8D = 3

Compound 49
pIC_{50} = 5.220; fringNC8B = 0

Figure 6. Illustration for molecular descriptor **fringNC8B**.

Two QSAR models proposed in the present study differ in one variable (molecular descriptor) only, viz. **fHringC4B** in model 1.1 (with positive coefficient) and **com_don_6A** (with negative coefficient) in model 1.2. All the relatively potent MDA-MB-231 cell antiproliferative compounds from the present dataset with $pIC_{50} \geq 7$ have **fHringC4B = 1** and **com_don_6A** = 0 (except compound 154 with **fOH5B** = 9, which more than compensates the counter effects of the rest of the molecular descriptors) and all the relatively inactive compounds with $pIC_{50} \leq 5.043$ have **fHringC4B** = 0 and **com_don_6A** > 0 (i.e., non-zero). This observation signifies the importance of combined effect of the molecular features represented with these molecular descriptors, and eventually broadens the applicability of the present QSAR evaluation studies (see supplementary information Table S3).

There are five outliers, viz. molecules 80, 154, 156, 160 and 183 (with >2.5 σ), which are revealed in the Williams plot (see Figure 1b–d, Figure 7 and supplementary information Tables S1–S3, Figures S1a and S2a). Lipophilic cyclohexyl substituent in compound 80 (Practical pIC_{50} = 4.991, Predicted pIC_{50} = 6.015 by Model 1.1 and Predicted pIC_{50} = 5.961 by Model 1.2) leads to the increase in values of **all_MSA3** and **fOH5B** molecular descriptors due to an increase in –CH_2- groups in the molecule. Hence, both the models predicted extremely high pIC_{50} values for molecule 80. Experimentally, the typical non-planar, chair conformation of the cyclohexyl group due to steric reasons might have restricted inhibitory interaction of compound 80 with the target (pIC_{50} = 4.991), and hence compound 80 turned out as an outlier. Compound 154 (Practical pIC_{50} = 7.097, Predicted pIC_{50} = 5.977 by Model 1.1 and Predicted pIC_{50} = 6.031 by Model 1.2) consists of multiple scaffolds with different pharmacophores which (pIC_{50} = 7.097) might practically have increased the possible inhibitory potency of the molecule through an increased number of favorable interactions, but owing to the same reason, there is an increase in the value of **fringNC8B** and **com_don_6A** molecular descriptors (both have negative coefficient), and this leads to the extremely low pIC_{50} value prediction by both of the models.

