Mol Divers (2016) 20:729–739
DOI 10.1007/s11030-016-9672-0

ORIGINAL ARTICLE

# In silico evaluation, molecular docking and QSAR analysis of quinazoline-based EGFR-T790M inhibitors

M. Asadollahi-Baboli[1]

Received: 21 December 2015 / Accepted: 2 May 2016 / Published online: 21 May 2016

**Abstract** Mutated epidermal growth factor receptor (EGFR-T790M) inhibitors hold promise as new agents against cancer. Molecular docking and QSAR analysis were performed based on a series of fifty-three quinazoline derivatives to elucidate key structural and physicochemical properties affecting inhibitory activity. Molecular docking analysis identified the true conformations of ligands in the receptor's active pocket. The structural features of the ligands, expressed as molecular descriptors, were derived from the obtained docked conformations. Non-linear and spline QSAR models were developed through novel genetic algorithm and artificial neural network (GA-ANN) and multivariate adaptive regression spline techniques, respectively. The former technique was employed to consider non-linear relation between molecular descriptors and inhibitory activity of quinazoline derivatives. The later technique was also used to describe the non-linearity using basis functions and sub-region equations for each descriptor. Our QSAR model gave a high predictive performance ($R^2_{\mathrm{p}} = 0.881$, $Q^2_{\mathrm{LOO}} = 0.923$, $R^2_{\mathrm{LSO}} = 0.828$ and $r^2_{\mathrm{m}} = 0.772$) using diverse validation techniques. Eight new compounds were designed using our QSAR model as potent EGFR-T790M inhibitors. Overall, the proposed in silico strategy based on docked derived descriptor and non-linear descriptor subset selection may help design novel quinazoline derivatives with improved EGFR-T790M inhibitory activity.

**Keywords** Quinazolines · Molecular docking · GA-ANN · MARS · EGFR · Descriptor selection

✉ M. Asadollahi-Baboli
asadollahi@nit.ac.ir

[1] Department of Science, Babol University of Technology, Babol, Mazandaran 47148-71167, Iran

## Introduction

The mutation or overexpression of EGFR (epidermal growth factor receptor) has been observed in many human tumors, and is under intense investigation as a novel anticancer molecular target [1]. The EGFR pathways can lead to the initiation of cancer proliferation and increased neoangiogenesis and metastatic potential. EGFR inhibition by inhibitors such as erlotinib and gefitinib results in a relatively effective treatment for patients for a while [2]. However, many patients will ultimately manifest disease progression because of drug resistance due to T790M mutation in the EGFR kinase domain especially in squamous head, neck carcinomas, ovarian, lung, and breast cancers [3]. Despite the rapid advances in EGFR oncology therapeutics over the past decade, substantial room for improvement remains. Several challenges exist in the effective inhibition of T790M clinically. Previously designed EGFR inhibitors did not demonstrate distinctly improved efficacy against T790M mutations which is partially attributed to the dosage limitation imposed by the drugs' toxicity [4]. The design of EGFR-T790M inhibitors with high activity and low dosage has proven to be quite challenging since only few reports exist [5]. Therefore, more studies are needed to find potent, promising, and comparably efficacious inhibitors against EGFR-T790M resistance tumor. Recently, some novel quinazoline derivatives have been shown to selectively inhibit EGFR-T790M with high potency and good binding affinity [6] which prompted us to explore the relationship between the structures of quinazoline derivatives and EGFR-T790M inhibition. The cost-effective computational routines can speed up the procedure of drug design by predicting the biological activity of drug-like molecules and identifying receptor–ligand interactions [7]. Molecular docking and QSAR (quantitative structure activity relationship) techniques can be used to guide lead optimization and in silico evaluation for EGFR-T790M inhibitors. This may help find and design new targets that fall within applicability domain.

To the best of our knowledge, no QSAR study has yet been reported on quinazoline-based EGFR-T790M inhibitors. The main aim of the present contribution was to develop validated QSAR models to predict the inhibition profiles of quinazolines. In silico evaluation was performed keeping the following points in mind. First, molecular docking studies have been used to identify the position and orientation of ligands (inhibitors) in the active pocket of the EGFR-T790M receptor and to facilitate an interpretation of receptor–ligand molecular interactions. Second, various classes of molecular descriptors were calculated from the optimum ligand conformation obtained from docking results. Third, a straightforward non-linear descriptor selection method was used based on a GA-ANN (genetic algorithm-artificial neural networks) strategy to investigate the non-linear dependencies between descriptors and inhibitory activities [8]. Fourth, an interpretable MARS (multivariate adaptive regression spline) algorithm was employed to find sub-region equations and spline functions describing the non-linearity in selected descriptors [9]. Fifth, as a critical step to evaluate the predictive power and reliability of QSAR models, internal and external validations were performed using an external prediction set, leave-one-out (LOO) cross-validation (C.V.), leave-some-out (LSO) procedure, diversity analysis, Y-scrambling, and applicability domain together with other statistical parameters.

## Methods and materials

### Biological data set and molecular optimization

In the present QSAR analysis, a series of fifty-three quinazoline-based EGFR-T790M inhibitors, together with their inhibitory concentrations, were taken from the literature [6]. The inhibitory concentration ($IC_{50}$) value was converted to its corresponding logarithmic scale $pIC_{50}$ value which was taken as the dependent parameter for QSAR analysis. The $IC_{50}$, an inhibition activity parameter, is a measure of a compound's inhibitory effectiveness and refers to the concentration required for a compound to reduce the level of EGFR activity by 50 %. The structures of EGFR inhibitors and their corresponding $pIC_{50}$ values are presented in Table 1. The 2D structures of the inhibitors in Table 1 were sketched and their 3D structures were subjected to energy minimization using the molecular mechanics force field (MMFF) method implemented in the Molecular Operating Environment (MOE2014.09, Chemical Computing Group, Inc.) package with a convergence criterion of 0.01 kcal $mol^{-1}$ and partial atomic charges. The optimized structures were used for molecular docking in the next section.

### Energy minimization and molecular docking

The coordinates of the EGFR crystal structure were obtained from the Protein Data Bank (www.rcsb.org) (PDB ID: 2JIU). Molecular docking studies were carried out using MOE2014.09 in order to explore the interaction mechanism and probable binding mode of quinazoline-based EGFR inhibitors at the active pocket of the 2JIU protein crystal. The energy minimization algorithm of MOE was used to minimize the energy of the protein using the MMFF94x force field with conjugant gradient method [10]. For molecular docking, the ligand was considered fully flexible, while the protein was considered rigid. All the torsion angles in the inhibitors were also set free to perform flexible docking with a grid-point spacing of 0.3 Å. The docking calculations were performed by setting a 70 × 70 × 70 $Å^3$ grid map centered in the binding pocket of EGFR using all the amino acid residues within the grid map. A total of 50 conformations were saved for each ligand using the default parameters in MOE. Simultaneously, all appropriate conformations and orientations were predicted with the best binding free energy. Then, the best docked conformation for each inhibitor was used to calculate molecular descriptors.

