Medicinal Chemistry Research (2019) 28:2079–2087
https://doi.org/10.1007/s00044-019-02437-y

MEDICINAL CHEMISTRY RESEARCH

ORIGINAL RESEARCH

QSAR study of human epidermal growth factor receptor (EGFR) inhibitors: conformation-independent models

Silvina E. Fioressi[1] · Daniel E. Bacelo[1] · Pablo R. Duchowicz[2]

Received: 9 July 2019 / Accepted: 31 August 2019 / Published online: 13 September 2019

Abstract

Many compounds have been proposed and tested as human epidermal growth factor receptor (EGFR) inhibitors for cancer treatment. Recently, new survival mechanisms of cancer cells have been discovered with the consequent resistance to therapy, which makes it necessary to search for new anticancer drugs. Here we perform a quantitative structure-activity relationship (QSAR) study on 290 compounds reported in the literature as EGFR inhibitors to analyze the molecular properties that may influence their activity. A large number of nonconformational descriptors (17,974) were explored including molecular descriptors, flexible molecular descriptors, and combination of both. To avoid ambiguities derived from the existence of several conformational states, only constitutional and topological molecular descriptors have been considered. The models were validated through Y-randomization, cross-validation, and mean absolute error criteria. A simple model involving flexible descriptors shows the best predictive performance and suggests that the presence of multiple aromatic rings and amino groups in a compound structure may increase its EGFR inhibitory activity.

Keywords Cancer · EGFR · QSAR · Tyrosine kinase protein · HER1 · Drug design

Abbreviations

EGFR	Epidermal growth factor receptor
QSAR	Quantitative structure-activity relationship
U.S. FDA	United State Food and Drug Administration
IC_{50}	The inhibitory activity was expressed as the concentration of the test compound that inhibited the activity of EGFR by 50%
pIC_{50}	The logarithmic molar IC_{50} values
SMILES	Molecular-input line-entry system
SR	Structural representation
HSG	Hydrogen-suppressed graph
HFG	Hydrogen-filled graph
GAO	Graph of atomic orbitals
SA	Structural attributes
DCW	Defined flexible descriptor
CW	Correlation weights
MC	Monte Carlo simulation
T	Threshold value
BSM	Balanced subsets method (BSM)
k-MCA	k-means cluster analysis
RM	Replacement method
MLR	Multivariable linear regression
Loo	Leave-one-out cross-validation
R^2_{Loo}	Loo variance
MAE	Mean absolute error
AD	Applicability domain
h_i	Calculated leverage value
h^*	Warning leverage value
S_{Val}	Standard deviation in the validation set
F	Fisher parameter
$o(2.5S)$	Number of outlier compounds in the training set

Supplementary information The online version of this article (https://doi.org/10.1007/s00044-019-02437-y) contains supplementary material, which is available to authorized users.

✉ Silvina E. Fioressi
sfioressi@yahoo.com

✉ Pablo R. Duchowicz
pabloducho@gmail.com

[1] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Facultad de Ciencias Exactas y Naturales, Universidad de Belgrano, Villanueva 1324 CP 1426, Buenos Aires, Argentina

[2] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA), CONICET, UNLP, Diag. 113y 64, Sucursal 4, C.C. 16, 1900 La Plata, Argentina

Introduction

The epidermal growth factor receptor (EGFR) tyrosine kinase protein, which includes the EGFR/HER1/Erb1, EGFR/ErbB2, HER3/ErbB3, and ErbB4 family of receptors, is among the most investigated cell signaling families in cancer research (Schlessinger 2000; Arteaga and Engelman 2014). The overexpression, deregulation, and mutations of EGFR appear to be crucial in tumor growth and progression in many malignancies, principally in the different kinds of epithelial cancers (Sigismund et al. 2018). Therefore, the use of tyrosine kinase inhibitors is one of the most clinically advanced approaches for the treatment of these cancers. The overexpression of EGFR is found in many malignancies particularly in carcinoma of lungs, glioblastoma, epithelial head and neck tumors, colon and breast cancer (Barber et al. 2004; Kalyankrishna and Grandis 2006; Mok et al. 2009; Walker et al. 2009; Ueno and Zhang 2011). Several drugs have already been launched as FDA approved kinase inhibitors including: gefitinib, erlotinib, afatinib, brigatinib, neratinib, vandetanib, lapatinib, osimertinib, icotinib among others (Gazit et al. 1993; Levitzki and Gazit 1995; Wu et al. 2016). For breast cancer, for example, trastuzumab (HerceptinH) (Orman and Perry 2007; Liang et al. 2010) has proven to be efficient in the inhibition of EGFR but induces EGFR overexpression and resistance after long term treatment (Pohlmann et al. 2009). For patients who are refractory to trastuzumab and chemotherapy, lapatinib (TykerbH) may be more effective because inhibits both, HER2 and EGFR (Curran 2010; Kroep et al. 2010). Recent findings of EGFR mutations and the dose-related toxicity of gefitinib, erlotinib, and afatinib, impose an urgent need for the design and development of new and more effective drugs. Most FDA approved inhibitors act towards the best known functions of the EGFR, which are in the context of ligand- and kinase-dependent activation (Lemmon and Schlessinger 2010). However, novel functions have been recently discovered both kinase dependent and independent, revealing unexpected roles of the EGFR in the regulation of autophagy and metabolism (Tan et al. 2016). Moreover, it was found that these new functions are stimulated by cellular stress, for example the stress caused by EGFR inhibitors. This may result in additional mechanisms of cancer cells survival with the consequential therapy resistance (Jutten et al. 2013). It is clear then, than the search for new anticancer agents is mandatory. At present, different families of structures have been suggested as EGFR inhibitors and their efficiency has been published (Gazit et al. 1991; Levitzki and Gazit 1995; Fink et al. 2005; Mastalerz et al. 2007; Mastalerz et al. 2007; Mastalerz et al. 2007; Xu et al. 2008; Xu et al. 2008; Cai et al. 2010; Li et al. 2010; Lv et al. 2010; Fink et al. 2011). Finding new molecular features that improve drug efficacy is an expensive and time-consuming task, although essential for the design of improved drugs.