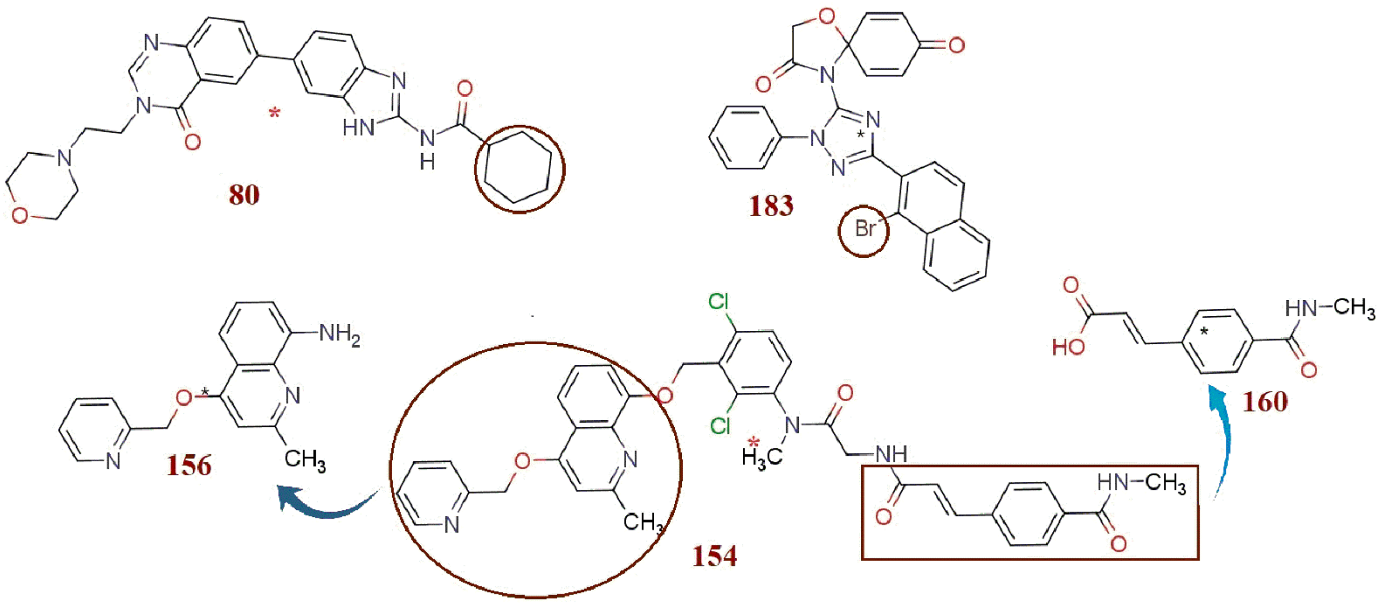


Figure 7. Representation and illustration of outliers.

Compound 156 (Practical pIC_{50} = 4.066, Predicted pIC_{50} = 5.239 by Model 1.1, Predicted pIC_{50} = 5.227 by Model 1.2) and compound 160 (Practical pIC_{50} = 4.658, Predicted pIC_{50} = 5.604 by Model 1.1 and Predicted pIC_{50} = 5.664 by Model 1.2) are developed from some of the structural fragments of compound 154 (Figure 7). These fragments retained their molecular features, which helped these compounds to attain enough large values for a few molecular descriptors, with which both the models predicted high pIC_{50} values for these molecules, but practically, these molecules fall short in a few other molecular features such as length of the molecule, which limited their inhibitory potency and set them in the outliers' category.

The **com_sp2N_2A** molecular descriptor does not take into consideration any such sp2-nitrogen, which is beyond 2Å from the center of the mass of the molecule. In compound 183, (Practical pIC_{50} = 7.301, Predicted pIC_{50} = 6.253 by Model 1.1, Predicted pIC_{50} = 6.113 by Model 1.2), presence of a Bromine (Br)-heavy atom with large atomic/ionic volume as a substituent caused shifting of the position of the center of the mass of the molecule, which reduced the value of **com_sp2N_2A** (positive coefficient) to 1, and hence both models predicted lower pIC_{50} values than was experimentally observed.

4. Materials and Methods

4.1. Dataset Selection

A total of 219 MDA-MB-231 cell antagonists having moderate anti-proliferation potency (experimental IC_{50} = 0.04 to 86 µM) have been selected for the present work [8–12,14–20]. The IC_{50} values were converted to pIC_{50} ($pIC_{50} = -\log IC_{50}$) before actual QSAR analysis. To demonstrate the variation in bioactivity with chemical features covered by the present dataset, ten least and ten most active molecules have been depicted in Figure 3a,b. The SMILES strings with reported IC_{50} and pIC_{50} values for all the molecules are present in Table S4 in the supplementary material.

4.2. Molecular Structure Drawing and Optimization

A free comprehensive chemical drawing package, ChemSketch 12 Freeware (www.acdlabs.com, accessed on 15 May 2021), was used to draw structures of all 219 molecules.

Subsequent conversion of these structures to corresponding 3D structures was achieved using another free and open-source chemistry toolbox, OpenBabel 2.4.0. Thereafter, an optimization and molecular alignment was carried out using the force field MMFF94 available in TINKER (default settings) and Open3DAlign, respectively.

4.3. Molecular Descriptor Calculation and Molecular Descriptor Pruning

PyDescriptor calculated various molecular descriptors for each molecule [38]. More than 15,000 molecular descriptors had been provided by PyDescriptor for each molecule. It is necessarily important to remove redundant molecular descriptors to steer clear of the impinging of multi-collinear and spurious variables in the GA-MLR model. Hence, molecular descriptors with nearly constant values (>95%) and co-linearity (|R|) >0.95 were removed using objective feature selection (OFS) in QSARINS v2.2.4 [39,40]. The contracted molecular descriptor pool thus resulted is comprised of 1370 molecular descriptors only.