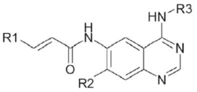
### Molecular descriptor generation

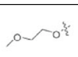
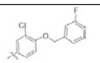
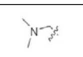
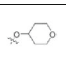
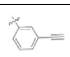
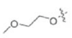
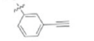
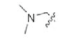
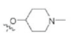
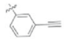
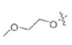
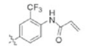
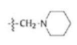
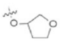
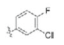
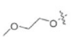
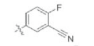
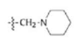
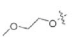
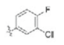
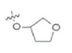
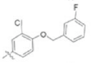
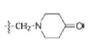
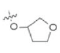
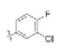
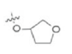
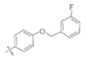
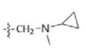
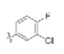
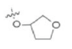
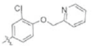
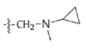
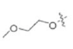
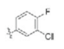
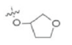
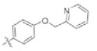
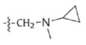
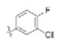
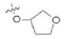
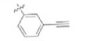
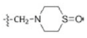
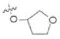
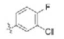
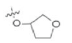
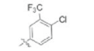
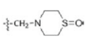
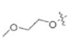
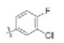
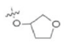
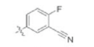
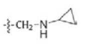
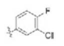
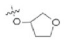
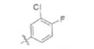
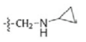
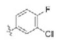
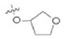
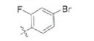
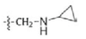
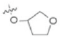
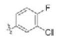
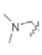
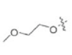
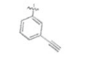
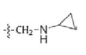
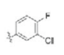
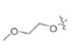
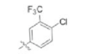
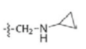
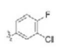
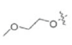
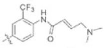
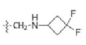
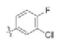
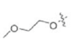
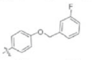
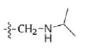
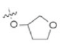
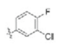
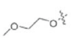
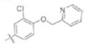
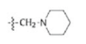
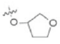
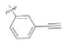
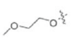
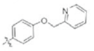
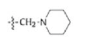
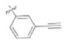
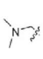
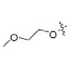
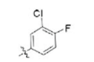
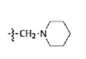
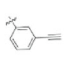
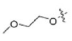
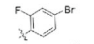
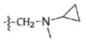
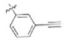
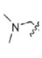
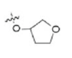
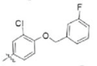
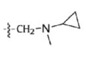
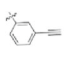
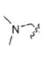
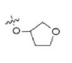
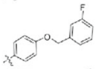
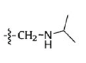
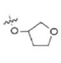
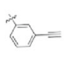
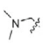
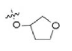
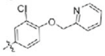
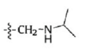
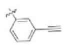
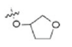
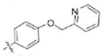
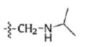
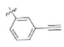
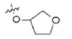
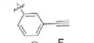
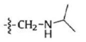
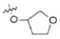
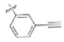
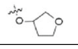
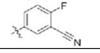
Different types of molecular descriptors (constitutional, one dimensional, two dimensional, and three dimensional) were used for QSAR analysis. These molecular descriptors encoded significant structural features of molecules. A total of 287 descriptors were calculated using the E-Dragon software [11] based on molecular structures and the quantum theory of atoms in molecules after removing invariable descriptors for all molecules and correlated descriptors ($R > 0.9$). The list of these descriptors, their meaning, and the calculation procedures are provided in the literature [12]. These descriptors can be categorized into eighteen different descriptor families: Constitutional, Topological, 2D autocorrelations, Walk and path counts, Connectivity indices, Information indices, Edge adjacency indices, Randic molecular profiles, Geometrical descriptors, RDF, 3D-MoRSE, BCUT, Topological charge indices, WHIM, GETAWAY, Charge, Molecular properties, and Eigenvalue-based indices.

### Descriptor selection and modeling tools

Descriptor selection is one of the most crucial steps in the development of QSAR models [13]. For the selection of $m$ descriptors from the descriptor poll ($n$ descriptors), there are $\frac{n!}{m!(n-m)!}$ possible combinations. Many conventional descrip-

**Table 1** Experimental and calculated inhibitor data using GA-ANN/MARS strategy for quinazoline derivatives together with their structures used in QSAR study



| No | Sub set[a] | R1 | R2 | R3 | Exp. $pIC_{50}$ | Pred. $pIC_{50}$ | No | Sub set | R1 | R2 | R3 | Exp. $pIC_{50}$ | Pred. $pIC_{50}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | S3 | H |  |  | 6.99 | 7.11 | 28[p] | S4 |  |  |  | 8.22 | 7.83 |
| 2[p] | S2 | H |  |  | 8.74 | 8.44 | 29 | S2 |  |  |  | 8.09 | 7.63 |
| 3 | S4 | H |  |  | 6.94 | 7.37 | 30 | S3 |  |  |  | 8.28 | 8.29 |
| 4 | S1 | H |  |  | 6.71 | 6.78 | 31 | S5 |  |  |  | 8.68 | 8.99 |
| 5 | S5 | H |  |  | 6.25 | 6.08 | 32[p] | S1 |  |  |  | 6.28 | 6.09 |
| 6[p] | S1 | H |  |  | 7.14 | 7.36 | 33 | S3 |  | -OCH$_3$ |  | 7.57 | 7.59 |
| 7 | S4 | H |  |  | 6.87 | 6.88 | 34 | S5 |  |  |  | 8.05 | 8.23 |
| 8 | S2 | H |  |  | 7.00 | 7.36 | 35 | S4 |  | -OCH$_2$CH$_3$ |  | 8.26 | 8.29 |
| 9 | S3 | H |  |  | 7.49 | 7.41 | 36 | S1 |  |  |  | 6.96 | 6.78 |
| 10[p] | S2 | H |  |  | 6.98 | 7.30 | 37[p] | S4 |  |  |  | 6.45 | 6.21 |
| 11 | S5 | H |  |  | 6.78 | 7.28 | 38 | S2 |  | -OCH$_2$CH$_3$ |  | 9.05 | 9.47 |
| 12 | S1 | H |  |  | 8.24 | 8.54 | 39 | S5 |  | -OCH$_3$ |  | 8.68 | 9.07 |
| 13 | S2 | H |  |  | 6.92 | 7.33 | 40 | S3 |  |  |  | 8.19 | 8.16 |
| 14 | S3 |  |  |  | 8.92 | 8.88 | 41 | S5 |  | -OCHF$_2$ |  | 8.02 | 7.94 |
| 15 | S2 | |  |  | 6.24 | 6.16 | 42 | S1 |  | -OCH$_2$CHF$_2$ |  | 8.16 | 8.45 |
| 16 | S3 | |  |  | 6.99 | 6.76 | 43[p] | S2 |  | -OCH$_3$ |  | 6.85 | 7.32 |
| 17 | S5 | |  |  | 7.04 | 7.01 | 44 | S5 |  |  |  | 8.15 | 8.43 |
| 18 | S1 | |  |  | 8.27 | 7.98 | 45 | S3 |  |  |  | 8.28 | 8.51 |
| 19 | S4 | |  |  | 8.00 | 7.89 | 46[p] | S4 |  | -OCH$_3$ |  | 8.57 | 8.86 |
| 20[p] | S3 |  |  |  | 8.92 | 9.18 | 47 | S1 |  | -OCH$_2$CH$_3$ |  | 8.21 | 8.32 |
| 21 | S1 | |  |  | 8.52 | 8.32 | 48 | S3 |  | -OCH$_3$ |  | 8.48 | 8.93 |
| 22 | S4 |  |  |  | 6.29 | 6.61 | 49 | S2 |  | -OCH$_2$CH$_3$ |  | 8.62 | 8.75 |
| 23 | S3 |  |  |  | 6.87 | 6.46 | 50 | S2 |  |  |  | 8.25 | 8.58 |
| 24[p] | S2 |  |  |  | 8.09 | 8.54 | 51[p] | S5 |  | -OCH$_3$ |  | 8.10 | 7.71 |
| 25 | S5 | |  |  | 8.05 | 7.59 | 52 | S1 |  | -OCH$_2$CH$_3$ |  | 8.38 | 8.46 |
| 26 | S4 | |  |  | 8.92 | 9.03 | 53 | S4 |  |  |  | 8.19 | 8.36 |
| 27 | S1 | |  |  | 6.83 | 6.77 | | | | | | | |