The application of QSAR techniques and computer-aided modeling are valuable tools to initiate the search for new drugs and are being intensively used for this purpose. Trustworthy models can provide insight into the molecular characteristics that may influence the drug inhibitory activity, drastically improving the success and the pace of the development of more effective drugs with weaker secondary effects. The identification of new EGFR inhibitor molecules has been intensely studied using different approaches including QSAR (Noolvi and Patel 2010; Marzaro et al. 2011; Chauhan et al. 2014; Sun et al. 2014; Singh et al. 2015; Faghih-Mirzaei et al. 2019), molecular docking (Nandi and Bagchi 2010; Bathini et al. 2016; Shinde et al. 2017; Gaber et al. 2018; Faghih-Mirzaei et al. 2019; Ruslin et al. 2019), and pharmacophore modeling (Gupta et al. 2011) studies.

The purpose of this study is to model the inhibitory activity of a large number of compounds towards EGFR using QSAR, considering only flexible molecular descriptors of 0, 1, and 2 dimensions. As a result, simple models were developed based solely on constitutional and topological molecular properties, thus avoiding the ambiguities that can arise if different conformational states are considered (Duchowicz et al. 2012; Talevi et al. 2012). Three different approaches were explored using different types of descriptors: (i) 0D, 1D, and 2D descriptors and fingerprints generated by popular and freely available programs: PaDEL-descriptor (version 2.20) (Yap 2011), EPI Suite (U.S. Environmental Protection Agency 2016), and Mold2 (Hong et al. 2008); (ii) flexible descriptors obtained through the CORALSEA program (Toropova et al. 2012) and (iii) both sets of descriptors, combined. Based on the analysis of the statistical parameters obtained using this methodology and always looking for the simplest models, linear relationships based on 1–8 descriptors have been selected as the best predictive combinations of independently selected variables. This strategy was successfully applied recently for the modeling of the inhibitory activity of HER2, obtaining a hybrid relationship that combines a flexible descriptor, three PaDel descriptors, and one fingerprint (Duchowicz et al. 2017).

Materials and methods

QSAR analysis was performed on 290 EGFR inhibitors (Table 1S). Their structures and in vitro activities, measured by different enzymological methods, were collected from recently published literature (Gazit et al. 1991; Gazit et al. 1993; Levitzki and Gazit 1995; Fink et al. 2005; Mastalerz et al. 2007; Mastalerz et al. 2007; Mastalerz et al. 2007; Xu

et al. 2008; Xu et al. 2008; Cai et al. 2010; Li et al. 2010; Lv et al. 2010; Fink et al. 2011). The inhibitory activity was expressed as the concentration of the test compound that inhibited the activity of EGFR by 50% (IC_{50}). The molar IC_{50} values were converted into the logarithmic (pIC_{50}, M) values to be used in the QSAR analysis.

Structural representation and molecular descriptors calculation

Compound structures were generated in both SMILES notation and bidimensional structures drawn with the free Discovery Studio software (Version 3.5) (Dassault Systèmes Biovia 2017) and saved in MDL-MOL format without performing any geometrical optimizations. Two different methods were used to calculate the descriptors: (a) Theoretical conformation-independent molecular descriptors and fingerprints were calculated using the freely available software PaDEL-descriptor (version 2.20) (Yap 2011), EPI Suite (U.S. Environmental Protection Agency 2016), and Mold2 (Hong et al. 2008). Overall, 1444 1D and 2D descriptors and 12 types of fingerprints (16,092) were obtained from PaDEL-descriptor, 184 descriptors from EPI Suite, and 254 descriptors from Mold2. A total of 17,974 descriptors were used to exhaustively explore the structural characteristics that may influence EGFR inhibitory activity. To minimize redundant information, descriptors found to be linearly dependent and constant values were excluded from the matrix of variables. (b) The CORAL freeware (Toropova et al. 2012) was used to calculate flexible molecular descriptors. The SMILES notations of the compounds were provided as input to the CORAL program along with the experimental pIC_{50} values. Three different structural representation (SR) approaches are available in the CORAL program: (i) a chemical graph, such as hydrogen-suppressed graph, hydrogen-filled graph (HFG), or graph of atomic orbitals; (ii) SMILES; and (iii) a hybrid of chemical graph and SMILES. Given that the SR selected defines the number and types of local descriptors to be included in the QSAR analysis, the most appropriate combination of structural attributes (local descriptors, SA) for modeling should be chosen. The CORAL framework searches for a QSAR model that correlates the experimental pIC_{50} and a properly defined flexible descriptor (DCW) through a one-variable linear relationship. The DCW descriptor is a linear combination of special coefficients called correlation weights (CW) whose value is calculated for each SA type in the training set via a Monte Carlo simulation (Table 2S). The DCW depends on the threshold value (T) and the number of epochs or iterations (N_{epochs}) used (Toropova et al. 2012). T defines rare SMILES attributes that do not contribute to the predicted activity. All SMILES attributes that take place in less than T SMILES notations of the training set are classified as rare instead of active. In this study, T ranges from 0 to 5 and the maximum number of iterations used is 50.

Model validation

To verify the predictive capability of the proposed models the dataset was split into a training set (99 compounds) for model development, a validation set (99 compounds) for model validation, and a test set (92 compounds) for external validation. The split of the dataset was carried out by the balanced subsets method (Rojas et al. 2016), a technique based on k-Means Cluster Analysis (k-MCA), to ensure that the training set is representative of the validation and test sets. The replacement method (Duchowicz et al. 2006), was applied to generate multivariable linear regression (MLR) models on the training set. The algorithms used in our calculations were programmed in MATLAB software (The MathWorks 2018). The MLR models were validated theoretically through the leave-one-out cross-validation (Loo) method to measure the stability of the QSAR model upon inclusion/exclusion of molecules. A general validation criterion is to accept the model if the Loo variance (R^2_{Loo}) is greater than 0.5. This is a necessary but not sufficient condition for predictive power (Golbraikh and Tropsha 2002). A more thorough validation was sought using the external test set of 92 compounds. To check that the model does not result from happenstance, the experimental values were scrambled through the Y-randomization method (Wold et al. 1995) in such a way as to not correspond to their respective compounds.