4.4. QSAR Model–Development and Validation

The contracted molecular descriptor set is a heterogeneous set of descriptors in the sense that it comprises 0D to 3D descriptors, charge descriptors and molecular properties, etc. that have covered an adequately comprehensive descriptor space. Subjective feature selection (SFS) in QSARINS v2.2.4 was executed in order to construct statistically robust GA-MLR-based QSAR models. Thereafter, thorough statistical validation (internal and external), Y-randomization and applicability domain analysis of the derived models was done, abiding by OECD [41–44] principles. A dataset of 219 molecules is large enough that even on splitting, the size of training dataset has covered the chemical space adequately and to great extent. Following are the steps in protocol for QSAR model construction using the divided dataset:

i. As per OECD guidelines, thorough internal as well as external validation of the developed QSAR model(s), for example, is necessarily mandatory. Hence, some molecules from the dataset were randomly kept aside as a prediction set, and remaining molecules (training set) were subjected to SFS treatment to develop the QSAR model. The QSAR model(s) generated is validated using molecules in the prediction set. Random splitting of the dataset using random splitting option in QSARINS v.2.2.4 into an 80% training set (175 molecules in training set) and a 20% prediction set (44 molecules in prediction set) was achieved. The training set was used for QSAR model development, and the prediction set was utilized for external validation.

ii. QSARINS v2.2.4 with default settings and Q^2_{LOO} as a fitness function for feature selection was deployed in genesis of the GA-MLR-based QSAR models with double cross validation. Up to six variables, there was a generous increase in the Q^2_{LOO} value, but minor augmentation was observed thereafter. Consequently, the selection of the molecular descriptor was confined to a set of six descriptors to foil the danger of over-fitting, and this additionally helped to derive easy and informative QSAR models (see supplementary information Table S3 values for all the selected molecular descriptors present in QSAR models).

iii. Abide by OECD guidelines; for ensured proper validation, all the models were subjected to internal and external validation, Y-randomization and model applicability domain (AD) analysis using QSARINS 2.2.4. Robustness of the GA-MLR-based QSAR model was adjudicated on the basis of (a) internal validation based on Leave-One-Out (LOO) and Leave-Many-Out (LMO) procedure; (b) external validation; (c) Y-randomization and (d) fulfilling of the respective threshold value for the statistical parameters: $R^2 \geq 0.6$, $Q^2_{LOO} \geq 0.5$, $Q^2_{LMO} \geq 0.6$, $R^2 > Q^2$, $R^2_{ex} \geq 0.6$, $RMSE_{tr} < RMSE_{cv}$, $\Delta K \geq 0.05$, CCC ≥ 0.80, $Q^2\text{-}F^n \geq 0.60$, $r^2m \geq 0.6$, $0.9 \leq k \leq 1.1$, $0.9 \leq k' \leq 1.1$ with RMSE and MAE close to zero. All QSAR models which failed to meet any of these criteria were omitted. Two QSAR models (1.1 and 1.2) with best values of these parameters and with best predicative ability ($Q^2\text{-}F^n > 0.71$) were selected.

5. Conclusions

A well-founded balance of statistical QSAR and descriptive QSAR with highly precise bioactivity predictability (external predictability) on incorporation of molecular features is furnished by both models. Various statistical parameters that are indicators of preciseness of the external predictability, especially R^2_{ext} and Q^2-F^n, resulted in extremely high values for both models. Molecular features which appeared in QSAR models, such as an increased number of four bond-distant ring carbons from hydrogen, five bond-distant hydrogen from oxygen, a lesser number of eight-bond distant carbons from ring nitrogen, etc., are easy to incorporate in a manner that will make optimization easy and scopeful. These models will help in optimizing present compounds to more potent MBA-MD-231 anti-proliferative leads to treat triple-negative breast cancer.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/molecules26164795/s1, Table S1: Details regarding performance of model 1.1. Table S2: Details regarding performance of model 1.2. Table S3: The values for selected molecular descriptors present in QSAR models. Table S4: The SMILES notation for two hundred and nineteen MDA-MB-231 cell anti-proliferative leads, along with their reported IC_{50} and pIC_{50} values; Figure S1: Different graphs associated with model 1.1 (a) graph of pred. endpoint vs. residual values (b) Y-scrambling plot. Figure S2: Different graphs associated with model 1.2 (a) graph of pred. endpoint vs. residual values (b) Y-scrambling plot. Statistical parameters for used for validation of QSAR models

