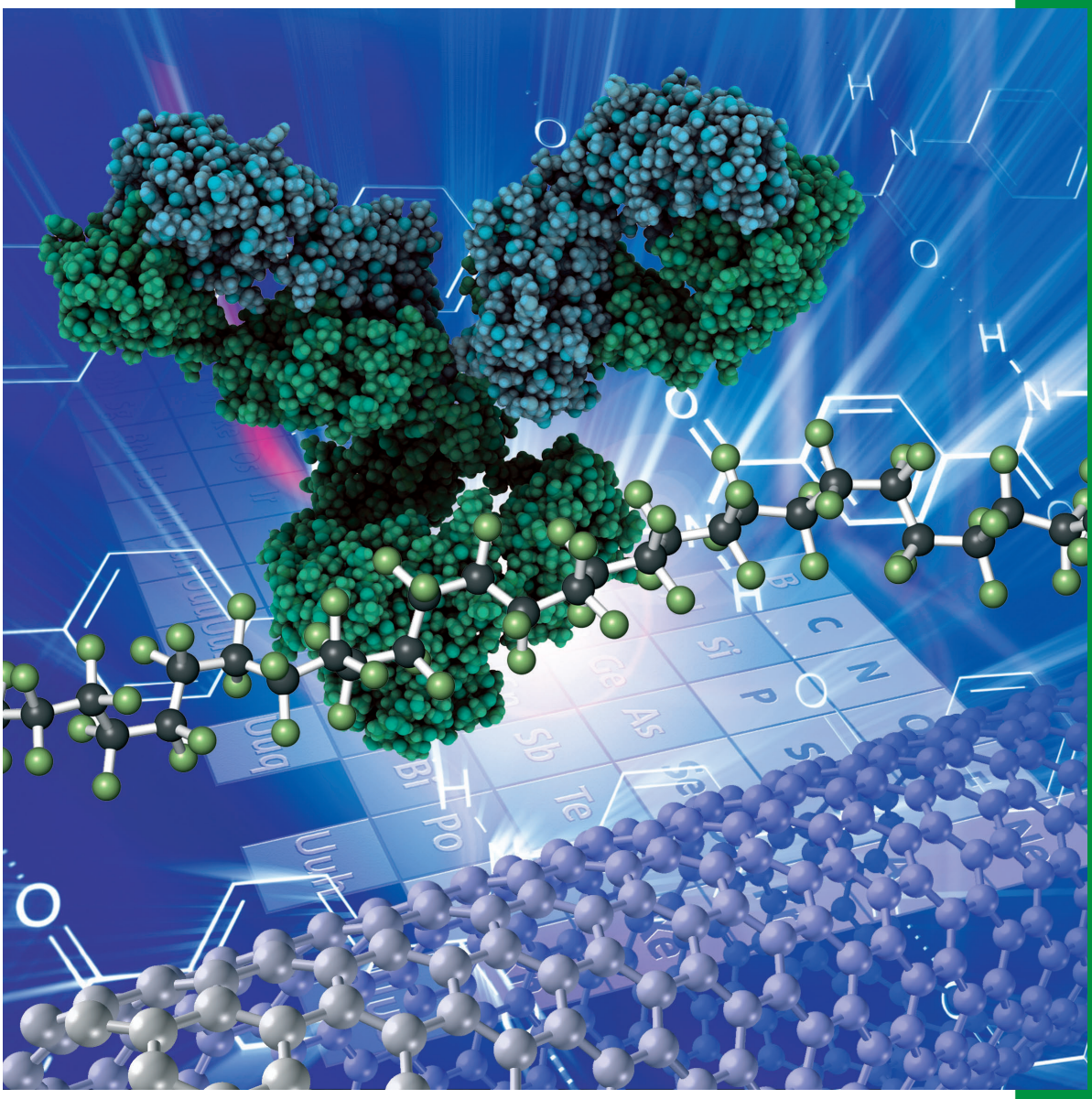
Chemistry SELECT

www.chemistryselect.org

A journal of ChemPubSoc Europe



REPRINT

WILEY-VCH

ChemPubSoc Europe

DOI: 10.1002/slct.201700436

Chemistry SELECT Full Papers

Medicinal Chemistry & Drug Discovery

Conformation-Independent QSAR Study on Human Epidermal Growth Factor Receptor-2 (HER2) Inhibitors