Mean absolute error (MAE) was the criterion applied to evaluate the predictive error values relative to the training set response range (Roy et al. 2016). A MAE of 10% of the training set range indicates that the model is adequately predictive. The criterion based on the value of the MAE + 3σ to be less than 25% of the training set range was also assessed (σ is the standard deviation of absolute error values for the test set).

Applicability domain

The applicability domain (AD) is a theoretically defined area such that only the molecules that are within this AD are not considered model extrapolations (Gramatica 2007; Gadaleta et al. 2016). The ADs for the proposed models were determined though two methodologies: the leverage approach (Eriksson et al. 2003) and a simple standardization method (Roy et al. 2015). In the leverage approach each compound i has a calculated leverage value h_i and a warning leverage value h^* (Table 2S), so that if h_i is greater than h^*, the prediction is considered a model extrapolation and is therefore, not reliable. In addition, the recently

proposed method for identifying compounds outside the AD, developed by Roy et al., was applied (Roy et al. 2015).

Results and discussion

We performed a QSAR analysis on 290 different compounds with proven inhibitory activity towards EGFR using experimental data reported in the literature. The three different approaches to QSAR described in the previous section were explored following a general methodology. First, the training set is used both for the selection of the best molecular descriptors and for the adjustment of the MLR parameters, while the validation set is used to determine the predictive capacity of the model. Then, the models are tested externally for the experimental pIC_{50} data in the test set. This allows to take maximum advantage of the available structural and response information and to enlarge the AD. The statistical parameters for each model are provided as supplementary information. The detailed results obtained for each approach are presented and thoroughly analyzed in the following sections.

Molecular descriptors models

Table 1 shows the results obtained for the best models found with the first approach using molecular descriptors and fingerprints. Models involving from one to eight descriptors were explored and the best predictive performance is observed for the five and six descriptors models. Both present similar values of S_{Val}. The five-descriptors model was selected as the best because the S_{Val} and S_{Train} values are closer to each other and it is a simpler model than the one with six descriptors. Figure 1 shows the calculated pIC_{50} versus the experimental values for the five-descriptors model represented by the equation:

$$\begin{aligned} pIC50 = {} & 1.695 - 1.529 GATS5c + 1.440 \\ & SpMax3\_Bhm - 0.883 PubchemFP260 \\ & +1.913\, PubchemFP577 + 1.085\, PubchemFP703 \end{aligned} \quad (1)$$

$$N_{Train} = 99, R^2_{Train} = 0.76, S_{Train} = 0.62, N_{Val} = 99, R^2_{Val} = 0.73, S_{Val} = 0.67, F = 59$$

$$N_{Test} = 92, R^2_{Test} = 0.73, S_{Test} = 0.65, o(2.5S) = 1, R^2_{Loo} = 0.68 \; S_{Loo} = 0.65, S_{Rand} = 1.18, h^* = 0.0909, \text{MAE} = 0.46,$$
Train range = 4.92.

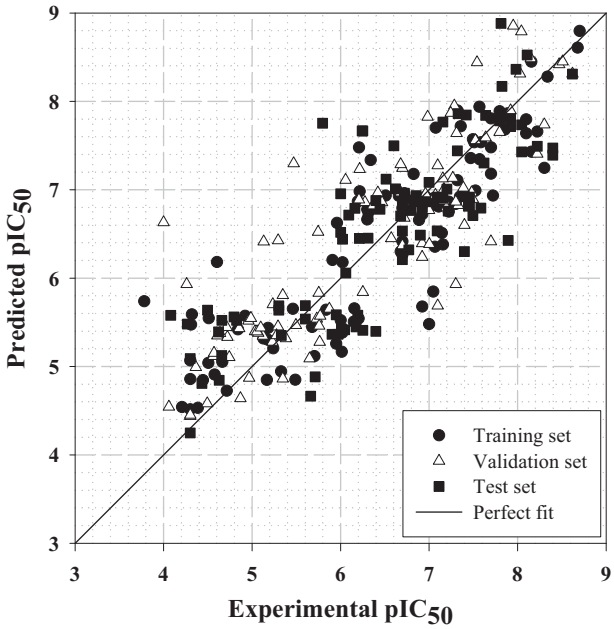


Fig. 1 Experimental and predicted values for the training, validation, and test sets for the five-descriptor model (Eq. 1) for 290 EGFR inhibitors

Table 1 Descriptors identified for modeling inhibitory EGFR activity together with the squared correlation coefficient and the standard deviation for the training, validation, and test sets

No. of Des.	Descriptors	R^2_{Train}	S_{Train}	R^2_{Val}	S_{Val}	R^2_{Test}	S_{Test}
1	*PubchemFP379*	0.55	0.83	0.62	0.76	0.59	0.71
2	*MACCSFP38, APC2D8_C_X*	0.61	0.78	0.69	0.71	0.59	0.71
3	*GATS5e, maxHBd, PubchemFP577*	0.68	0.70	0.73	0.66	0.63	0.68
4	*GATS5c, PubchemFP259, PubchemFP260, PubchemFP577*	0.73	0.65	0.73	0.67	0.65	0.67
5	***GATS5c, SpMax3_Bhm, PubchemFP260, PubchemFP577, PubchemFP703***	**0.76**	**0.62**	**0.73**	**0.67**	**0.65**	**0.68**
6	*SpMax3_Bhm, SpMin5_Bhm, nHBDon_Lipinski, EStateFP32, PubchemFP577, PubchemFP703*	0.80	0.57	0.74	0.66	0.63	0.71
7	*SaaaC, MIC1, EStateFP18, Pubchem FP192, PubchemFP577, KRFP480, Estimated BCF*	0.83	0.52	0.72	0.69	0.58	0.63
8	*MIC2, MACCSFP38, Pubchem FP192, PubchemFP260, PubchemFP577, KRFP480, KRFP4254, Estimated BCF*	0.86	0.48	0.71	0.69	0.62	0.74