Author Contributions: Conceptualization, A.G.; Data curation, M.E.A.Z., S.A.A.-H. and A.B.G.; Formal analysis, A.G., V.M. and A.C.; Methodology, A.G., V.M. and A.B.G.; Writing—original draft, A.G. and S.A.A.-H.; Writing—review and editing, V.M., M.E.A.Z. and A.C. All authors have read and agreed to the published version of the manuscript.

Funding: The authors extend their appreciation to the Deanship of Scientific Research at Imam Mohammad bin Saud Islamic University for funding this work through Research Group no-21-09-77.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data are available in the supplementary section.

Acknowledgments: The authors are thankful to Paola Gramatica and her team for providing QSARINS-v2.2.4 and developers of TINKER, ChemSketch 12 Freeware (ACD labs), and PyDescriptor for providing the free versions of their software. Archana Chapolikar and Ajaykumar Gandhi are grateful to the Government College of Arts and Science, Aurangabad, Maharashtra, India. Magdi E.A. Zaki, Sami A. Al-Hussain and Anis Ben Ghorbal are thankful to the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University for funding this work through Research Group no-21-09-77.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CADD	Computer-Aided Drug Designing
SMILES	Simplified Molecular-Input Line-Entry System
GA	Genetic Algorithm
MLR	Multiple Linear Regression
QSAR	Quantitative Structure–Activity Relationship
OLS	Ordinary Least Square
QSARINS	QSAR Insubria
OECD	Organization for Economic Co-operation and Development
OFS	Objective Feature Selection
SFS	Subjective Feature Selection
HER2	Human Epidermal growth factor Receptor 2
TNBC	Triple Negative Breast Cancer

ER	Estrogen Receptor
PR	Progesterone Receptor
BCCL	Breast Cancer Cell Line
LOF	Lack of Fit (Friedmann Parameter)
RMSE	Root Mean Square Error
MAE	Mean Absolute Error
RSS	Residual Sum of Squares
CCC	Concordance Correlation Coefficient
PRESS	Predictive Residual Sum of Squares
LOO	Leave One Out
LMO	Leave Many Out