Pablo R. Duchowicz,*[a] Silvina E. Fioressi,*[b] Eduardo Castro,[a] Karolina Wróbel,[c] Nnenna E. Ibezim,[d] and Daniel E. Bacelo[b]

Inhibition of HER2 (human epidermal growth factor receptor 2) expression and function is required in several cancer treatments. Numerous compounds with very different molecular structures have been suggested as HER2 inhibitors. Here we perform quantitative structure-activity relationship (QSAR) analysis on 444 of such compounds to investigate the molecular properties that may influence its efficiency. Models based on 1D and 2D flexible molecular descriptors are proposed to develop simple models based solely on constitutional and topological molecular features. A large number of non-conformational descriptors (17974) was used to thoroughly explore the structural characteristics that influence the HER2 inhibitory activity. Three different approaches were explored using: 1) Molecular Descriptors, 2) Flexible Molecular Descriptors, and 3) Hybrid Descriptors. A QSAR model for HER2 inhibitors was successfully developed. Some properties such as electronegativity, aromatic character, and the presence of amino groups appear as molecular characteristics that may have influence in the HER2 inhibitory activity.

Introduction

Several human cancers are categorized by elevated levels of proteins that regulate cell cycle progression and proliferation. HER2 are constituents of the epidermal growth factor receptor tyrosine kinase protein family which includes HER1/EGFR, HER2/ErbB2, HER3/ErbB3, and ErbB4.[1] This family of receptors is among the most investigated cell signaling families in cancer research.[2] Receptor over-production is caused by abnormalities in HER2 gene regulation, resulting in various cancers including breast,[3] prostate,[4] ovarian,[5] and gastric cancer[6]. In these cases treatment of the cancer and prevention of its spread must include inhibition of HER2 expression and function. For breast cancer, Trastuzumab (HerceptinH)[7,8] has proven to be efficient in overexpression inhibition of HER2 and Lapatinib (TykerbH)[9,10] is a second-line treatment for patients who are refractory to Trastuzumab and chemotherapy.

The discovery of new molecular cancer drug targets and the development of specific agents directed to these targets is an active area of research. At the present different families of structures have been suggested as HER2 inhibitors and their efficiencies reported.[11–31] The QSAR statistical approach, which correlates quantitative response activities with numerical descriptors from a set of training molecules, has proved to be an essential technique in the discovery of new drugs.[32–34] Trustworthy models can improve drastically the proficiency and pace of detection of more effective drugs with weaker secondary effects. Zhu and coworkers developed a 3D-QSAR model using sets of 12 molecules of 3-substituted indolin-2-ones and 19 compounds of benzylidene malononitriles with low-to-high affinity for HER2.[35] Docking and 3D- QSAR analyses were employed to explore differences in binding mode preferences at the ATP site and the selectivity of the dihydroxy compounds as inhibitors of HER-2 using 50 benzylidene malonitrile tyrphostins.[36] A set of 32 C4- and C5- substituted pyrrolotriazines showing inhibition activity toward HER2 protein tyrosine kinases were studied by 3D-QSAR comparative molecular field analysis (CoMFA); the model found showed good predictive ability.[37] The investigation of QSAR models and pharmacophore features for designing HER2/HSP90 dual-targeted inhibitors was reported by Chen and Chen. Their models based on 48 compounds proved highly predictive, with correlation coefficients (r^2) in the range of 0.93-0.96.[38] The same research group proposed a 3D-QSAR model of 36 protein kinase inhibitor ligands. In this case CoMFA models and comparative molecular similarity indices analysis yielded r^2 values of 0.9547 and 0.9226, respectively.[39] They also published multiple linear regression (MLR) and support vector machine (SVM) models constructed from an extensive set of 298 ligands.