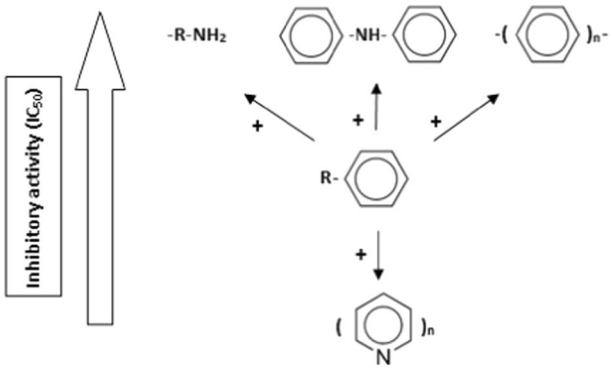
The best model is in bold text

The Fisher parameter (F) accounts for statistical significance and $o(2.5S)$ (Verma and Hansch 2005) indicates the number of outlier compounds in the training set. A compound having an absolute residual value (difference between experimental and calculated pIC_{50}) greater than 2.5 times S_{Train} is considered an outlier. All compounds in this model are within the AD and only one outlier (compound 248) was found. The MAE result including all test compounds is under 10% of the training compounds' value range (Roy et al. 2015). Equation 1 satisfies the external validation conditions (Roy 2007).

$$-1 - R_0^2/R_{Test}^2 < 0.1(0.0028) \text{ or } 1 - R_0'^2/R_{Test}^2 < 0.1(0.12)$$
$$\text{and, } -0.85 \le k \le 1.15(0.98) \text{ and } 0.85 \le k' \le 1.15(1.01)$$
$$\text{and } - R_M^2 > 0.5\,(0.67).$$

Two descriptors in this model present a negative effect on the inhibitory activity: *GATS5c* and *PubchemFP260*. The *GATS5c* is a 2D-autocorrelation descriptor originating in autocorrelation of the topological structure of Geary (GATS). This descriptor encodes both the molecular structure and a physicochemical property as a vector, associating the topology of a structure with a selected physicochemical attribute. The two indices following the descriptor symbol represent the topological distance between pairs of atoms, lag 5 in this case, and the physicochemical property considered in the weighting component for its computation. For the *GATS5c* descriptor the letter *c* means weighted by charge. The *PubchemFP260* is a fingerprint that indicates the presence of three hetero-aromatic rings in the structure.

On the other hand, the *SpMax3_Bhm* descriptor and the *PubchemFP577* and *PubchemFP703* fingerprints present a positive correlation with the pIC_{50}. These three descriptors also appear in the six descriptors model with positive coefficients supporting the conclusion that higher values for these descriptors should increase the EGFR inhibitory activity. Fingerprints *PubchemFP577* and *PubchemFP703* both test for the presence of an amino group. The *FP577* indicates the existence of a secondary amino group bridging two aromatic carbons and the *FP703* the presence of a primary amine bonded to an aliphatic carbon. Secondary and tertiary amino groups have been recently reported as the most frequent moieties found in anticancer drugs tested against NCI-60 cell lines in a QSAR study using a dataset of 8565 molecules (Singh et al. 2016). The *SpMax3_Bhm* descriptor represents the largest absolute eigenvalue of the Burden modified matrix – n 3/weighted by relative mass. The Burden matrix is an adjacency matrix and its eigenvalues reflect relevant aspects of molecular structure useful for similarity searching. In the case of the *SpMax3_Bhm* descriptor, the diagonal elements of the matrix are given by the carbon normalized atomic mass. The molecular



Scheme 1 Influence of the presence of functional groups on the EGFR inhibitory activity

descriptors and fingerprints appearing in Eq. 1 suggest that the inhibitory activity of EGFR is decreased by the presence of multiple hetero-aromatic rings and increased by the occurrence of atoms heavier than carbon and also by the presence of primary or secondary amines in the compounds structure (Scheme 1).

Flexible molecular descriptors model

The best linear regression models were selected by performing a QSAR analysis on the training set of 99 compounds. The flexible-descriptor design requires finding the most efficient structural attributes for each SR. To achieve this, the DCW flexible descriptor is optimized by increasing R_{Train}^2, until the predictive capability of the model in the validation set begins to decrease. This is the same procedure followed to select the most predictive MLR model searching for the best combination among thousands of descriptors. In all cases, the test set is not involved in the development of the model. The main statistical parameters for the QSAR models with better predictive performance are presented in Table 2. It can be seen from these results that the best choice is a CORAL combination that contains HFG representations. Three variable types based on 168 active attributes compose the optimal descriptor (shown in Table 10S). Figure 2 shows that the predicted and experimental values for the training, validation, and test sets follow a straight line. The resulting equation for this model with one DCW is:

$$pIC_{50} = 3.699 + 0.042 * DCW \quad (2)$$

$$N_{Train} = 99, R_{Train}^2 = 0.73, S_{Train} = 0.64, N_{Val} = 99,$$
$$R_{Val}^2 = 0.74, S_{Val} = 0.64, F = 260$$

$$N_{Test} = 92, R_{Test}^2 = 0.67, S_{Test} = 0.64, o(2.5S) = 3,$$
$$R_{Loo}^2 = 0.79 S_{Loo} = 0.56, S_{Rand} = 1.12,$$

$$h^* = 0.06, \text{MAE} = 0.46, \text{ Train range} = 4.92.$$

Table 2 The search for the best QSAR model using flexible molecular descriptors

Structural attributes	R^2_{Train}	S_{Train}	R^2_{Val}	S_{Val}	R^2_{Test}	S_{Test}
^1EC_j	0.64	0.74	0.72	0.66	0.65	0.65
^2EC_j	0.80	0.55	0.74	0.64	0.58	0.74
^1EC_j, ^2EC_j	0.75	0.61	0.72	0.66	0.70	0.62
^2EC_j, $Pt2_k$	0.76	0.60	0.71	0.67	0.68	0.64
^1EC_j, ^2EC_j, $Pt2_k$,	**0.73**	**0.64**	**0.74**	**0.64**	**0.67**	**0.64**
^0EC_j, ^1EC_j, ^2EC_j	0.76	0.61	0.72	0.66	0.66	0.66
^1EC_j, ^2EC_j, $Pt2_k$, NNC	0.74	0.63	0.73	0.65	0.68	0.62
^0EC_j, ^1EC_j, ^2EC_j, ^3S_k	0.81	0.53	0.73	0.66	0.69	0.64
^1EC_j, ^2EC, $Pt2_k$, NNC_j, ^3S_k	0.81	0.53	0.74	0.65	0.67	0.67
^0EC_j, ^1EC_j, ^2EC_j, $Pt2_k$, ^3S_k	0.79	0.56	0.73	0.65	0.68	0.64