References

1. Ferlay, J.; Ervik, M.; Lam, F.; Colombet, M.; Mery, L.; Piñeros, M.; Znaor, A.; Soerjomataram, I.; Bray, F.; Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. 2020. Available online: https://gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf (accessed on 29 July 2021).
2. Ferlay, J.; Ervik, M.; Lam, F.; Colombet, M.; Mery, L.; Piñeros, M.; Znaor, A.; Soerjomataram, I.; Bray, F.; Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. 2020. Available online: https://www.who.int/news-room/fact-sheets/detail/cancer (accessed on 29 July 2021).
3. European College of Authenticated Cell cultures Cell line profile MDA-MB-231. *Eur. Collect. Authenticated Cell Cult.* **2017**, *231*, 1–3, MDA-MB-231 (ECACC 92020424).
4. Mendelsohn, J.; Howley, P.M.; Israel, M.A.; Gray, J.W.; Thompson, C.B. *The Molecular Basis of Cancer*; Elsevier Saunders: Philadelphia, PA, USA, 2015; ISBN 978-1-4557-4066-6.
5. Holliday, D.L.; Speirs, V. Choosing correct breast cancer cell line for breast cancer research. *Breast Cancer Res.* **2011**, *13*, 1–7. [CrossRef]
6. Freischel, A.R.; Damaghi, M.; Cunningham, J.J.; Ibrahim-Hashim, A.; Gillies, R.J.; Gatenby, R.A.; Brown, J.S. Frequency-dependent interactions determine outcome of competition between two breast cancer cell lines. *Sci. Rep.* **2021**, *11*, 1–18. [CrossRef]
7. Zhou, W.X.; Chen, C.; Liu, X.Q.; Li, Y.; Lin, Y.L.; Wu, X.T.; Kong, L.Y.; Luo, J.G. Discovery and optimization of withangulatin A derivatives as novel glutaminase 1 inhibitors for the treatment of triple-negative breast cancer. *Eur. J. Med. Chem.* **2021**, *210*, 112980. [CrossRef] [PubMed]
8. Ashraf-Uz-Zaman, M.; Shahi, S.; Akwii, R.; Sajib, M.S.; Farshbaf, M.J.; Kallem, R.R.; Putnam, W.; Wang, W.; Zhang, R.; Alvina, K.; et al. Design, synthesis and structure-activity relationship study of novel urea compounds as FGFR1 inhibitors to treat metastatic triple-negative breast cancer. *Eur. J. Med. Chem.* **2021**, *209*, 112866. [CrossRef] [PubMed]
9. Zhao, R.; Fu, L.; Yuan, Z.; Liu, Y.; Zhang, K.; Chen, Y.; Wang, L.; Sun, D.; Chen, L.; Liu, B.; et al. Discovery of a novel small-molecule inhibitor of Fam20C that induces apoptosis and inhibits migration in triple negative breast cancer. *Eur. J. Med. Chem.* **2021**, *210*, 113088. [CrossRef]
10. Rassias, G.; Leonardi, S.; Rigopoulou, D.; Vachlioti, E.; Afratis, K.; Piperigkou, Z.; Koutsakis, C.; Karamanos, N.K.; Gavras, H.; Papaioannou, D. Potent antiproliferative activity of bradykinin B2 receptor selective agonist FR-190997 and analogue structures thereof: A paradox resolved? *Eur. J. Med. Chem.* **2021**, *210*, 112948. [CrossRef] [PubMed]
11. Huang, T.; Wu, X.; Yan, S.; Liu, T.; Yin, X. Synthesis and in vitro evaluation of novel spiroketopyrazoles as acetyl-CoA carboxylase inhibitors and potential antitumor agents. *Eur. J. Med. Chem.* **2021**, *212*, 113036. [CrossRef]
12. Luo, L.; Jia, J.J.; Zhong, Q.; Zhong, X.; Zheng, S.; Wang, G.; He, L. Synthesis and anticancer activity evaluation of naphthalene-substituted triazole spirodienones. *Eur. J. Med. Chem.* **2021**, *213*, 113039. [CrossRef]
13. An, L.; Wang, C.; Zheng, Y.G.; Liu, J.D.; Huang, T.H. Design, synthesis and evaluation of calix[4]arene-based carbonyl amide derivatives with antitumor activities. *Eur. J. Med. Chem.* **2021**, *210*, 112984. [CrossRef] [PubMed]
14. Fan, C.; Zhong, T.; Yang, H.; Yang, Y.; Wang, D.; Yang, X.; Xu, Y.; Fan, Y. Design, synthesis, biological evaluation of 6-(2-amino-1H-benzo[d]imidazole-6-yl)quinazolin-4(3H)-one derivatives as novel anticancer agents with Aurora kinase inhibition. *Eur. J. Med. Chem.* **2020**, *190*, 112108. [CrossRef]
15. Nepali, K.; Chang, T.Y.; Lai, M.J.; Hsu, K.C.; Yen, Y.; Lin, T.E.; Lee, S.B.; Liou, J.P. Purine/purine isoster based scaffolds as new derivatives of benzamide class of HDAC inhibitors. *Eur. J. Med. Chem.* **2020**, *196*, 112291. [CrossRef] [PubMed]
16. Petreni, A.; Bonardi, A.; Lomelino, C.; Osman, S.M.; ALOthman, Z.A.; Eldehna, W.M.; El-Haggar, R.; McKenna, R.; Nocentini, A.; Supuran, C.T. Inclusion of a 5-fluorouracil moiety in nitrogenous bases derivatives as human carbonic anhydrase IX and XII inhibitors produced a targeted action against MDA-MB-231 and T47D breast cancer cells. *Eur. J. Med. Chem.* **2020**, *190*, 112112. [CrossRef]
17. Sana, S.; Reddy, V.G.; Bhandari, S.; Reddy, T.S.; Tokala, R.; Sakla, A.P.; Bhargava, S.K.; Shankaraiah, N. Exploration of carbamide derived pyrimidine-thioindole conjugates as potential VEGFR-2 inhibitors with anti-angiogenesis effect. *Eur. J. Med. Chem.* **2020**, *200*, 112457. [CrossRef] [PubMed]

