4D-QSAR Analyses for EGFR Inhibitors Based on CDDA-OPS-GA Method

Daogang Qin

Nanchang University

Xiaoqi Zeng

Jiangxi Provincial Drug Administration

Tiansheng Zhao

Nanchang University

Biyang Cai

Nanchang University

Bowen Yang

Nanchang University

Guogang Tu (✉ tugg199@yahoo.com)

Nanchang University <https://orcid.org/0000-0003-1264-9475>

Research Article

Keywords: 4D-QSAR, comparative distribution detection algorithm (CDDA), ordered predictors selection (OPS), genetic algorithm (GA), EGFR Inhibitors

Posted Date: December 6th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1120590/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Epidermal growth factor receptor is a preferred target for treating cancer. Compared to 3D-QSAR, 4D-QSAR has the feature of conformational flexibility and free alignment for individual ligands. In present studies, the 4D-QSAR of 131 analogs of 4-anilino quinazoline for EGFR inhibitors was built. The GROMACS package was employed to yield the conformational ensemble profile. The field descriptors of Coulomb and Lennard-Jones potentials were calculated by LQTA-QSAR. The filter descriptors and variable selection is very important, which was performed by means of comparative distribution detection algorithm (CDDA), ordered predictors selection (OPS) and genetic algorithm (GA) method. Best 4D-QSAR model yielded satisfactory statistics ($R^2 = 0.71$), good performance in internal ($Q^2_{LOO} = 0.60$) and external prediction ($R^2_{pred} = 0.69$, $k = 0.97$, $k' = 1.01$). The 4D-QSAR was shown to be robust ($Q^2_{LMO} = 0.59$) and was not built by chance ($R^2_{YS} = 0.17$, $Q^2_{YS} = -0.25$). The model has a good potential for rational design new EGFR inhibitors.

Introduction

Epidermal growth factor receptor (EGFR) is the prototype of receptor tyrosine kinases (TKs) family (He et al. 2016). As a transmembrane glycoprotein, several signal transduction cascades are initiated, and lead to DNA synthesis and cell proliferation when EGFR is activated (Wang et al. 2016b). So EGFR plays an important role in the regulation of several cellular functions such as survival, cell growth, differentiation, proliferation, and apoptosis (Yewale et al. 2013). The mutation or amplification of EGFR was found in solid tumors, such as lung cancer, glioma, breast cancer, and ovarian cancer. Currently, EGFR is a potential target for cancer therapy (Cho et al. 2016; Wang et al. 2016a; Imamura et al. 2016; Holdman et al. 2015). The EGFR inhibitor, lapatinib, is approved for the treatment of breast cancer by the FDA (Oda et al. 2005). Other EGFR inhibitors, such as lomustine, temozolomide, gefitinib, and erlotinib, have also been approved by the FDA for the treatment of glioma (Minkovsky and Berezov 2008). However, in many cancer types as breast cancer, hepatocellular carcinoma, pancreatic cancer, non-small cell lung cancer, colorectal carcinoma, glioblastoma, and melanoma, there are significant resistance to the used EGFR inhibitors (Burotto et al. 2014). All these findings make requires to design and synthesize new potent EGFR inhibitors.

Before the synthesis, it is necessary to develop a prediction method for biological activities. Quantitative structure-activity relationship (QSAR) is an important part for modern drug design, including risk assessment, drug discovery and predictive toxicology (Dearden 2017; Cherkasov et al. 2014). Initially, two-dimensional quantitative structure-activity relationship (2D-QSAR) and three-dimensional quantitative structure activity relationship (3D-QSAR) were extensively explored in medicinal chemistry study. However, the major constraint of 3D-QSAR is its dependency and sensitivity to conformations and alignments of compounds (Shim and Mackerell Jr 2011), because only one conformation, not a conformational ensemble profile, is considered for each compound (Ghasemi et al. 2011). To overcome inherent constraint of 3D-QSAR, the four-dimensional quantitative structure activity relationship (4D-

QSAR) was originally developed which includes the freedom of alignment and the conformational flexibility to build 3D-QSAR by performing molecular state ensemble averaging, i.e., the fourth “dimension” (Hopfinger et al. 1997).

The LQTA-QSAR (*Laboratório de Quimiometria Teórica e Aplicada*), a 4D-QSAR approach, often calculates a large number of descriptors (variable), frequently several thousands (Martins et al. 2009). Hence, the variable selection is very important when generating 4D-QSAR model. The comparative distribution detection algorithm (CDDA) can classify descriptors according to distribution profile, which compares individual distributions of a descriptor with dependent variables, and computes dissimilarity statistics (Barbosa and Ferreira 2012). CDDA enables numerical inspection of bivariate scatter plots and helps in filtering (selection) of descriptors which is suitable to establish 4D-QSAR model. The ordered predictors selection (OPS) is able to get a informative vector that contains information about the location of the best response variables for prediction. The OPS was shown to avoid overfitting and chance correlation, and be robust for QSAR model (Teofilo et al. 2009). The genetic algorithm (GA), a stochastic method, enables to solve optimization problems of fitness criteria, which applies different genetic functions and evolution hypothesis of Darwin, i.e. mutation and crossover (Hemmateenejad et al. 2002).