[a] S1–S5 subsets

[b] Prediction set in MARS

tor selection methods provide one solution which may fall into a local minima and miss the global minima. Despite these methods, GA is an evolutionary stochastic method that can find the population of solutions. In each population, the chromosome of length *k* (the number of features) consists of zeros and ones indicating selected descriptors. The population in each generation is the result of the genetic manipulation of the chromosome found in the previous population through one-point crossover and mutation based on their fitness scores. This cyclic process was performed to obtain the best fitness

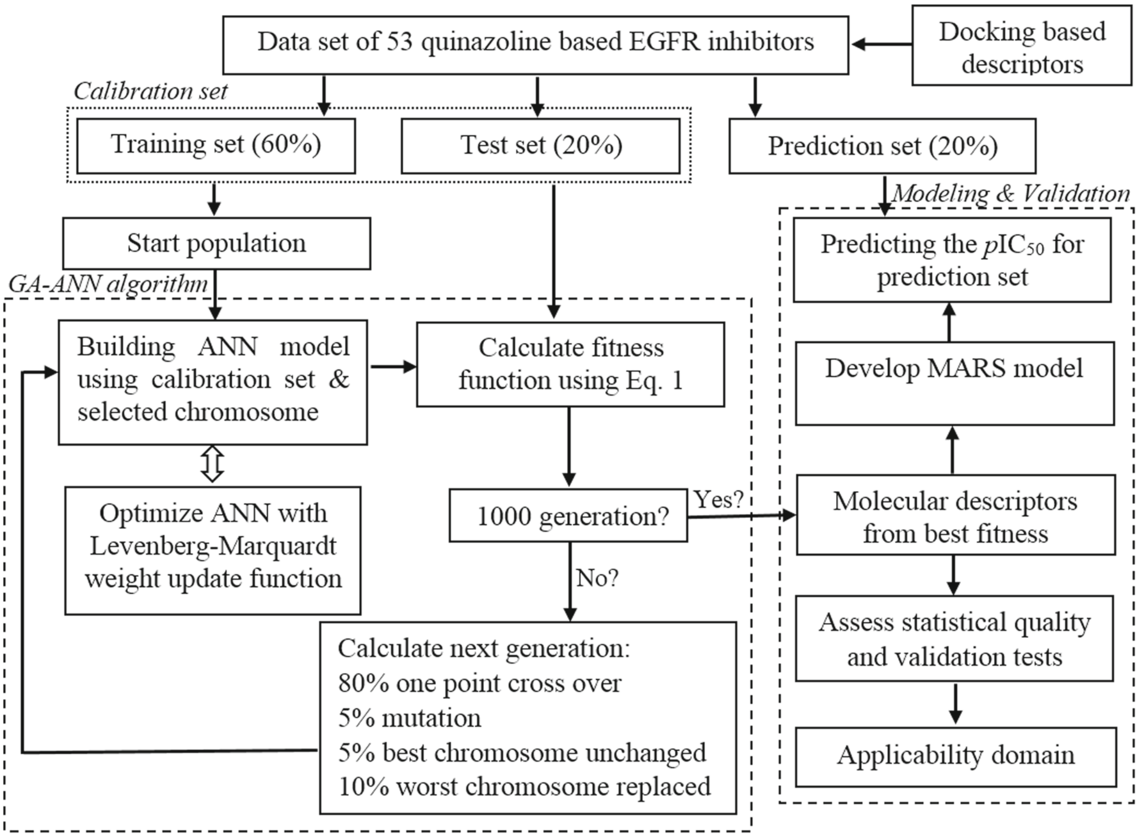

Data set of 53 quinazoline based EGFR inhibitors
Docking based descriptors
Calibration set
Training set (60%)
Test set (20%)
Prediction set (20%)
Modeling & Validation
Start population
GA-ANN algorithm
Building ANN model using calibration set & selected chromosome
Calculate fitness function using Eq. 1
Optimize ANN with Levenberg-Marquardt weight update function
1000 generation?
Yes?
No?
Calculate next generation:
80% one point cross over
5% mutation
5% best chromosome unchanged
10% worst chromosome replaced
Predicting the pIC50 for prediction set
Develop MARS model
Molecular descriptors from best fitness
Assess statistical quality and validation tests
Applicability domain

**Fig. 1** Flowchart of our GA-ANN/MARS strategy together with validations used in the present QSAR study

score or satisfactory result. In the present study, artificial neural network (ANN) technique is defined as the fitness function in the GA due to its ability to measure a non-linear dependency between the descriptors (in the chromosome) and target variable ($p\text{IC}_{50}$). To calculate the fitness score, a multi-objective fitness function was used to account for both the residual errors and the number of selected descriptors according to the following equation:

$$\text{Fitness score} = \frac{1}{\text{RMSE}_{\text{Train}} + \text{RMSE}_{\text{Test}} + \sqrt{\text{k}}}, \quad (1)$$

where $\text{RMSE}_{\text{Train}}$ and $\text{RMSE}_{\text{Test}}$ are the root mean square errors of the train and test sets, respectively, and k is the number of selected descriptors in the corresponding chromosome. Previous reports of GA-ANN strategies [14] were modified for effective descriptor selection as shown in Fig. 1. It should be mentioned that the Levenberg–Marquardt weight update function [15] was used in ANN due to its fast and remarkable performance. For more accurate descriptor selection, the dataset was divided into five subsets (S1, S2, S3, S4, and S5) as shown in Table 1, and a variable selection process was performed for different combinations of these subsets. Four subsets were used for calibration (train and test subsets) in developing the model and one subset was used for predicting and evaluating the generated model. The ANN models were developed using the train set, while the test set was used to prevent over-fitting. Ten possible combinations of train, test, and prediction subsets were used in the present study for descriptor selection. The most frequent selected descriptors were chosen as the most important descriptors describing the variance of the inhibitory activities.