The best model is in bold text

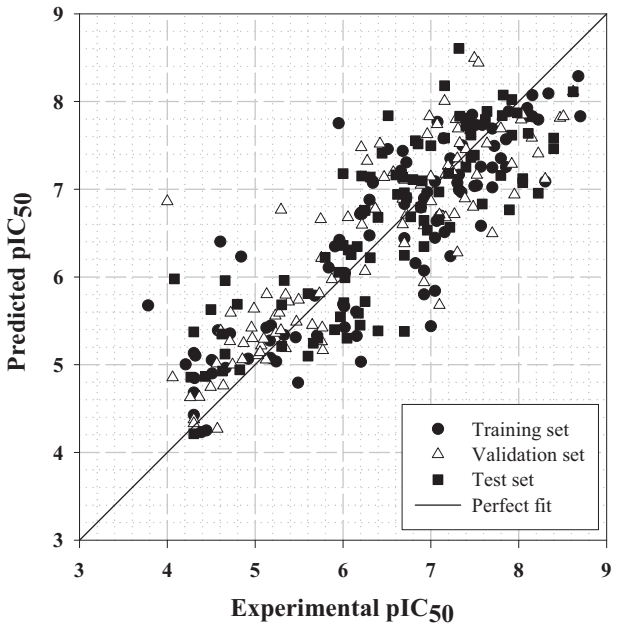


Fig. 2 Experimental and predicted values for the training, validation, and test sets for the one-flexible-descriptor model (Eq. 2) for 290 EGFR inhibitors

All compounds are within the AD according to both methods used, and systematic error is absent. The MAE calculated including all test compounds is lower than 10 % of the training compounds' value range. Three compounds in the training set (67, 213, and 248) show absolute residuals greater than 2.5 times S_{Train} and are considered outliers, however the three present absolute residuals lower than 3 times S_{Train}. We applied both Y-randomization to demonstrate that $S_{Train} < S_{Rand}$ and also the external validation criterion (Roy 2007) to ensure that a valid structure-activity relationship is achieved:

$$-1 - R^2_0/R^2_{Test} < 0.1(0.0001) \text{ or } 1 - R'^2_0/R^2_{Test} < 0.1(0.12)$$
$$\text{and, } -0.85 \le k \le 1.15(0.98) \text{ and } 0.85 \le k' \le 1.15(1.01)$$
$$\text{and } - R^2_M > 0.5(0.73).$$

Table 11S includes an example of a DCW calculation for compound **2**. The local descriptors that contribute to the DCW calculation are listed in Table 10S and are all structural attributes. Higher positive CW values tend to predict higher activity values.

Hybrid descriptors model

The last approach explores models combining PaDEL, EPI Suite, Mold2, and the flexible CORAL descriptors and fingerprints. The combination of various flexible descriptors or flexible descriptors with molecular descriptors results in a significant increase in the model's complexity and does not improve their predictive quality. The best hybrid descriptors model involves five descriptors (see Table 3). It is worth mentioning that, in this model, the descriptor called *CoralA* is the best descriptor found in the previous model (flexible molecular descriptors model, Eq. 2) whereas *CoralB* refers to the descriptors obtained using Coral in HFG representations for the attributes ^2EC_j, NNC_j, ^3S_k. Figure 3 shows the predicted and experimental values for the training, validation, and test sets using the best hybrid model represented by the equation:

$$pIC_{50} = 3.417 - 0.343 PubchemFP199 + 0.496 PubchemFP577 + 1.086 KRFP480 - 0.0001 Estimated\ BCF + 0.0366\ CoralB \quad (3)$$

$$N_{Train} = 99, R^2_{Train} = 0.89, S_{Train} = 0.41, N_{Val} = 99, R^2_{Val} = 0.76, S_{Val} = 0.64, F = 155$$

$$N_{Test} = 92, R^2_{Test} = 0.68, S_{Test} = 0.69, o(2.5S) = 0, R^2_{Loo} = 0.87$$
$$S_{Loo} = 0.46, S_{Rand} = 1.19, h^* = 0.18, \text{MAE} = 0.50,$$
$$\text{Train Trainrange} = 4.92.$$

According to the two methods used, four of the compounds in the test set are outside the AD and no systematic error was observed. The MAE result, including all test compounds, is slightly more than 10% of the training range indicating moderate predictive performance. The *CoralB* descriptor, the *KRFP490* fingerprint (indicates presence of the fragment $[CH_2][NH]C(=S)[NH]$) and the *PubchemFP577*, which denotes the presence of a secondary amino group, present a positive correlation with the

Table 3 Descriptors identified for modeling inhibitory EGFR activity together with the squared correlation coefficient and the standard deviation for training, validation, and test sets

No. of Des.	Descriptors	R^2_{Train}	S_{Train}	R^2_{Val}	S_{Val}	R^2_{Test}	S_{Test}
1	*CoralA*[a]	0.73	0.64	0.74	0.64	0.67	0.65
2	*CoralB*[b], *PubchemFP199*	0.84	0.49	0.74	0.65	0.67	0.67
3	*CoralB, PubchemFP199, KRFP480*	0.86	0.47	0.74	0.65	0.64	0.72
4	*CoralB, PubchemFP199, KRFP480, Estimated BCF*	0.88	0.43	0.73	0.67	0.66	0.71
5	***CoralB, PubchemFP199, PubchemFP577, KRFP480, Estimated BCF***	**0.89**	**0.41**	**0.76**	**0.64**	**0.68**	**0.69**
6	*CoralB, ATSC6c, PubchemFP199, PubchemFP577, KRFP480, Estimated BCF*	0.91	0.38	0.75	0.65	0.68	0.72
7	*CoralB, ATSC6c, PubchemFP199, PubchemFP577, KRFP480, KRFP4028, Estimated BCF*	0.92	0.36	0.75	0.66	0.67	0.74
8	*CoralB, ATSC6c, PubchemFP199, PubchemFP577, KRFP480, KRFP4028 APC2D6_C_Br, Estimated BCF*	0.92	0.35	0.76	0.64	0.58	0.85