In present work, we aimed to build a 4D-QSAR model of the EGFR inhibitors by means of CDDA-OPS-GA method for descriptors selection. The regression methods used were multiple linear regression (MLR). It is the first to report the 4D-QSAR model of 4-anilino quinazoline derivatives as EGFR inhibitors.

Methods

Data set

All inhibitors of EGFR were taken from literature (Bridges et al. 1996; Li et al. 2012; Suzuki et al. 2011; Waterson et al. 2009). In order to provide numerically larger data values, the biological activities expressed as IC_{50} values in units of molarity were transformed to pIC_{50} ($-\log IC_{50}$) which were used as dependent variables. All of the compounds were divided into the training set of 105 compounds and the test set of 26 compounds taking into account both the distribution of dependent variables and the structural diversity. The training set was used to construct 4D-QSAR model, and the test set was used to evaluate the predictive quality. The chemical structures and IC_{50} values of the data set are presented in Fig. 1 and Tab. 1.

Table 1
Structures of training set and test set

Comp.	R ₁	R ₂	R ₃	X	Y	Z	pIC ₅₀
1	H	H	H	N	N	NH	6.46
2	H	2'-F	H	N	N	NH	7.25
3 ^a	H	2'-Cl	H	N	N	NH	7.64
4	H	2'-Br	H	N	N	NH	7.57
5	H	2'-I	H	N	N	NH	7.10
6	H	2'-CF ₃	H	N	N	NH	6.24
7	7-OMe	H	H	N	N	NH	6.92
8 ^a	7-OMe	2'-Br	H	N	N	NH	8.00
9	7-NH ₂	H	H	N	N	NH	7.00
10	7-NH ₂	2'-F	H	N	N	NH	8.70
11	7-NH ₂	2'-Cl	H	N	N	NH	9.60
12	7-NH ₂	2'- Br	H	N	N	NH	10.00
13 ^a	7-NH ₂	2'-I	H	N	N	NH	9.46
14	7-NH ₂	2'-CF ₃	H	N	N	NH	8.48
15	7-NO ₂	H	H	N	N	NH	4.92
16	7-NO ₂	2'-F	H	N	N	NH	5.21
17	7-NO ₂	2'-Cl	H	N	N	NH	6.09
18 ^a	7-NO ₂	2'-Br	H	N	N	NH	6.00
19	7-NO ₂	2'-I	H	N	N	NH	6.27
20	7-OH	2'-Br	H	N	N	NH	8.33
21	7-NHAc	2'-Br	H	N	N	NH	7.40
22	7-NHMe	2'-Br	H	N	N	NH	8.15
23 ^a	7-NHEt	2'-Br	H	N	N	NH	7.92

^a The test set compounds

Comp.	R ₁	R ₂	R ₃	X	Y	Z	pIC ₅₀
24	7-NMe ₂	2'-Br	H	N	N	NH	7.96
25	6-OMe	H	H	N	N	NH	7.26
26	6-OMe	2'-Br	H	N	N	NH	7.52
27 ^a	6-NH ₂	H	H	N	N	NH	6.11
28	6-NH ₂	2'-CF ₃	H	N	N	NH	6.24
29	6-NH ₂	2'-Br	H	N	N	NH	9.11
30	6-NO ₂	H	H	N	N	NH	5.30
31	6-NMe ₂	2'-Br	H	N	N	NH	7.08
32	6-NHCO ₂ Me	2'-Br	H	N	N	NH	7.92
33 ^a	6, 7-diOMe	H	H	N	N	NH	7.54
34	6, 7-diOMe	2'-F	H	N	N	NH	8.42
35	6, 7-diOMe	2'-Cl	H	N	N	NH	9.51
36	6, 7-diOMe	2'-Br	H	N	N	NH	10.60
37	6, 7-diOMe	2'-I	H	N	N	NH	9.05
38 ^a	6, 7-diOMe	2'-CF ₃	H	N	N	NH	9.62
39	6-NO ₂ , 7-NH ₂	2'-Br	H	N	N	NH	7.28
40	6-NO ₂ , 7-NHAc	2'-Br	H	N	N	NH	7.55
41	6-NO ₂ , 7-OMe	2'-Br	H	N	N	NH	7.82
42	6-NO ₂ , 7-Cl	2'-Br	H	N	N	NH	7.60
43 ^a	6-NH ₂ , 7-NHMe	2'-Br	H	N	N	NH	7.96
44	6-NH ₂ , 7-OMe	2'-Br	H	N	N	NH	8.42
45	6-NH ₂ , 7-Cl	2'-Br	H	N	N	NH	8.19
46	6, 7-diOH	2'-Br	H	N	N	NH	9.77
47	6, 7-diOEt	2'-Br	H	N	N	NH	11.22