For modeling purpose, MARS was employed as a simple and interpretable technique [16]. The main idea of MARS is dividing the whole space of each independent descriptor into various sub-regions and then defining a different mathematical equation for each region. This makes the MARS regression technique more flexible than the other regression techniques. For each descriptor, some break points (knot points) are defined using right and left side splines. These two left and right spline functions (basis functions) are used for relating the independent descriptors to the dependent variable. Generalized cross-validation (GCV) was employed to determine the importance of each basis function. Moreover, the MARS algorithm can be used to describe the non-linearity with little data preparation. The theory of this algorithm has been described elsewhere [9]. MATLAB Version 11.1 (Mathwork) [17] was used to run the GA-ANN algorithm and MARS toolbox.

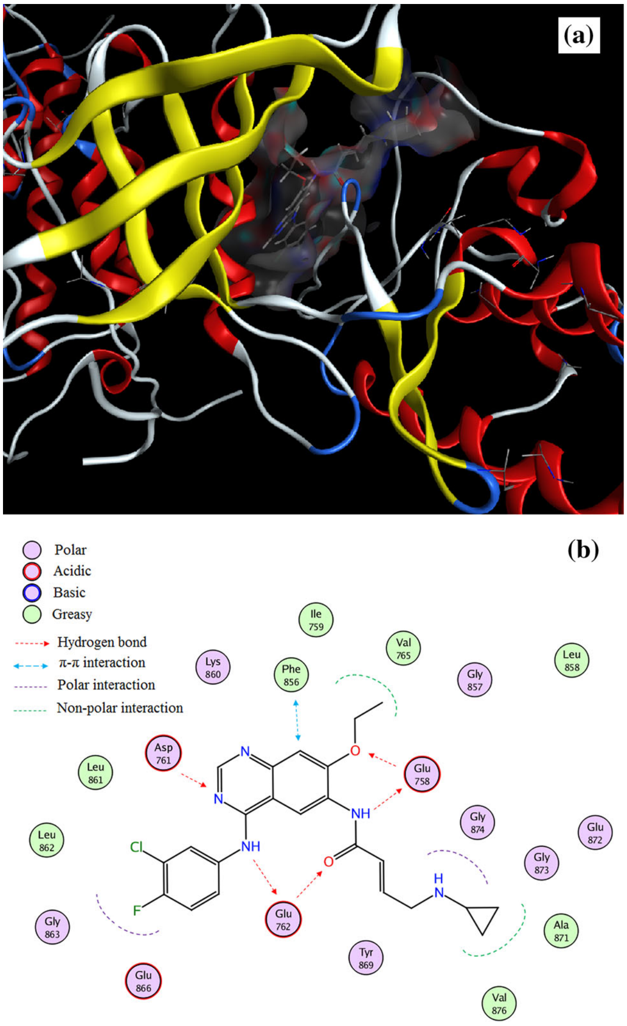


**Fig. 2** **a** Docking pose of the most potent inhibitor (compound **38**) in the EGFR-T790M active site. **b** 2D interaction of receptor compound **38** and surrounding amino acid residues

## Results and discussion

As a necessary step in drug design, molecular docking studies can be used to identify receptor–ligand interactions in a protein's binding pocket. The inhibitors listed in Table 1 were docked into the EGFR receptor using MOE. The most important interactions of EGFR with the most potent inhibitor (compound **38**) from our molecular docking results are shown in Fig. 2a. A graphical receptor–ligand interaction is shown in Fig. 2b. As this figure shows, there are two hydrogen bonds between Glu762 and the inhibitor (compound **38**). In addition, the amide group and ethyl ether also form two hydrogen bonds with Glu758. Another important hydrogen bond is between Asp761 and the N atom in the quinazoline ring. The inhibitor approaches residues Ala871 and Val876 from the α-face and makes direct non-polar contacts (van der Waals interaction); however, residues Tyr869, Gly873, Gly874, and Glu872 make polar contacts with the inhibitor. The two quinazoline rings can make a $\pi - \pi$ interaction with Phe856. The polar interactions between the halide atoms (F & Cl) and residues Gly863 and Glu866 have been detected in docking results as shown in Fig. 2b. Lastly, the optimized conformation for each inhibitor was obtained from the receptor–ligand docked results. The obtained conformations of the inhibitors were used for molecular descriptor generation and further QSAR studies.

In the descriptor selection procedure, a GA-ANN strategy was used with a population of 30 chromosomes for 1000 generations. All parts of this process are illustrated in Fig. 1. As mentioned before, the dataset was randomly split into five subsets (S1, S2, S3, S4, and S5) in such a way that they consisted of all range of $p\text{IC}_{50}$ values considering chemical diversity. A diversity analysis was performed as illustrated in Fig. 3 to show that the inhibitors in each subset (S1–S5) can represent the whole data. In this figure, the mean distances of inhibitors were plotted against $p\text{IC}_{50}$ illustrating the diversity of the inhibitors in the five subsets. The descriptor selection and model generation will be more stable using the diversity analysis. For ten possible combinations of subsets (10 runs), the whole GA-ANN process was carried out using different calibration (train and test sets) and prediction sets. The statistical results of the calibration and prediction sets are shown in Table 2 for ten runs using the GA-ANN algorithm. The average values of $R^2_{\text{Cal}}$, $R^2_{\text{p}}$, $\text{RMSE}_{\text{Cal}}$, and $\text{RMSE}_{\text{p}}$ are 0.793, 0.735, 0.438, and 0.544, respectively, which indicate that the performance of the GA-ANN is acceptable. The selected descriptors in each run and the repetition frequency for each descriptor in the GA-ANN are shown in Table 2. Each descriptor was at least repeated in seven runs from ten runs, which shows that the descriptor selection process is consistent and reliable. Inspection of this table also reveals that the five RDF030p descriptors (RDF descriptor, Radial Distribution Function weighted by polarizability), G2s (WHIM descriptor, 2nd component symmetry directional WHIM index weighted by electrotopological state), Mor04v (3D-MoRSE descriptor, 3D-MoRSE weighted by van der Waals volume), MATS2e (2D autocorrelation descriptor, Moran autocorrelation weighted by electronegativity) and RDF055m (RDF descriptor, Radial Distribution Function weighted by mass) appeared more frequently in 10 runs compared to the other descriptors.

The information about the selected descriptors is easily accessible from the $x$, $y$, $z$ coordinates of the inhibitor atoms and other quantities attached to these coordinates, such as molecular volume, electronegativity, polarizability, covalent radii, interatomic distances, and distances from a specified origin. They reflect aspects related to molecular size, shape, charge, and the steric interactions. Hence, they encode different aspects of their inhibition mechanism. RDF030p and RDF055m are the radial distribution functions on a spheri-