18. Wang, R.; Yu, S.; Zhao, X.; Chen, Y.; Yang, B.; Wu, T.; Hao, C.; Zhao, D.; Cheng, M. Design, synthesis, biological evaluation and molecular docking study of novel thieno[3,2-d]pyrimidine derivatives as potent FAK inhibitors. *Eur. J. Med. Chem.* **2020**, *188*, 112024. [CrossRef] [PubMed]
19. Elkhalifa, D.; Siddique, A.B.; Qusa, M.; Cyprian, F.S.; El Sayed, K.; Alali, F.; Al Moustafa, A.E.; Khalil, A. Design, synthesis, and validation of novel nitrogen-based chalcone analogs against triple negative breast cancer. *Eur. J. Med. Chem.* **2020**, *187*, 111954. [CrossRef] [PubMed]
20. Diao, P.C.; Lin, W.Y.; Jian, X.E.; Li, Y.H.; You, W.W.; Zhao, P.L. Discovery of novel pyrimidine-based benzothiazole derivatives as potent cyclin-dependent kinase 2 inhibitors with anticancer activity. *Eur. J. Med. Chem.* **2019**, *179*, 196–207. [CrossRef]
21. Liu, T.; Wan, Y.; Liu, R.; Ma, L.; Li, M.; Fang, H. Improved antiproliferative activities of a new series of 1,3,4-thiadiazole derivatives against human leukemia and breast cancer cell lines. *Chem. Res. Chin. Univ.* **2016**, *32*, 768–774. [CrossRef]
22. Alkhaldi, A.A.M.; Al-Sanea, M.M.; Nocentini, A.; Eldehna, W.M.; Elsayed, Z.M.; Bonardi, A.; Abo-Ashour, M.F.; El-Damasy, A.K.; Abdel-Maksoud, M.S.; Al-Warhi, T.; et al. 3-Methylthiazolo[3,2-a]benzimidazole-benzenesulfonamide conjugates as novel carbonic anhydrase inhibitors endowed with anticancer activity: Design, synthesis, biological and molecular modeling studies. *Eur. J. Med. Chem.* **2020**, *207*, 112745. [CrossRef]
23. Baldi, A. Computational approaches for drug design and discovery: An overview. *Syst. Rev. Pharm.* **2010**, *1*, 99–105. [CrossRef]
24. Joy, S.; Vijayakumar, Y.M.; Sunhye, G. Role of computer-aided drug design in modern drug discovery. *Arch. Pharm. Res.* **2015**, *38*, 1686–1701. [CrossRef]
25. Cherkasov, A.; Muratov, E.N.; Fourches, D.; Varnek, A.; Baskin, I.I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y.C.; Todeschini, R.; et al. QSAR modeling: Where have you been? Where are you going to? *J. Med. Chem.* **2014**, *57*, 4977–5010. [CrossRef]
26. Chirico, N.; Gramatica, P. Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. *J. Chem. Inf. Model.* **2012**, *52*, 2044–2058. [CrossRef] [PubMed]
27. Martin, T.M.; Harten, P.; Young, D.M.; Muratov, E.N.; Golbraikh, A.; Zhu, H.; Tropsha, A. Does rational selection of training and test sets improve the outcome of QSAR modeling? *J. Chem. Inf. Model.* **2012**, *52*, 2570–2578. [CrossRef] [PubMed]
28. Fujita, T.; Winkler, D.A. Understanding the Roles of the "two QSARs". *J. Chem. Inf. Model.* **2016**, *56*, 269–274. [CrossRef]
29. Huang, J.; Fan, X. Why QSAR fails: An empirical evaluation using conventional computational approach. *Mol. Pharm.* **2011**, *8*, 600–608. [CrossRef] [PubMed]
30. Masand, V.H.; Mahajan, D.T.; Nazeruddin, G.M.; Hadda, T.B.; Rastija, V.; Alfeefy, A.M. Effect of information leakage and method of splitting (rational and random) on external predictive ability and behavior of different statistical parameters of QSAR model. *Med. Chem. Res.* **2015**, *24*, 1241–1264. [CrossRef]