^a The test set compounds

Comp.	R ₁	R ₂	R ₃	X	Y	Z	pIC ₅₀
48	6, 7-diOPr	2'-Br	H	N	N	NH	9.77
49 ^a	H	2'-OMe	H	N	N	NH	6.07
50	H	2'-Me	H	N	N	NH	6.04
51	5, 6-diOMe	2'-Br	H	N	N	NH	5.86
52	6, 7-diOMe	2'-Br	NH ₂	N	N	NH	6.33
53 ^a	5, 6, 7-triOMe	2'-Br	H	N	N	NH	9.17
54	6, 7-diOMe	1'-Br	H	N	N	NH	6.89
55	6, 7-diOMe	2',4'-diBr	H	N	N	NH	6.95
56	5-NO ₂	2'-Br	H	N	N	NH	6.45
57 ^a	8-NH ₂	2'-Br	H	N	N	NH	6.98
58	8-OMe	2'-Br	H	N	N	NH	6.01
59	6-NO ₂	2'-Br	H	N	N	NH	6.05
60	6-NHMe	2'-Br	H	N	N	NH	8.40
61	6, 7-diNH ₂	2'-Br	H	N	N	NH	9.92
62	6-NH ₂ , 7-NMe ₂	2'-Br	H	N	N	NH	6.80
63 ^a	6-NO ₂ , 7- NHMe	2'-Br	H	N	N	NH	7.17
64	6-NO ₂ , 7- NMe ₂	2'-Br	H	N	N	NH	5.70
65	6, 7-diOBu	2'-Br	H	N	N	NH	6.98
66	6, 7-diOMe	3'-Br	H	N	N	NH	9.02
67	6, 7-diOMe	2', 3'-diBr	H	N	N	NH	10.14
68 ^a	5-NH ₂	2'-Br	H	N	N	NH	6.36
69	6, 7-diOMe	2'-Br	H	N	N	NMe	6.82
70	H	H	H	N	N	NH(CH ₂) ₂	5.39
71	H	H	H	N	N	NHCH ₂	6.49

^a The test set compounds

Comp.	R ₁	R ₂	R ₃	X	Y	Z	pIC ₅₀
72 ^a	H	3'-OMe	H	N	N	NHCH ₂	5.00
73	H	H	H	N	N	NMe	4.00
74	5-NO ₂	H	H	N	N	NHCH ₂	5.10
75	5-OMe	H	H	N	N	NHCH ₂	5.31
76	6-OMe	H	H	N	N	NHCH ₂	6.70
77	7-NO ₂	H	H	N	N	NHCH ₂	5.23
78 ^a	7-OMe	H	H	N	N	NHCH ₂	7.24
79	6-OMe, 7-OH	H	H	N	N	NHCH ₂	6.23
80	6-OH, 7-OMe	H	H	N	N	NHCH ₂	7.25
81	H	3'-Cl	H	N	N	NHCH ₂	5.15
82	6-NO ₂	H	H	N	N	NHCH ₂	5.00
83 ^a	6-NH ₂	H	H	N	N	NHCH ₂	5.85
84	6, 7-diOMe	H	H	N	N	NHCH ₂	8.00
85	H	2'-Br	H	N	N	O	6.12
86	H	2'-Br	H	N	CH	NH	5.26
87	6, 7-diOH	2'-Br	H	N	C-CN	NH	6.46
88 ^a	6, 7-diOMe	2'-Br	H	N	C-CN	NH	6.72
89	6, 7-diOMe	2'-Cl	H	N	C-CN	NH	6.74
90	6, 7-diOMe	2'-CF ₃	H	N	C-CN	NH	5.84
91	6, 7-diOMe	2', 3'-diOMe	H	N	C-CN	NH	4.30
92	6, 7-diOMe	2'-F	H	N	C-CN	NH	6.13
93 ^a	6, 7-diOMe	3'-CN	H	N	C-CN	NH	4.85
94	6, 7-diOMe	3'-F	H	N	C-CN	NH	4.32
95	6, 7-diOMe	2'-COMe	H	N	C-CN	NH	5.13

^a The test set compounds

Comp.	R ₁	R ₂	R ₃	X	Y	Z	pIC ₅₀
96	6, 7-diOMe	2'-NH ₂	H	N	C-CN	NH	6.09
97	6, 7-diOMe	2'-N(Me) ₂	H	N	C-CN	NH	6.07
98 ^a	6, 7-diOMe	2'-NO ₂	H	N	C-CN	NH	6.06
99	6, 7-diOMe	2'-Cl, 3'-F	H	N	C-CN	NH	6.27
100	6, 7-diOMe	3'-Cl, 1'-F	H	N	C-CN	NH	4.86
101	6, 7-diOMe	2'-OH	H	N	C-CN	NH	5.20
102	6, 7-diOMe	3'-Me	H	N	C-CN	NH	5.56
103 ^a	6, 7-diOMe	2'-CONH ₂	H	N	C-CN	NH	5.11
104	6, 7-diOMe	2'-Br, 3'-Me	H	N	C-CN	NH	5.75
105	6, 7-diOMe	2'-Cl, 3'-OH	H	N	C-CN	NH	5.04
106	6, 7-diOMe	2', 4'-diCl, 3'-OH	H	N	C-CN	NH	5.50
107	6, 7-diOMe	1'-OH, 4'-Cl	H	N	C-CN	NH	4.77
108	6, 7-diOMe	2'-SMe	H	N	C-CN	NH	6.12
109 ^a	6, 7-diOMe	2'-Cl, 3'-Me	H	N	C-CN	NH	5.83
110	6, 7-diOMe	1'-F, 3'- Br	H	N	C-CN	NH	4.04
111	6, 7-diOMe	3'- CH(Me) ₂	H	N	C-CN	NH	6.64
112	6, 7-diOMe	1'- CH(Me) ₂	H	N	C-CN	NH	6.44
113 ^a	6, 7-diOMe	2'- CH(Me) ₂	H	N	C-CN	NH	4.21
114	6, 7-diOMe	3'-Br	H	N	C-CN	NH	5.18
115	6, 7-diOMe	1'-Br	H	N	C-CN	NH	5.38
116	6, 7-diOMe	2'-CF ₃ , 3'-F	H	N	C-CN	NH	6.80
117	6, 7-diOMe	1'-Me, 2'- Br	H	N	C-CN	NH	6.08
118	6, 7-diOMe	2'-Me, 3'- Br	H	N	C-CN	NH	6.21
119 ^a	6, 7-diOMe	2'- N ₃	H	N	C-CN	NH	5.66
120	6, 7-diOMe	2'- NHCOMe	H	N	C-CN	NH	4.16