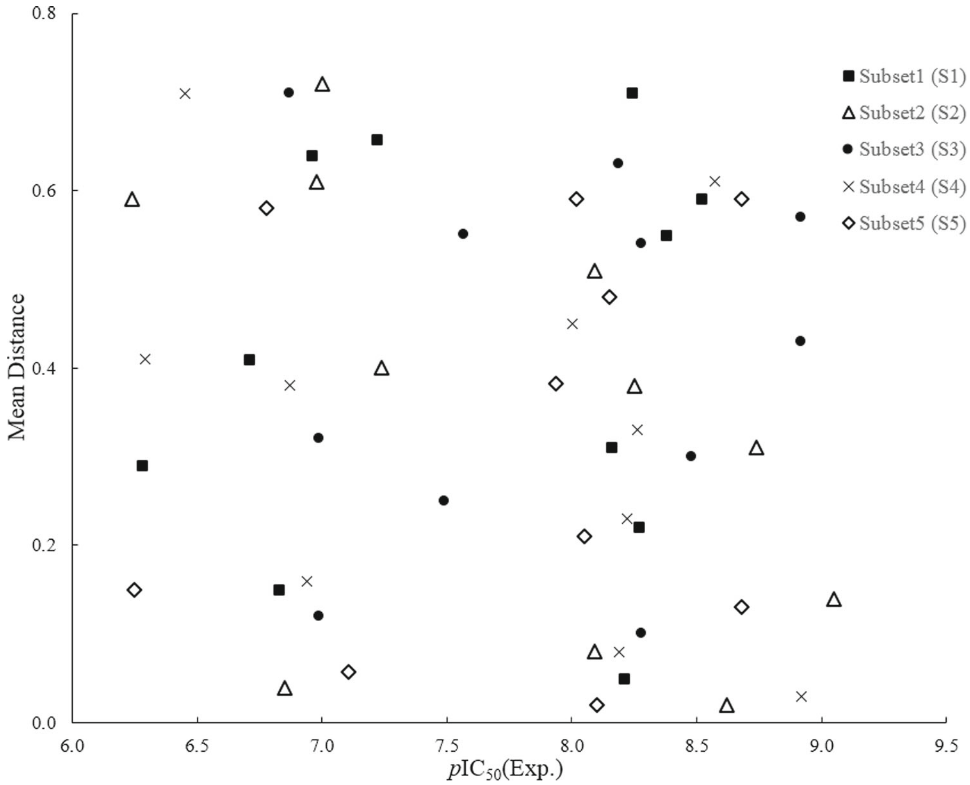


**Fig. 3** Scatter plot of five subsets (S1–S5) obtained from diversity analysis

cal volume of radius of 3.0 and 5.5 angstroms weighted by atomic polarizability and atomic mass, respectively. These descriptors are based on geometrical interatomic distance and provide valuable information about interatomic distances and ring types [18]. G2s is the geometric mean of the directional symmetries and it estimates dispersion and distribution of molecular electronic and topological state properties around the second component symmetry geometric center. This WHIM descriptor is built in such a way as to capture relevant molecular 3D information regarding electrotopological state. It should be mentioned that the electrotopological state is a measure of the electronic accessibility and can be interpreted as a probability of interaction with another molecule [19]. The Mor04v descriptor is among the 3D-MoRSE descriptors. In order to calculate this descriptor, the sum of all atomic van der Waals volumes is calculated using different angular scattering functions. The Mor04v descriptor shows the three-dimensional arrangement of the atoms influencing the size and shape of a molecule [20]. The MATS2e descriptor can be calculated using the Moran algorithm by summing up the products of atomic electronegativity of the terminal atoms in molecular structural graph. This descriptor encodes information of adjacency and distance of atoms and also describes how electronegativity is distributed along the topological structure of molecules [21]. To ensure that the selected descriptors encode different aspects of inhibitors, the correlation between the selected descriptors is shown in Table 3. According to this table, there are no significant correlations between these descriptors except for the RDF030p and RDF055m descriptors (correlation = 0.582).

The five selected descriptors by the GA-ANN algorithm were used as inputs for developing the final MARS model to predict the $p\mathrm{IC}_{50}$ values of the fifty-three quinazoline-based EGFR inhibitors. Each descriptor in the MARS model has one or two basis functions. The coefficients, importance, and $P$ values of these basis functions are given in Table 4. Each basis function in the MARS model is a regression equation with specific coefficients. One can easily calculate the relationship between the selected molecular descriptors and the inhibitory concentration of the quinazoline inhibitors using these basis functions. The importance of basis functions can be evaluated using the GCV parameter as shown in Table 4. The ranking order of the GCV values for the selected descriptors is RDF030p > MATS2e > Mor04v > RDF055m > G2s > G2s-RDF055m. The five selected descriptors, together with G2s-RDF055m interactions, have the highest effects on $p\mathrm{IC}_{50}$ values. As Table 4 shows, all descriptors are statistically significant with $P$ values less than 0.05. The performance of the GA-ANN/MARS strategy was evaluated by plotting the predicted $p\mathrm{IC}_{50}$ versus experimental $p\mathrm{IC}_{50}$ for calibration and prediction sets as illustrated in Fig. 4. The predicted values of $p\mathrm{IC}_{50}$ by the GA-ANN/MARS strategy are listed in Table 1. This table,

**Table 2** Results of ten runs and the descriptors selected in each run using GA-ANN strategy

| Run[a] | Train subset[b] | Test subset | $R^2_{\mathrm{Cal}}$ | $\mathrm{RMSE}_{\mathrm{Cal}}$ | Pred. subset | $R^2_{\mathrm{P}}$ | $\mathrm{RMSE}_{\mathrm{P}}$ | Selected descriptors[c] |
|---|---|---|---|---|---|---|---|---|
| 1 | S1 + S2 + S3 | S4 | 0.823 | 0.420 | S5 | 0.748 | 0.503 | RDF030p, G2s, MATS2e, RDF055m, R2e |
| 2 | S1 + S2 + S4 | S5 | 0.780 | 0.447 | S3 | 0.718 | 0.611 | RDF030p, Mor04v, X2sol, MATS2e, RDF055m |
| 3 | S1 + S2 + S5 | S3 | 0.749 | 0.453 | S4 | 0.716 | 0.584 | RDF030p, H3p, ICR, G2s, MATS2e, RDF055m |
| 4 | S1 + S3 + S4 | S5 | 0.817 | 0.417 | S2 | 0.721 | 0.575 | RDF030p,G2s, Mor04v, MATS2e, RDF055m |
| 5 | S1 + S3 + S5 | S2 | 0.764 | 0.453 | S4 | 0.743 | 0.523 | RDF030p, De, G2s, Mor04v, MSD, MATS2e |
| 6 | S1 + S4 + S5 | S3 | 0.839 | 0.402 | S2 | 0.776 | 0.481 | RDF030p, G2s, Mor04v, MATS2e, RDF055m |
| 7 | S2 + S3 + S4 | S1 | 0.778 | 0.448 | S5 | 0.737 | 0.523 | RDF030p, G2s, MATS2e, More04v, S2K |
| 8 | S2 + S3 + S5 | S4 | 0.752 | 0.487 | S1 | 0.703 | 0.617 | G2s, E2v, Mor04v, P2m, MATS2e, RDF055m |
| 9 | S2 + S4 + S5 | S1 | 0.824 | 0.416 | S3 | 0.754 | 0.486 | RDF030p, Mor04v, MATS2e, RDF055m, FDI |
| 10 | S3 + S4 + S5 | S2 | 0.804 | 0.438 | S1 | 0.735 | 0.541 | RDF030p, G2s, Mor04v, MATS2e, LDip, R4u |
| Mean | | | 0.793 | 0.438 | | 0.735 | 0.544 | |

[a] Ten runs were performed using different subsets S1–S5 from Table 1
[b] Calibration set contains train and test sets
[c] The repetition frequencies of RDF030p, G2s, Mor04v, MATS2e, and RDF055m descriptors in ten runs are 9, 8, 8, 7, and 7, respectively