31. Gramatica, P.; Cassani, S.; Roy, P.P.; Kovarich, S.; Yap, C.W.; Papa, E. QSAR modeling is not "Push a button and find a correlation": A case study of toxicity of (Benzo-)triazoles on Algae. *Mol. Inform.* **2012**, *31*, 817–835. [CrossRef]
32. Gramatica, P. On the development and validation of QSAR models. *Methods Mol. Biol.* **2013**, *930*, 499–526. [CrossRef]
33. Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the Definition of the Q2 Parameter for QSAR Validation. *J. Chem. Inf. Model.* **2009**, *49*, 1669–1678. [CrossRef]
34. Consonni, V.; Todeschini, R.; Ballabio, D.; Grisoni, F. On the Misleading Use of QF32 for QSAR Model Comparison. *Mol. Inform.* **2019**, *38*, 1800029. [CrossRef] [PubMed]
35. Chirico, N.; Gramatica, P. Real external predictivity of QSAR models: How to evaluate It? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *J. Chem. Inf. Model.* **2011**, *51*, 2320–2335. [CrossRef] [PubMed]
36. Lawal, H.A.; Uzairu, A.; Uba, S. QSAR, molecular docking studies, ligand-based design and pharmacokinetic analysis on Maternal Embryonic Leucine Zipper Kinase (MELK) inhibitors as potential anti-triple-negative breast cancer (MDA-MB-231 cell line) drug compounds. *Bull. Natl. Res. Cent.* **2021**, *45*. [CrossRef]
37. Shukla, A.; Tyagi, R.; Meena, S.; Datta, D.; Srivastava, S.K.; Khan, F. 2D- and 3D-QSAR modelling, molecular docking and in vitro evaluation studies on 18β-glycyrrhetinic acid derivatives against triple-negative breast cancer cell line. *J. Biomol. Struct. Dyn.* **2020**, *38*, 168–185. [CrossRef]
38. Masand, V.H.; Rastija, V. PyDescriptor: A new PyMOL plugin for calculating thousands of easily understandable molecular descriptors. *Chemom. Intell. Lab. Syst.* **2017**, *169*, 12–18. [CrossRef]
39. Gramatica, P.; Cassani, S.; Chirico, N. QSARINS-chem: Insubria datasets and new QSAR/QSPR models for environmental pollutants in QSARINS. *J. Comput. Chem.* **2014**, *35*, 1036–1044. [CrossRef] [PubMed]
40. Gramatica, P.; Chirico, N.; Papa, E.; Cassani, S.; Kovarich, S. QSARINS: A new software for the development, analysis, and validation of QSAR MLR models. *J. Comput. Chem.* **2013**, *34*, 2121–2132. [CrossRef]
41. OECD Validation of (Q)SAR Models–OECD. Available online: https://www.oecd.org/env/ehs/riskassessment/validationofqsarmodels.htm (accessed on 17 July 2021).
42. OECD. *Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models*; OECD Series on Testing and Assessment; No. 69; OECD Publishing: Paris, France, 2014. [CrossRef]

43. Group, Q.E. The report from the expert group on (Quantitative) Structure-Activity Relationships [(Q)SARs] on the principles for the validation of (Q)SARs. *Organ. Econ. CO-OPERATION Dev. Paris* **2004**, *49*, 206.
44. 37th Joint Meeting of the Chemicals Committee, OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationship models These principles were agreed by OECD member countries at the 37. *Biotechnology* **2004**, 3–4. Available online: https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf (accessed on 29 July 2021).