^a The test set compounds

Comp.	R ₁	R ₂	R ₃	X	Y	Z	pIC ₅₀
121	6, 7-diOEt	2'-Br	H	N	C-CN	NH	6.38
122	6, 7-diOEt	2'-Cl, 3'-F	H	N	C-CN	NH	6.01
123	6-OEt, 7-OMe	2'-Br	H	N	C-CN	NH	4.67
124 ^a	6-OMe, 7-OEt	2'-Br	H	N	C-CN	NH	4.48
125	6, 7-diOMe	2'-Br	H	N	C-CO ₂ Et	NH	4.09
126	6, 7-diOMe	2'-Br	H	N	C-CH ₂ OH	NH	3.99
127	6, 7-diOMe	2'-Br	H	N	C-CHO	NH	5.47
128	6, 7-diOMe	2'-Br	H	N	C-CO ₂ H	NH	4.47
129 ^a	6, 7-diOMe	2'-Br	H	N	C-CONH ₂	NH	4.82
130	6, 7-diOMe	2'-Br	H	C-CN	N	NH	4.15
131	6, 7-diOMe	2'-Cl	H	C-CN	N	NH	4.05
^a The test set compounds							

4d-qsar Study

The 3D structures of all of the compounds were built by means of Ghemical program (Hassinen and Peräkylä 2001). The structures were optimized with the ffG43a1 force field. Then partial atomic charges of AM1-BCC method were computed with AMBER ff03 atom types using UCSF Chimera (Pettersen et al. 2004). The topology files of compounds were obtained by the topobuild program. In order to obtain conformational ensemble profile (CEP), the molecular dynamics (MD) simulations of all compounds were performed by the GROMACS software (Van Der Spoel et al. 2005). All compounds were put in the dodecahedron box which filled with water molecules. Long-range electrostatics and van der Waals interaction energies were computed by means of Particle Mesh Ewald method with a cut off radius of 10 Å (Darden et al. 1993). System temperature was controlled by Berendsen thermostat coupling, and pressure was kept by Parrinello Rahman coupling (Berendsen et al. 1984). System was optimized by the steepest descent and conjugated gradient method. Using the script of LQTAgrid software (Patil and Sawant 2015), the stepwise heating method was run which included heating the system at 50 K, 100 K, 200 K and 350 K for 20 ps in 1 fs step size. The system was then backed to 300 K for 500 ps. The trajectory file was recorded every 10 ps simulation time.

The compound 126 was chosen as the reference of alignment due to the most active compound among all compounds. All conformations generated in MD simulations at 300 K were superimposed to the reference using the index number of common atoms. The atoms, which were selected for alignment, are shown in Fig. 1. During the alignment, the initial conformer generated at 20 ps was selected, then other trajectories, which were generated up to 100 ps times with 2 ps increment, were subjected to alignment using the least squares method to compute the minimum root mean- square of the distances (RMSD). The aligned CEPs of the most active compound 126 (reference) and alignment with CEPs of least active compound 47 are shown in Fig. 2.

The grid box was defined as $20 \times 19 \times 17$ Å which was large enough to accommodate all conformers. The aligned molecules were submitted to the LQTAgrid program to calculate the energy descriptors of intermolecular interaction every grid point of a 1 Å grid cell lattice. The NH_3^+ probe was selected to simulated N-terminus moiety of protein. The Coulomb interaction descriptors (C descriptors) and Lennard-Jones potential descriptors (LJ descriptors) were calculated. The dimension of the descriptor matrix was $131 \times 15,120$, where each row is a compound and each column is a descriptor.

Descriptor Selection And Model Construction

First, it is necessary to truncate both LJ and C descriptors, in order to avoid large values with high orders of magnitude, and to keep information in the region close to the compounds (Ma et al. 2019). When the distance between the atoms of compound and probe is close to zero, interaction energy generates a large value which do not benefit to the model. Based on equation (1) and (2) (Fig. 3), if the absolute value of interaction energy was more than 30 kcal/mol, the logarithmic value of residual was added to 30 kcal/mol.

Second, if the variance of descriptors is below of 0.01, then the descriptors are excluded. Because the descriptors are far from compounds, and contains very little information.

Third, the pearson correlation coefficients between descriptors and dependent variables were calculated (r vector) using correlation coefficient cut-off according to the equation (3) (Barbosa and Ferreira 2012), where $Z_{0.99}$ is the number of standard deviations equal to 2.33, extending from the mean of normal distribution ($\mu = 0$) required to contain 99% of the area, and σ is the standard deviation of r_{rand} . When the absolute value of $|r|_{\text{cut-off}}$ was lower than 0.3, the descriptors were excluded. This method can eliminate most of noise.

$$|r|_{\text{cut-off}} = Z_{0.99} \sigma \quad (3)$$

Forth, the CDDA was performed to exclude descriptors whose distribution is inconsistent with dependent variables [31]. The descriptors were sorted in descending order according to their absolute value of correlation coefficients. The hyperparameter m (0.05 - 1, step 0.01) was applied to adjust the number of descriptors which were used to build 4D-QSAR model.