**Table 3** Correlation matrix of descriptors

| Descriptors | RDF030p | G2s | Mor04v | MATS2e | RDF055m |
|---|---|---|---|---|---|
| RDF030p | 1 | 0.145 | 0.163 | −0.074 | 0.582 |
| G2s | | 1 | 0.279 | 0.210 | 0.083 |
| Mor04v | | | 1 | −0.238 | 0.262 |
| MATS2e | | | | 1 | −0.388 |
| RDF055m | | | | | 1 |

**Table 4** Coefficients of basis functions for each descriptor in MARS model

| Parameters | RDF030p[b] | G2s | Mor04v | MATS2e | RDF055m | G2s-RDF055m |
|---|---|---|---|---|---|---|
| Right-side Eq.[a] | 0.43 [X1-8.42] | 3.33 [X2-0.47] | 1.81 [X3-1.04] | −3.20 [X4-0.29] | 0.31 [X5-5.08] | – |
| Left-side Eq.[a] | 0.25 [8.42-X1] | 4.35 [0.47-X2] | 1.47 [1.04-X3] | −2.59 [0.29-X4] | – | 4.91 [0.38-X2]<br>0.22 [4.06-X5] |
| Basis functions | 2 | 2 | 2 | 2 | 1 | 1 |
| *P* value[b] | <0.01 | <0.01 | <0.01 | <0.01 | 0.017 | 0.026 |
| GCV[c] | 8.43 | 3.52 | 4.24 | 4.74 | 3.91 | 2.63 |

[a] Descriptors of RDF030p, G2s, Mor04v, MATS2e, and RDF055m were shown as X1, X2, X3, X4, and X5, respectively
[b] Statistical significance test was performed using ANOVA
[c] The importance of descriptors as generalized cross-validation values

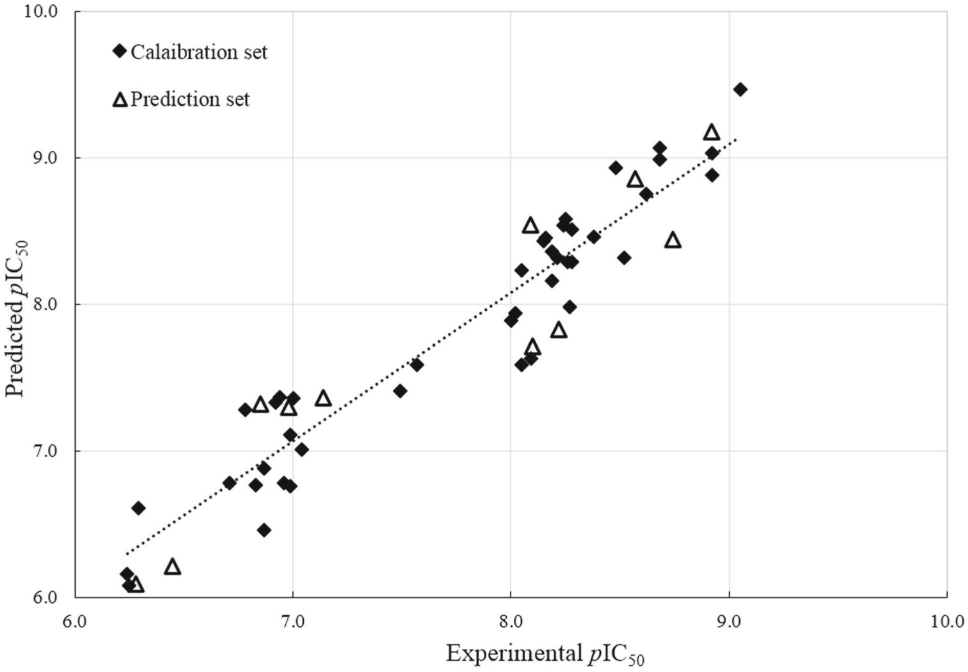


**Fig. 4** GA-ANN/MARS plot of calculated versus experimental $p\mathrm{IC}_{50}$ values for the calibration and prediction sets

**Table 5** Statistical results of QSAR models

| Model | Calibration | | Prediction | | LOO C.V. | | LSO C.V. | | Y-Scrambling[a] | | $r^2_{\mathrm{m}}$ | $F$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $R^2$ | RMSE | $R^2_{\mathrm{p}}$ | RMSE | $Q^2$ | RMSE | $R^2$ | RMSE | $R^2$ | RMSE | | |
| GA-ANN/MARS | 0.917 | 0.266 | 0.881 | 0.332 | 0.923 | 0.253 | 0.828 | 0.436 | 0.275 | 3.629 | 0.772 | 73.6 |
| GA-PLS/MARS[b] | 0.809 | 0.497 | 0.705 | 0.628 | 0.724 | 0.520 | 0.542 | 1.306 | 0.315 | 2.482 | 0.625 | 46.1 |

[a] LSO was repeated 200 cycles with random 20 % subset selection
[b] Selected descriptors: RDF030p, H3v, G2s, MATS2e, RDF060m, and IDDE

together with Fig. 4, shows that the predicted $p\mathrm{IC}_{50}$ is in agreement with the experimental $p\mathrm{IC}_{50}$.

The statistical performance and validation of the QSAR models were judged by means of (a) the external prediction set; (b) internal validation by LOO cross-validation; (c) complementary cross-validation by LSO procedure; (d) examining if the following conditions are satisfied: (1) $R^2_{\mathrm{p}} > 0.6$, (2) $Q^2_{\mathrm{LOO}} > 0.5$, (3) $r^2_{\mathrm{m}} > 0.5$, (4) $r^2_0$ (i.e., imply regression through the origin) is close to $R^2_{\mathrm{p}}$ such that $[(R^2_{\mathrm{p}} - r^2_0)/R^2_{\mathrm{p}}] < 0.1$; and (e) Y-scrambling [22,23]. The results of these assessment procedures are given in Table 5 for the GA-ANN/MARS strategy. It should be noted that the external prediction set was not used in any step of the model development. The RMSE (=0.332) and $R^2$ (=0.881) values for the prediction set indicate that 88.1 % of the variance in the inhibitory concentration of the quinazolines can be explained by the QSAR model. The internal validation was conducted by the LOO technique to ensure the relia-