[a]CoralA refers to the descriptors obtained using Coral in HFG representations for the attributes ^1EC_j, ^2EC_j, $Pt2_k$

[b]CoralB refers to the descriptors obtained using Coral in HFG representations for the attributes ^2EC_j, NNC_j, ^3S_k

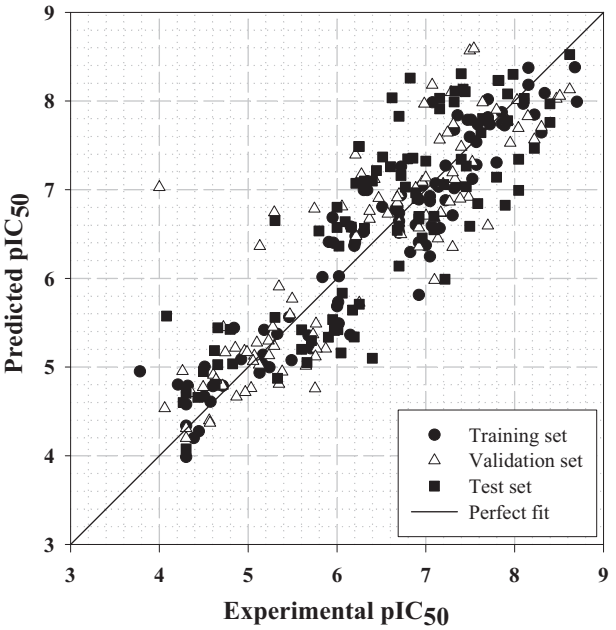


Fig. 3 Experimental and predicted values for the training, validation, and test set for the five-descriptors hybrid model (Eq. 3) for 290 EGFR inhibitors

predicted activity. This last fingerprint was also found in the best model of the molecular descriptors approach (Eq. 1). The *PubchemFP19* (presence of four aromatic rings) and the EPI suite descriptor *Estimated BCF* have negative coefficients in this model. Equation 3 also satisfies the external validation conditions (Roy 2007).

$$-1 - R_0^2/R_{\text{Test}}^2 < 0.1(0.023) \text{ or } 1 - R_0'^2/R_{\text{Test}}^2 < 0.1(0.036)$$
$$\text{and, } -0.85 \le k \le 1.15(0.98) \text{ and } 0.85 \le k' \le 1.15(1.01)$$
$$\text{and } -R_M^2 > 0.5(0.63)$$

Conclusion

A simple structure-inhibitory activity relationship for EGFR inhibitors was developed through a strategy that does not require knowledge of the molecular conformation as part of the SR. A model that uses flexible descriptors calculated with the Coral software shows the best predictive performance for the EGFR inhibitory activity. This model was validated through Y-randomization, cross-validation and MAE criteria, and satisfies AD analysis. Calculations involving molecular descriptors and fingerprints also result in an acceptable model involving five descriptors. However, combinations of Coral flexible descriptors with fingerprints and molecular descriptors did not lead to better models and substantially increases their complexity. The descriptors involved in the models here proposed suggest that the presence of aromatic rings and amino groups in the inhibitors molecular structure have a positive effect on the EGFR inhibitory activity. The results obtained here could be useful in the design of new anticancer drugs that effectively inhibit the expression and function of EGFR and lead to more efficient treatments.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgements We are grateful for financial support provided by the National Research Council of Argentina (CONICET) project PIP11220130100311 and to the Ministerio de Ciencia, Tecnología e Innovación Productiva for access to electronic library facilities. SEF, DEB, and PRD are members of the scientific researcher career of CONICET.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

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

References

Arteaga CL, Engelman JA (2014) ERBB receptors: from oncogene discovery to basic science to mechanism-based cancer therapeutics. Cancer Cell 25:282–303

Barber TD, Vogelstein B, Kinzler KW, Velculescu VE (2004) Somatic mutations of EGFR in colorectal cancers and glioblastomas. N Engl J Med 351:2883

Bathini R, Sivan SK, Fatima S, Manga V (2016) Molecular docking, MM/GBSA and 3D-QSAR studies on EGFR inhibitors. J Chem Sci 128:1163–1173

Cai X, Zhai H-X, Wang J, Forrester J, Qu H, Yin L, Lai C-J, Bao R, Qian C (2010) Discovery of 7-(4-(3-ethynylphenylamino)-7-methoxyquinazolin-6-yloxy)-N-hydroxyheptanamide (CUDC-101) as a potent multi-acting HDAC, EGFR, and HER2 inhibitor for the treatment of cancer. J Med Chem 53:2000–2009

Chauhan J, Dhanda S, Singla D (2014) The open source drug discovery; Agarwal, SM; Raghava, GPS QSAR-based models for designing quinazoline/imidazothiazoles/pyrazolopyrimidines based inhibitors against wild and mutant EGFR. PLoS ONE 9:e101079

Curran MP (2010) Lapatinib: in postmenopausal women with hormone receptor-positive, HER2-positive metastatic breast cancer. Drugs 70:1411–1422

Dassault Systèmes Biovia (2017) Discovery Studio Modeling Environment. https://www.3dsbiovia.com/products/collaborative-science/biovia-discovery-studio/visualization.html Accessed 28 July 2018.