Fifth, the OPS method attaches importance to each descriptor based on a vector; then the matrix of descriptors is rearranged according to their relevancy. The most relevant/important descriptors are represented by the first column of the matrix (de Campos and de Melo 2014). Successive partial least squares (PLS) regressions are performed by increasing the descriptors number in order to select the set that build the best latent variables (LVs) for correlation with the endpoints (de Melo et al. 2016). The process is repeated in an iterative manner

Finally, GA in QSRINS package (Gramatica et al. 2013), which is a software for the development and validation of multiple linear regression (MLR) model, was applied to choose the most appropriate descriptors for model. The GA performs its optimization make use of variation and selection via the evaluation of the fitness function. GA is a stochastic technique well suited to the problem of variable selection and optimization, and is proved to be effective as a variable selection method.

Model Evaluation

The internal validation of 4D-QSAR model was performed to establish robustness and internal stability. For internal validation, leave-one-out cross-validation (Q^2_{LOO}) is the most preferred technique in which each compound of the training set was removed once from the dataset, and the biological activity of removed compound was predicted from the model. Leave-many-out cross-validation (Q^2_{LMO}) method was also used which carried out for 30% of data out of training each run. If the average R^2_{LMO} and Q^2_{LMO} is close to R^2 and Q^2 , the model is considered robust (Patil et al. 2018). In order to check the chance correlation, Y-Scrambling testing was performed in which the dependent variable vector, Y-vector, is randomly shuffled many times, then a new QSAR model is built making use of the original independent variable matrix and the R^2_{YS} and Q^2_{YS} values are calculated each time. The averages R^2_{YS} and Q^2_{YS} should be much smaller than the values of the original model. The external validation was used to evaluate the predictive accuracy. The model equation, built using the training set compounds, was applied to test set compounds, and the biological activity of test set compounds was computed. The predictive accuracy is checked in terms of the values of root mean square error external ($RMSE_{EXT}$), r^2_m (Average), Δr^2_m , mean absolute error in fitting (MAE), standard error of estimate (s), concordance correlation coefficient (CCC), predictive residual sum of squares (PRESS) and Golbraikh-Tropsha statistics which calculates the slopes (i.e. k and k') of the regression lines of the external validation (Golbraikh and Tropsha 2002a; Golbraikh and Tropsha 2002b).

Results And Discussion

Analysis of MD trajectories

Following successful simulation, the MD trajectories were investigated for dynamic behavior. The last trajectories, obtained for the 500 ps simulations, were analyzed. For some more active compounds

(compounds 73, 110, 125, 126, 130 and 131) and less active compounds (compounds 12, 36, 47, 48, 61 and 67), the root mean square deviations (RMSD) values, which were calculated from the reference trajectories obtained at the end of simulations of 350 K with the trajectories obtained during simulations of 300 K, stayed within 0.20 nm range (Fig. 4). The RMSD fluctuation of compound 125 is greater than that of other compound, due to compound 125 containing the flexible ester group at position 3. The conformations of target compounds did not drastically change during the MD simulations. This indicates that an equilibrium state of target compounds was reached characterized by the RMSD profile.

4d-qsar Model

The 4D-QSAR model with 18 variables was shown to be the best model by the leave-one-out cross-validation, which resulted in $R^2 = 0.71$ and $Q^2_{\text{LOO}} = 0.60$ and $R^2_{\text{Pred}} = 0.69$. The details of the 4D-QSAR model statistics are shown in Tab. 2. It can be concluded that the model presents relatively high predictive power, as the model satisfied the external and internal validation criteria: $R^2 > 0.6$, Q^2_{LMO} and $Q^2_{\text{LOO}} > 0.5$, $\text{CCC} > 0.85$, Q^2_{F1} , Q^2_{F2} and $Q^2_{\text{F3}} > 0.6$, $\bar{r}^2_{\text{m}} > 0.5$, $\Delta r^2_{\text{m}} < 0.2$ and $0.85 \leq k \leq 1.15$ or $0.85 \leq k' \leq 1.15$ (Balupuri et al. 2020). The difference between R^2 and Q^2_{LOO} is 0.11 units, less than 0.2, indicating the absence of overfitting (Kiralj and Ferreira 2009). In addition, other validation parameters were within acceptable limits.