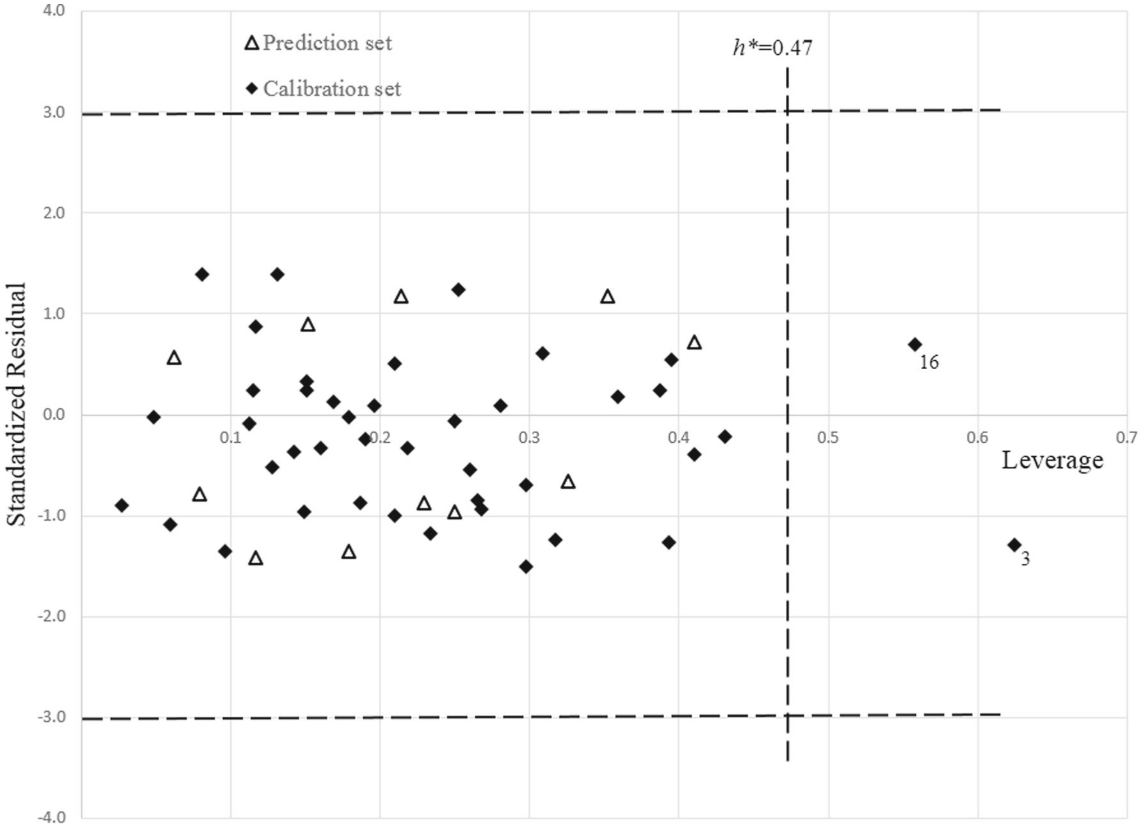
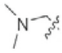
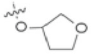
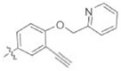
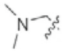
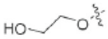
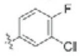
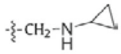
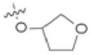
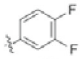
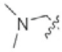
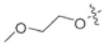
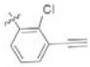
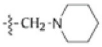
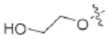
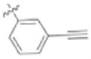
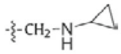
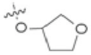
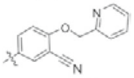
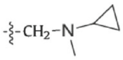
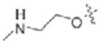
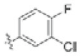
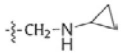
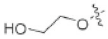
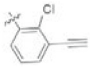


**Fig. 5** Applicability domain plot for the calibration and prediction sets

bility of the QSAR model. According to Table 5, it can be noticed that the $Q^2_{\mathrm{LOO}}$ (=0.923) and $\mathrm{RMSE_{LOO}}$ (=0.253) values are comparable. In the case of LSO, 80 % of the data were used to develop the QSAR model and the remaining 20 % of the data were used to evaluate the generated model. This repetitive process was performed 500 times with random subsets selected. The overall mean for this process ($R^2_{\mathrm{LSO}} = 0.828$ and $\mathrm{RMSE_{LSO}} = 0.436$) indicates that the constructed model is stable and valid. Other statistical parameters ($[(R^2_{\mathrm{p}} - r^2_0)/R^2_{\mathrm{p}}] = 0.017$ and $r^2_{\mathrm{m}} = 0.772$) have acceptable values for the QSAR models. These results suggest that both interpolations and extrapolations of the inhibitory concentrations in the QSAR model are reasonably adequate. Y-scrambling was performed by the analysis of 200 cases of repetitive randomization of the response data ($p\mathrm{IC}_{50}$). The low $R^2_{\mathrm{S}}$ (=0.275) (scrambled results) and high $\mathrm{RMSE_S}$ (=3.629) indicate that the QSAR model is not due to chance correlation. For the sake of comparison, a GA-PLS/MARS technique [24] was also used to develop the QSAR model. The statistical results of the GA-ANN/MARS strategy are better than those of the GA-PLS/MARS technique, especially the results of $\mathrm{R^2_p}$ and $\mathrm{Q^2_{LOO}}$ (Table 5). It is clear that the RMSE of the prediction set has been reduced by about 50 % using a GA-ANN/MARS strategy. Furthermore, the statistical parameters of $R^2_{\mathrm{p}}$, $F$, $[(R^2_{\mathrm{p}} - r^2_0)/R^2_{\mathrm{p}}]$, and $r^2_{\mathrm{m}}$ have also been improved using this strategy.

In this work, we use a leverage approach for the applicability domain definition [25]. The applicability domain is a defined squared area that depends on the descriptor values and the dataset. A Williams plot of standardized residuals vs. leverage values ($h$) was used for an immediate and simple graphical detection of the response outliers. The applicability domain is a measure to validate the reliability of the GA-ANN/MARS strategy for future predictions. The prediction is valid if $h < h^*$ (critical value: $h^* = 3$ p/n, where p is the number of descriptors in the model plus one and n is the number of inhibitors in the calibration set). The five selected descriptors in this work were used to calculate the leverage values. In the Williams graph (Fig. 5), the standardized residuals are plotted against the leverage value ($h$). Here, no compound has been identified as an outlier in prediction set and just two compounds in the calibration set (compounds **3** and **16**) have $h$ values higher than $h^*$(=0.47). Such compounds are somewhat structurally different from the other inhibitors (with substituted amide group at the 4-position of phenyl group in R3 branch as shown in Table 1). Therefore, they are considered as good leverage points since the information that these two compounds encode makes the QSAR

**Table 6** Chemical structures of newly designed EGFR-T790M inhibitors based on QSAR models

| No. | R1 | R2 | R3 | $p\mathrm{IC}_{50}$ | $h_i$ |
|---|---|---|---|---|---|
| 1 |  |  |  | 8.85 | 0.32 |
| 2 |  |  |  | 9.02 | 0.17 |
| 3 |  |  |  | 8.73 | 0.27 |
| 4 |  |  |  | 8.48 | 0.11 |
| 5 |  |  |  | 8.93 | 0.08 |
| 6 |  |  |  | 9.42 | 0.41 |
| 7 |  |  |  | 9.21 | 0.17 |
| 8 |  |  |  | 8.78 | 0.18 |

model more precise. Other inhibitors in the calibration and prediction sets fall within the applicability domain.

Finally, several new inhibitors were proposed to show the practical application of the developed models. The $p\mathrm{IC}_{50}$ values of the new design inhibitors were predicted using the established QSAR models. Compounds **14**, **20**, **26**, and **38**, with the highest inhibitory concentrations, were considered as reference chemicals to design new inhibitors. The R1, R3, and R2 branches in the dataset were modified to propose new inhibitors. The predicted $p\mathrm{IC}_{50}$ values of eight promising new inhibitors, together with their structures, are shown in Table 6. The leverage values (*hi*) of the proposed inhibitors are lower than the critical value ($h^*$), which shows that the predicted $p\mathrm{IC}_{50}$ values are acceptable. These results reveal that the proposed QSAR strategy is good enough to be considered as an alternative way for evaluating new potent EGFR inhibitors.