Duchowicz PR, Castro EA, Fernández FM (2006) Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies. MATCH Commun Math Comput Chem 55:179–192

Duchowicz PR, Comelli NC, Ortiz EV, Castro EA (2012) QSAR study for carcinogenicity in a large set of organic compounds. Curr Drug Saf 7:282–288

Duchowicz PR, Fioressi SE, Castro E, Wróbel K, Ibezim NE, Bacelo DE (2017) Conformation-independent QSAR study on human epidermal growth factor receptor-2 (HER2) inhibitors. ChemistrySelect 2:3725–3731

Eriksson L, Jaworska J, Worth AP, Cronin MT, Mcdowell RM, Gramatica P (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs. Environ Health Perspect 111:1361

Faghih-Mirzaei E, Sabouri S, Zeidabadinejad L, Abdolahramazani S, Abaszadeh M, Khodadadi A, Shamsadinipour M, Jafari M, Pirhadi S (2019) Metronidazole aryloxy, carboxy and azole derivatives: synthesis, anti-tumor activity, QSAR, molecular docking and dynamics studies. Bioorg Med Chem 27:305–314

Fink BE, Vite GD, Mastalerz H, Kadow JF, Kim S-H, Leavitt KJ, Du K, Crews D, Mitt T, Wong TW (2005) New dual inhibitors of EGFR and HER2 protein tyrosine kinases. Bioorg Med Chem Lett 15:4774–4779

Fink BE, Norris D, Mastalerz H, Chen P, Goyal B, Zhao Y, Kim S-H, Vite GD, Lee FY, Zhang H (2011) Novel pyrrolo [2, 1-f][1, 2, 4] triazin-4-amines: Dual inhibitors of EGFR and HER2 protein tyrosine kinases. Bioorg Med Chem Lett 21:781–785

Gaber AA, Bayoumi AH, El-Morsy AM, Sherbiny FF, Mehany AB, Eissa IH (2018) Design, synthesis and anticancer evaluation of 1H-pyrazolo [3, 4-d] pyrimidine derivatives as potent EGFRWT and EGFRT790M inhibitors and apoptosis inducers. Bioorg Chem 80:375–395

Gadaleta D, Mangiatordi GF, Catto M, Carotti A, Nicolotti O (2016) Applicability domain for QSAR models: where theory meets reality. Int J Quant Struct-Prop Relat 1:45–63

Gazit A, Osherov N, Posner I, Yaish P, Poradosu E, Gilon C, Levitzki A (1991) Tyrphostins. II. heterocyclic and. alpha.-substituted benzylidenemalononitrile tyrphostins as potent inhibitors of EGF receptor and ErbB2/neu tyrosine kinases. J Med Chem 34:1896–1907

Gazit A, Osherov N, Posner I, Bar-Sinai A, Gilon C, Levitzki A (1993) Tyrphostins. 3. Structure-activity relationship studies of alpha-substituted benzylidenemalononitrile 5-S-aryltyrphostins. J Med Chem 36:3556–3564

Golbraikh A, Tropsha A (2002) Beware ofq2! J Mol Graph Model 20:269–276

Gramatica P (2007) Principles of QSAR models validation: internal and external. QSAR Comb Sci 26:694–701

Gupta A, Bhunia S, Balaramnavar V, Saxena A (2011) Pharmacophore modelling, molecular docking and virtual screening for EGFR (HER 1) tyrosine kinase inhibitors. SAR QSAR Environ Res 22:239–263

Hong H, Xie Q, Ge W, Qian F, Fang H, Shi L, Su Z, Perkins R, Tong W (2008) Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. J Chem Inf Model 48:1337–1344

Jutten B, Keulers TG, Schaaf MB, Savelkouls K, Theys J, Span PN, Vooijs MA, Bussink J, Rouschop KM (2013) EGFR overexpressing cells and tumors are dependent on autophagy for growth and survival. Radiother Oncol 108:479–483

Kalyankrishna S, Grandis JR (2006) Epidermal growth factor receptor biology in head and neck cancer. J Clin Oncol 24:2666–2672

Kroep JR, Linn SC, Boven E, Bloemendal HJ, Baas J, Mandjes IA, Van Den Bosch J, Smit WM, De Graaf H, Schroder CP, Vermeulen GJ, Hop WC, Nortier JW (2010) Lapatinib: clinical benefit in patients with HER 2-positive advanced breast cancer. Neth J Med 68:371–376

Lemmon MA, Schlessinger J (2010) Cell signaling by receptor tyrosine kinases. Cell 141:1117–1134

Levitzki A, Gazit A (1995) Tyrosine kinase inhibition: an approach to drug development. Science 267:1782–1788

Li H-Q, Yan T, Yang Y, Shi L, Zhou C-F, Zhu H-L (2010) Synthesis and structure–activity relationships of N-benzyl-N-(X-2-hydroxybenzyl)-N′-phenylureas and thioureas as antitumor agents. Bioorg Med Chem 18:305–313

Liang K, Esteva FJ, Albarracin C, Stemke-Hale K, Lu Y, Bianchini G, Yang CY, Li Y, Li X, Chen CT, Mills GB, Hortobagyi GN, Mendelsohn J, Hung MC, Fan Z (2010) Recombinant human erythropoietin antagonizes trastuzumab treatment of breast cancer cells via Jak2-mediated Src activation and PTEN inactivation. Cancer Cell 18:423–435

Lv P-C, Zhou C-F, Chen J, Liu P-G, Wang K-R, Mao W-J, Li H-Q, Yang Y, Xiong J, Zhu H-L (2010) Design, synthesis and biological evaluation of thiazolidinone derivatives as potential EGFR and HER-2 kinase inhibitors. Bioorg Med Chem 18:314–319

Marzaro G, Chilin A, Guiotto A, Uriarte E, Brun P, Castagliuolo I, Tonus F, González-Díaz H (2011) Using the TOPS-MODE approach to fit multi-target QSAR models for tyrosine kinases inhibitors. Eur J Med Chem 46:2185–2192

Mastalerz H, Chang M, Chen P, Dextraze P, Fink BE, Gavai A, Goyal B, Han WC, Johnson W, Langley D, Lee FY, Marathe P, Mathur A, Oppenheimer S, Ruediger E, Tarrant J, Tokarski JS, Vite GD, Vyas DM, Wong H, Wong TW, Zhang H, Zhang G (2007) New C-5 substituted pyrrolotriazine dual inhibitors of EGFR and