Table 2
The statistical parameters of 4D-QSAR model

Statistical parameter	Model with 18 variables	Statistical parameter	Model with 18 variables
R^2	0.71	MAE_{cv}	0.84
R^2_{adj}	0.65	$PRESS_{cv}$	13.82
$RMSE_{tr}$	0.87	Q^2_{LMO}	0.59
MAE_{tr}	0.69	R^2_{YS}	0.17
CCC_{tr}	0.89	Q^2_{YS}	-0.25
S	0.96	$RMSE_{ext}$	0.84
F	11.84	MAE_{ext}	0.70
Q^2_{Loo}	0.60	$PRESS_{ext}$	8.47
$RMSE_{cv}$	1.04	R^2_{Pred}	0.69
$\bar{r}^2_m$	0.57	CCC_{ext}	0.86
Q^2_{F1}	0.67	Δr^2_m	0.16
Q^2_{F3}	0.73	Q^2_{F2}	0.67
k'	1.01	k	0.97

The model equation with 18 variables is given in Equation 4. In terms of least squares curve fitting method, the values of the regression coefficient were calculated. The variable with positive values is favorably contributing to the model, whereas the variable with negative coefficients is inversely contributing to the model. The plot for experimental pIC_{50} against predicted pIC_{50} is shown in Fig. 5.

$$pIC_{50} = 0.0167*(C1) - 0.0009*(C2) - 0.0114*(C3) - 0.0076*(C4) - 0.0051*(C5) + 0.0160*(LJ1) + 0.0125*(LJ2) + 0.0024*(LJ3) + 0.2405*(LJ4) + 0.0037*(LJ5) + 0.0147*(LJ6) + 1.4731*(LJ7) - 0.0113*(LJ8) - 0.0239*(LJ9) - 0.0096*(LJ10) - 0.0146*(LJ11) - 0.0241*(LJ12) - 0.0157*(LJ13) + 8.7873 \quad (4)$$

The descriptors are designated as C1: X24_21_15_NH₃⁺_C, C2: X21_27_11_NH₃⁺_C, C3: X25_29_10_NH₃⁺_C, C4: X29_21_17_NH₃⁺_C, C5: X29_24_14_NH₃⁺_C, LJ1: X21_22_16_NH₃⁺_LJ, LJ2: X27_26_17_NH₃⁺_LJ, LJ3: X28_18_15_NH₃⁺_LJ, LJ4: X28_27_6_NH₃⁺_LJ, LJ5: X29_18_13_NH₃⁺_LJ, LJ6:

X32_22_16_NH₃⁺_LJ, LJ7: X34_29_10_NH₃⁺_LJ, LJ8: X25_21_14_NH₃⁺_LJ, LJ9: X28_29_9_NH₃⁺_LJ, LJ10: X29_24_16_NH₃⁺_LJ, LJ11: X29_26_13_NH₃⁺_LJ, LJ12: X31_21_19_NH₃⁺_LJ, LJ13: X32_21_11_NH₃⁺_LJ. Each descriptor denotes the specific interaction energy at the specific grid point. The C1 represents the Coulomb descriptor at the grid point 24, 21 and 15 along the X, Y and Z axes respectively. Similarly, the LJ1 represents the Lennard–Jones descriptor at the grid point 21, 22 and 16 along the X, Y and Z axes respectively. Compare with the R² and Q² of the original model, R²_{YS} and Q²_{YS} was lowest in the Y-scrambling test which implies no chance correlation in the model (Fig. 6).

Model Applicability Domain

The leverage method was used to verify the chemical applicability domain (AD) and the robustness of model. The leverage, h^* , for each molecule was calculated by this method. The warning leverage is generally fixed at $3LV/m$, where LV is the latent variable and m is the number of training set compounds. It can be seen from the AD analysis results presented in Fig. 7 that there are no outliers for all compounds. More importantly, the test compounds which were not applied to build model are predicted with similar accuracy of the training compounds.

Graphical representations of the 4D-QSAR model are shown in Fig. 8. Blue regions are the electrostatic descriptors corresponding to negative regression coefficients and red regions are the electrostatic descriptors related to positive regression coefficients. Likewise, Yellow regions and green regions denote steric descriptors with negative and positive regression coefficients, respectively. The descriptors LJ7 and LJ11 at 3 position (the numbers of compounds as shown in Fig. 1) indicate that bulky group (such as compound 125) is favorable for biological activity. The descriptors LJ4 and LJ9 near the R₃ position indicate that bulky group (such as compound 52) is favorable, a small group unfavorable, for biological activity. The yellow descriptor C3 near 1 position describes electron-withdrawing substituent is preferred, like compound 130, and 131. The descriptors C2 and LJ1 are related to R₁ group, which suggests bulky group at 6 position (such as compound 65 contain butoxyl group) and electron-withdrawing group at 7 position (such as compounds 15 – 19 containing nitro group) is preferred. The R₂ group at phenyl ring describes structural information related to conformational flexibility of substituent group. The descriptors C4 and LJ12 show small group with negative-charge increase the compound's bioactivity. The descriptors LJ5 and LJ3 near R₂ position of compound 111 containing isopropyl group indicate that bulky group is favorable. The descriptors LJ8 and C1 describe that a small group with positive-charge is preferred, such as compound 103. The descriptors LJ10 and C5 prefer small groups with negative-charge like compound 110. The descriptors LJ6 and LJ13 show sterically favorable and unfavorable region respectively, such as compounds 106 and 108, the former contains the 2'-Cl substituent and the latter contains the 2'-SMe substituent. The coefficient of LJ7 is the largest in the model equation, so the green descriptor LJ7 can be mostly connected with the activity.

Conclusion

In this paper, the CEPs were constructed by using GROMACS dynamics simulation and interaction energy descriptors, i.e. Lennard–Jones interaction energy descriptors and Coulomb interaction energy descriptors were calculated using LQTAgrid program. Through the combination of CDDA-OPS-GA method for filtration and selection descriptors, this 4D-QSAR model had achieved satisfactory results.

Declarations

Conflict of interest:

The author(s) confirm that this article content has no conflict of interest.