## Conclusions

In the present study, a non-linear descriptor selection strategy was employed to generate QSAR models to predict the inhibitory concentration of fifty-three newly synthesized quinazoline-based EGFR-T790M inhibitors. Conventional descriptor selection methods are relatively inaccurate and they only propose one solution without considering non-linearity. Therefore, a GA-ANN strategy was implemented as the non-linear descriptor selection technique with different subset combinations. A MARS strategy was used to define sub-region equations for each descriptor. Various methods were also used to validate the QSAR models including internal and external cross-validation methods. The ranking of the five molecular descriptors obtained from GA-ANN/MARS strategy is RDF030p > MATS2e > Mor04v > RDF055m > G2s > G2s-RDF055m. Furthermore, a molecular docking analysis revealed important receptor–ligand interactions for the most active inhibitor. Finally, the molecular information obtained from docking and QSAR analysis was employed to propose new inhibitors. The results of this study provide a guideline about the structural features of EGFR-T790M inhibitors, which can be used to design novel inhibitors and predict the inhibitory activity of non-examined chemicals.

### Compliance with ethical standards

**Conflict of interest** The author declares that there are no conflicts of interest.

## References

1. Wang C, Gao H, Dong J, Wang F, Li P, Zhang J (2014) Insight into the medicinal chemistry of EGFR and HER-2 inhibitors. Curr Med Chem 21:1336–1350. doi:10.2174/09298673206661311119124646
2. Chaft JE, Oxnard GR, Sima CS, Kris MG, Miller VA, Riely GJ (2011) Disease flare after tyrosine kinase inhibitor discontinuation in patients with EGFR-mutant lung cancer and acquired resistance to erlotinib or gefitinib: implications for clinical trial design. Clin Cancer Res 17:6298–6303. doi:10.1158/1078-0432.CCR-11-1468
3. Suda K, Onozato R, Yatabe Y, Mitsudomi T (2009) EGFR T790M mutation: a double role in lung cancer cell survival? J Thorac Oncol 4:1–4. doi:10.1097/JTO.0b013e3181913c9f
4. Lacouture ME (2006) Mechanisms of cutaneous toxicities to EGFR inhibitors. Nat Rev Cancer 6:803–812. doi:10.1038/nrc1970
5. Choong NW, Dietrich S, Seiwert TY, Tretiakova MS, Nallasura V, Davies GC, Lipkowitz S, Husain AN, Salgia R, Ma PC (2006) Gefitinib response of erlotinib-refractory lung cancer involving meninges role of EGFR mutation. Nat Clin Pract Oncol 3:50–57. doi:10.1038/ncponc0400

6. Zhang L, Yang Y, Zhou H, Zheng Q, Li Y, Zheng S, Zhao S, Chen D, Fan C (2015) Structure-activity study of quinazoline derivatives leading to the discovery of potent EGFR-T790M inhibitors. Eur J Med Chem 102:445–463. doi:10.1016/j.ejmech.2015.08.026
7. Muratov EN, Varlamova EV, Artemenko AG, Polishchuk PG, Kuz'min VE (2012) Existing and developing approaches for QSAR analysis of mixtures. Mol Inform 31:202–221. doi:10.1002/minf.201100129
8. Kovalishyn V, Poda G (2015) Efficient variable selection batch pruning algorithm for artificial neural networks. Chemom Intell Lab Syst 149:10–16. doi:10.1016/j.chemolab.2015.10.005
9. Kooperberg C (2006) Multivariate adaptive regression splines, in encyclopedia of environmetrics. Wiley, Chichester
10. Heinzerling L, Klein R, Rarey M (2012) Fast force field-based optimization of protein-ligand complexes with graphics processor. J Comput Chem 33:2554–2565. doi:10.1002/jcc.23094
11. VCCLAB, Virtual Computational Chemistry Laboratory, Milan. http://www.vcclab.org/lab/edragon/
12. Todeschini R, Consonni V, Mannhold R, Kubinyi H, Folkers G (2009) Molecular descriptors for chemoinformatics, vol 41. Wiley, New York
13. Asadollahi-Baboli M, Mani-Varnosfaderani A (2015) Therapeutic index modeling and predictive QSAR of novel thiazolidin-4-one analogs against *Toxoplasma gondii*. Eur J Pharm Sci 70:117–124. doi:10.1016/j.ejps.2015.01.014
14. Jalali-Heravi M, Asadollahi-Baboli M (2008) QSAR analysis of platelet-derived growth inhibitors using GA-ANN and shuffling crossvalidation. QSAR Comb Sci 27:750–757. doi:10.1002/qsar.200710138
15. Yamashita N, Fukushima M (2001) On the rate of convergence of the Levenberg-Marquardt Method. In: Alefeld G, Chen X (eds) Topics in numerical analysis. Springer, Vienna, pp 239–249
16. Friedman JH (1991) Multivariate adaptive regression splines. Ann Stat 19:1–67
17. MATLAB 11.1. The Mathworks, Inc., Natick. http://www.mathworks.com/
18. Von Lilienfeld OA, Ramakrishnan R, Rupp M, Knoll A (2015) Fourier series of atomic radial distribution functions: a molecular fingerprint for machine learning models of quantum chemical properties. Int J Quantum Chem 115:1084–1093. doi:10.1002/qua.24912
19. Todeschini R, Gramatica P (1998) New 3D molecular descriptors: the WHIM theory and QSAR applications. In: Kubinyi H, Folkers G, Martin YC (eds) 3D QSAR in drug design. Springer, Dordrecht
20. Devinyak O, Havrylyuk D, Lesyk R (2014) 3D-MoRSE descriptors explained. J Mol Graph Modell 54:194–203. doi:10.1016/j.jmgm.2014.10.006
21. Saíz-Urra L, Pérez González M, Teijeira M (2007) 2D-autocorrelation descriptors for predicting cytotoxicity of naphthoquinone ester derivatives against oral human epidermoid carcinoma. Biorg Med Chem 15:3565–3571. doi:10.1016/j.bmc.2007.02.032
22. Gramatica P (2014) External Evaluation of QSAR Models, in Addition to Cross-Validation: Verification of Predictive Capability on Totally New Chemicals. Mol Inform 33:311–314. doi:10.1002/minf.201400030
23. Asadollahi-Baboli M (2012) Quantitative structure-activity relationship analysis of human neutrophil elastase inhibitors using shuffling classification and regression trees and adaptive neuro-fuzzy inference systems. SAR & QSAR Environ Res 23:505–520. doi:10.1080/1062936X.2012.696552
24. Sagrado S, Cronin MTD (2008) Application of the modelling power approach to variable subset selection for GA-PLS QSAR models. Anal Chim Acta 609:169–174. doi:10.1016/j.aca.2008.01.013
25. Roy K, Kar S, Ambure P (2015) On a simple approach for determining applicability domain of QSAR models. Chemom Intell Lab Syst 145:22–29. doi:10.1016/j.chemolab.2015.04.013