HER2 protein tyrosine kinases. Bioorg Med Chem Lett 17:2036–2042
Mastalerz H, Chang M, Chen P, Fink BE, Gavai A, Han W-C, Johnson W, Langley D, Lee FY, Leavitt K (2007) 5-((4-Aminopiperidin-1-yl) methyl) pyrrolotriazine dual inhibitors of EGFR and HER2 protein tyrosine kinases. Bioorg Med Chem Lett 17:4947–4954
Mastalerz H, Chang M, Gavai A, Johnson W, Langley D, Lee FY, Marathe P, Mathur A, Oppenheimer S, Tarrant J, Tokarski JS, Vite GD, Vyas DM, Wong H, Wong TW, Zhang H, Zhang G (2007) Novel C-5 aminomethyl pyrrolotriazine dual inhibitors of EGFR and HER2 protein tyrosine kinases. Bioorg Med Chem Lett 17:2828–2833
Mok TS, Wu Y-L, Thongprasert S, Yang C-H, Chu D-T, Saijo N, Sunpaweravong P, Han B, Margono B, Ichinose Y (2009) Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma. N Engl J Med 361:947–957
Nandi S, Bagchi MC (2010) 3D-QSAR and molecular docking studies of 4-anilinoquinazoline derivatives: a rational approach to anticancer drug design. Mol Diversity 14:27–38
Noolvi MN, Patel HM (2010) 3d QSAR studies on a series of quinazoline derrivatives as tyrosine kinase (egfr) inhibitor: the k-nearest neighbor molecular field analysis approach. J Basic Clin Pharm 1:153
Orman JS, Perry CM (2007) Trastuzumab. Drugs 67:2781–2789
Pohlmann PR, Mayer IA, Mernaugh R (2009) Resistance to trastuzumab in breast cancer. Clin Cancer Res 15:7479–7491
Rojas C, Tripaldi P, Duchowicz PR (2016) A new QSPR study on relative sweetness. Int J Quant Struct-Prop Relat 1:78–93
Roy K (2007) On some aspects of validation of predictive quantitative structure–activity relationship models. Expert Opin Drug Discov 2:1567–1577
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
Roy K, Das RN, Ambure P, Aher RB (2016) Be aware of error measures. Further studies on validation of predictive QSAR models. Chemom Intell Lab Syst 152:18–33
Ruslin R, Amelia R, Yamin Y, Megantara S, Wu C, Arba M (2019) 3D-QSAR, molecular docking, and dynamics simulation of quinazoline–phosphoramidate mustard conjugates as EGFR inhibitor. J Appl Pharm Sci 9:089–097
Schlessinger J (2000) Cell signaling by receptor tyrosine kinases. Cell 103:211–225
Shinde MG, Modi SJ, Kulkarni VM (2017) QSAR and molecular docking of phthalazine derivatives as epidermal growth factor receptor (EGFR) inhibitors. J Appl Pharm Sci 7:181–191
Sigismund S, Avanzato D, Lanzetti L (2018) Emerging functions of the EGFR in cancer. Mol Oncol 12:3–20
Singh H, Singh S, Singla D, Agarwal SM, Raghava GP (2015) QSAR based model for discriminating EGFR inhibitors and non-inhibitors using random forest. Biol Direct 10:10
Singh H, Kumar R, Singh S, Chaudhary K, Gautam A, Raghava GP (2016) Prediction of anticancer molecules using hybrid model developed on molecules screened against NCI-60 cancer cell lines. BMC Cancer 16:77
Sun X-Q, Chen L, Li Y-Z, Li W-H, Liu G-X, Tu Y-Q, Tang Y (2014) Structure-based ensemble-QSAR model: a novel approach to the study of the EGFR tyrosine kinase and its inhibitors. Acta Pharm Sin 35:301
Talevi A, Bellera CL, Di Ianni M, Duchowicz PR, Bruno-Blanch LE, Castro EA (2012) An integrated drug development approach applying topological descriptors. Curr Comput Aided Drug Des 8:172–181
Tan X, Lambert PF, Rapraeger AC, Anderson RA (2016) Stress-induced EGFR trafficking: mechanisms, functions, and therapeutic implications. Trends Cell Biol 26:352–366
The Mathworks I (2018) MATLAB 7.0 and Statistics Toolbox 7.1. http://www.mathworks.com Accessed 29 Mar 2019
Toropova A, Toropov A, Martyanov S, Benfenati E, Gini G, Leszczynska D, Leszczynski J (2012) CORAL: QSAR modeling of toxicity of organic chemicals towards Daphnia magna. Chemom Intell Lab Syst 110:177–181
U.S. Environmental Protection Agency (2016) Estimation Programs Interface Suite. https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface Accessed 6 June 2019
Ueno NT, Zhang D (2011) Targeting EGFR in triple negative breast cancer. J Cancer 2:324
Verma RP, Hansch C (2005) An approach toward the problem of outliers in QSAR. Bioorg Med Chem 13:4597–4621
Walker F, Abramowitz L, Benabderrahmane D, Duval X, Descatoire V, Hénin D, Lehy T, Aparicio T (2009) Growth factor receptor expression in anal squamous lesions: modifications associated with oncogenic human papillomavirus and human immunodeficiency virus. Hum Pathol 40:1517–1527
Wold S, Eriksson L, Clementi S (1995) Statistical validation of QSAR results. In: van de Waterbeemd H (ed) Chemometric methods in molecular design. Wiley-VCH, Weinheim, p 309–338
Wu P, Nielsen TE, Clausen MH (2016) Small-molecule kinase inhibitors: an analysis of FDA-approved drugs. Drug Discov Today 21:5–10
Xu G, Abad MC, Connolly PJ, Neeper MP, Struble GT, Springer BA, Emanuel SL, Pandey N, Gruninger RH, Adams M (2008) 4-Amino-6-arylamino-pyrimidine-5-carbaldehyde hydrazones as potent ErbB-2/EGFR dual kinase inhibitors. Bioorg Med Chem Lett 18:4615–4619
Xu G, Searle LL, Hughes TV, Beck AK, Connolly PJ, Abad MC, Neeper MP, Struble GT, Springer BA, Emanuel SL (2008) Discovery of novel 4-amino-6-arylaminopyrimidine-5-carbaldehyde oximes as dual inhibitors of EGFR and ErbB-2 protein tyrosine kinases. Bioorg Med Chem Lett 18:3495–3499
Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474