Acknowledgments

The project was supported by the Undergraduate Innovation and Entrepreneurship Foundation (2021CX242, 2020CX289), the Nanchang University Teaching Reform Foundation (NCUJGLX-19-130, NCUJGLX-19-124) and the Jiangxi Province Science Foundation (20171BAB205104).

References

1. Balupuri A, Balasubramanian PK, Cho SJ (2020) 3D-QSAR, docking, molecular dynamics simulation and free energy calculation studies of some pyrimidine derivatives as novel JAK3 inhibitors. *Arab J Chem* 13:1052-1078.
2. Barbosa EG, Ferreira MM (2012) Digital filters for molecular interaction field descriptors. *Mol Inform* 31:75-84.
3. Berendsen HJ, Postma Jv, van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. *J Chem Phys* 81:3684-3690.
4. Bridges AJ, Zhou HR, Cody DR, Rewcastle GW, McMichael A, Showalter HD, Fry DW, Kraker AJ, Denny WA (1996) Tyrosine kinase inhibitors. 8. an unusually steep structure-activity relationship for analogues of 4-(3-bromoanilino)-6,7-dimethoxyquinazoline (PD 153035), a potent inhibitor of the epidermal growth factor receptor. *J Med Chem* 39:267-276.
5. Burotto M, Chiou VL, Lee JM, Kohn EC (2014) The MAPK pathway across different malignancies: a new perspective. *Cancer* 120:3446-3456.
6. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II, Cronin M, Dearden J, Gramatica P, Martin YC, Todeschini R, Consonni V, Kuz'min VE, Cramer R, Benigni R, Yang C, Rathman J, Terfloth L, Gasteiger J, Richard A, Tropsha A (2014) QSAR modeling: where have you been? where are you going to? *J Med Chem* 57:4977-5010.
7. Cho A, Hur J, Moon YW, Hong SR, Suh YJ, Kim YJ, Im DJ, Hong YJ, Lee HJ, Kim YJ, Shim HS, Lee JS, Kim JH, Choi BW (2016) Correlation between EGFR gene mutation, cytologic tumor markers, 18F-FDG uptake in non-small cell lung cancer. *BMC Cancer* 16:1471-2407

8. Darden T, York D, Pedersen L (1993) Particle mesh ewald: an $N \cdot \log(N)$ method for ewald sums in large systems. *J Chem Phys* 98:10089-10092.
9. de Campos LJ, de Melo EB (2014) Modeling structure–activity relationships of prodiginines with antimalarial activity using GA/MLR and OPS/PLS. *J Mol Graphics Model* 54:19-31.
10. de Melo EB, Martins JPA, Miranda EH, Ferreira MMC (2016) A best comprehension about the toxicity of phenylsulfonyl carboxylates in *vibrio fischeri* using quantitative structure activity/property relationship methods. *J Hazard Mater* 304:233-241.
11. Dearden JC 2017. The history and development of quantitative structure-activity relationships (QSARs). *Oncology: breakthroughs in research and practice*: IGI Global, p. 67-117.
12. Ghasemi J, Salahinejad M, Rofouei M (2011) Review of the quantitative structure–activity relationship modelling methods on estimation of formation constants of macrocyclic compounds with different guest molecules. *Supramol Chem* 23:614-629.
13. Golbraikh A, Tropsha A (2002a) Beware of q^2 ! *J Mol Graph Model* 20:269-276.
14. Golbraikh A, Tropsha A (2002b) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Mol Divers* 5:231-243.
15. Gramatica P, Chirico N, Papa E, Cassani S, Kovarich S (2013) QSARINS: a new software for the development, analysis, and validation of QSAR MLR models. *J Comput Chem* 34:2121-2132.
16. Hassinen T, Peräkylä M (2001) New energy terms for reduced protein models implemented in an off-lattice force field. *J Comput Chem* 22:1229-1242.
17. He Y, Harrington BS, Hooper JD (2016) New crossroads for potential therapeutic intervention in cancer - intersections between CDCP1, EGFR family members and downstream signaling pathways. *Oncoscience* 3:5-8.
18. Hemmateenejad B, Miri R, Akhond M, Shamsipur M (2002) QSAR study of the calcium channel antagonist activity of some recently synthesized dihydropyridine derivatives. an application of genetic algorithm for variable selection in MLR and PLS methods. *Chemometrics Intellig Lab Syst* 64:91-99.
19. Holdman XB, Welte T, Rajapakshe K, Pond A, Coarfa C, Mo Q, Huang S, Hilsenbeck SG, Edwards DP, Zhang X, Rosen JM (2015) Upregulation of EGFR signaling is correlated with tumor stroma remodeling and tumor recurrence in FGFR1-driven breast cancer. *Breast Cancer Res* 17:1465-1481.
20. Hopfinger A, Wang S, Tokarski JS, Jin B, Albuquerque M, Madhav PJ, Duraiswami C (1997) Construction of 3D-QSAR models using the 4D-QSAR analysis formalism. *J Am Chem Soc* 119:10509-10524.
21. Imamura F, Uchida J, Kukita Y, Kumagai T, Nishino K, Inoue T, Kimura M, Oba S, Kato K (2016) Monitoring of treatment responses and clonal evolution of tumor cells by circulating tumor DNA of heterogeneous mutant EGFR genes in lung cancer. *Lung Cancer* 94:68-73.
22. Kiralj R, Ferreira M (2009) Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. *J Braz Chem Soc* 20:770-787.

23. Li S, Guo C, Zhao H, Tang Y, Lan M (2012) Synthesis and biological evaluation of 4-[3-chloro-4-(3-fluorobenzyloxy)anilino]-6-(3-substituted-phenoxy)pyrimidines as dual EGFR/ErbB-2 kinase inhibitors. *Bioorg Med Chem* 20:877-885.
24. Ma W, Wang Y, Chu D, Yan H (2019) 4D-QSAR and MIA-QSAR study on the bruton's tyrosine kinase (Btk) inhibitors. *J Mol Graphics Model* 92:357-362.
25. Martins JP, Barbosa EG, Pasqualoto KF, Ferreira MM (2009) LQTA-QSAR: a new 4D-QSAR methodology. *J Chem Inf Model* 49:1428-1436.
26. Minkovsky N, Berezov A (2008) BIBW-2992, a dual receptor tyrosine kinase inhibitor for the treatment of solid tumors. *Curr Opin Investig Drugs* 9:1336-1346.
27. Oda K, Matsuoka Y, Funahashi A, Kitano H (2005) A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol Syst Biol* 1:1-17.
28. Patil R, Sawant S (2015) Molecular dynamics guided receptor independent 4D QSAR studies of substituted coumarins as anticancer agents. *Curr Comput Aided Drug Des* 11:39-50.
29. Patil RB, Barbosa EG, Sangshetti JN, Sawant SD, Zambre VP (2018) LQTA-R: a new 3D-QSAR methodology applied to a set of DGAT1 inhibitors. *Comput Biol Chem* 74:123-131.
30. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera: a visualization system for exploratory research and analysis. *J Comput Chem* 25:1605-1612.
31. Shim J, Mackerell Jr AD (2011) Computational ligand-based rational design: role of conformational sampling and force fields in model development. *Medchemcomm* 2:356-370.
32. Suzuki N, Shiota T, Watanabe F, Haga N, Murashi T, Ohara T, Matsuo K, Oomori N, Yari H, Dohi K, Inoue M, Iguchi M, Sentou J, Wada T (2011) Synthesis and evaluation of novel pyrimidine-based dual EGFR/Her-2 inhibitors. *Bioorg Med Chem Lett* 21:1601-1606.
33. Teofilo RF, Martins JPA, Ferreira MM (2009) Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression. *J Chemometrics* 23:32-48.
34. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ (2005) GROMACS: fast, flexible, and free. *J Comput Chem* 26:1701-1718.
35. Wang K, Li D, Sun L (2016a) High levels of EGFR expression in tumor stroma are associated with aggressive clinical features in epithelial ovarian cancer. *Onco Targets Ther* 9:377-386.
36. Wang R, Wang X, Wu JQ, Ni B, Wen LB, Huang L, Liao Y, Tong GZ, Ding C, Mao X (2016b) Efficient porcine reproductive and respiratory syndrome virus entry in MARC-145 cells requires EGFR-PI3K-AKT-LIMK1-COFILIN signaling pathway. *Virus Res* 225:23-32.
37. Waterson AG, Petrov KG, Hornberger KR, Hubbard RD, Sammond DM, Smith SC, Dickson HD, Caferro TR, Hinkle KW, Stevens KL, Dickerson SH, Rusnak DW, Spehar GM, Wood ER, Griffin RJ, Uehling DE (2009) Synthesis and evaluation of aniline headgroups for alkynyl thienopyrimidine dual EGFR/ErbB-2 kinase inhibitors. *Bioorg Med Chem Lett* 19:1332-1336.

38. Yewale C, Baradia D, Vhora I, Patil S, Misra A (2013) Epidermal growth factor receptor targeting in cancer: a review of trends and strategies. *Biomaterials* 34:8690-8707.

Figures

Figure 1

The structure of data set and red atoms used for alignment

Figure 2

The aligned CEPs generated in MD simulations. A): aligned CEPs of compound 126 (represented by licorice), B): alignment of reference with the least active compound 47 (represented by line)

$$LJ'_{x,y,z} = \begin{cases} LJ_{x,y,z} & LJ_{x,y,z} \leq 30 \\ 30 + \log_{10}(LJ_{x,y,z} - 29) & LJ_{x,y,z} > 30 \end{cases} \quad (1)$$

$$C'_{x,y,z} = \begin{cases} C_{x,y,z} & |C_{x,y,z}| \leq 30 \\ -30 + \log_{10}(-C_{x,y,z} - 29) & C_{x,y,z} < -30 \\ 30 + \log_{10}(C_{x,y,z} - 29) & C_{x,y,z} > 30 \end{cases} \quad (2)$$

Figure 3

The equation for truncating both LJ and C descriptors

Figure 4

Graphs showing RMSD value. A): RMSD value of most active compounds. B): RMSD value of less active compounds.

Figure 5

Predicted pIC₅₀ versus experimental pIC₅₀ for training set and test set compounds

Figure 6

Plot of Y-scrambled models compared with the original model.

Figure 7

Plot of leverages against standardized residuals. Dashed lines and dotted lines represent ± 3.0 standardized residual and warning leverage ($h^* = 0.543$).

Figure 8

Different perspectives of the steric and electrostatic contour maps of 4D-QSAR model