[a] Dr. P. R. Duchowicz, Dr. E. Castro
Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (CCT La Plata-CONICET, UNLP)
Diag. 113 y 64, Sucursal 4, C.C. 16, 1900 La Plata, Argentina
E-mail: pabloducho@gmail.com
[b] Dr. S. E. Fioressi, Dr. D. E. Bacelo
Departamento de Química
Facultad de Ciencias Exactas y Naturales
Universidad de Belgrano
Villanueva 1324 CP 1426, Buenos Aires, Argentina
E-mail: sfioressi@yahoo.com
[c] K. Wróbel
Medical University of Lodz
Kosciuszki 4, 90-419 Lodz, Poland
[d] Dr. N. E. Ibezim
Department of Computer Education
University of Nigeria
Nsukka, Nigeria

Supporting information for this article is available on the WWW under https://doi.org/10.1002/slct.201700436

The reported r^2 values were 0.795 for MLR and 0.862 for SVM.[40] Self-Organizing Molecular Field Analysis (SOMFA), a grid-based and alignment-dependent 3D-QSAR method, was applied to a series of 24 new quinazoline derivatives, and r^2 of 0.815 was obtained.[41] Recently 115 HER2/ErbB2 inhibitors were studied with 3D-pharmacophoric descriptors in QSAR analysis; ligand efficiency as the response variable was applied because the logarithmic transformation of bioactivities failed to provide self-consistent QSAR models.[42]

The aim of the present study is to analyze by QSAR the inhibitory HER2 activity of compounds reported in the literature to identify molecular properties that influence activity in order to contribute to the design of compounds more effective in treatment. The objective is to find simple models based on a sizable, varied set of molecules, employing a large and heterogeneous set of descriptors. Therefore, QSAR models based on 1D and 2D flexible molecular descriptors are proposed to develop simple models based solely on constitutional and topological molecular characteristics.[43,44] Exclusion of 3D structural aspects avoids the ambiguities arising from the existence of various conformational states. On the other hand exclusion of quantum-chemical descriptors avoids the high computational costs and long calculation times associated with the calculations of optimal molecular geometries. Three different QSAR approaches were explored to develop models for the prediction of the inhibitory activity of HER2. In one approach the popular freely available descriptor generators PaDEL-Descriptor (version 2.20),[45,46] EPI Suite,[47] and Mold2[48] were used to generate 0D and 1D descriptors and fingerprints. In the other approach the CORALSEA program[49,50] was used to obtain flexible descriptors. Finally we also explored models combining both sets of descriptors. Using those descriptors, simple models based on from 1–6 descriptors have been chosen as the best predictive combinations of independently selected variables.

2. Methodology

QSAR analysis was performed on 444 HER2 inhibitors (Table 1S). Their structures and in vitro activities were collected from recently published literature.[11–31] The bioactivities were expressed as concentrations of the test compounds that inhibited the activity of HER2 by 50% (IC_{50}). The IC_{50} values were converted into molarities (M), and then the logarithmic (pIC_{50}, M) values were used in the QSAR analysis.

2.1 Structural representation and molecular descriptors calculation

The structures of the compounds were generated in both SMILES notations and 2D structures drawn with Discovery Studio (Version 3.5) freeware,[51] without performing geometrical optimizations, and saved in MDL-MOL format. The descriptors were calculated using two different methodologies: a) Theoretical conformation-independent molecular descriptors and fingerprints were calculated using the freely-available software PaDEL-Descriptor (version 2.20),[45,46] EPI Suite,[47] and Mold2[48]. In total, 1444 1D and 2D descriptors and 12 types of fingerprints (16092) were obtained from Padel-Descriptor, 184 descriptors from EPI Suite, and 254 descriptors from Mold2. In total a large number of non-conformational descriptors (17974) were used to thoroughly explore the structural characteristics that influence HER2 inhibitory activity. Constant values and descriptors found to be linearly-dependent were identified and excluded from the original matrix of variables to minimize redundant information. b) Flexible molecular descriptors were calculated with CORAL freeware.[49,50] At first 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 (HSG), hydrogen-filled graph (HFG) or graph of atomic orbitals (GAO); ii. SMILES; and iii. a hybrid of chemical graph and SMILES.[50] The most appropriate combination of structural attributes (local descriptors, SA) should be chosen for modeling because the selected SR defines the number and types of local descriptors to be included in the QSAR analysis. The CORAL framework searches for a QSAR model through a one-variable linear correlation between pIC_{50} and a properly defined flexible descriptor (DCW). The DCW descriptor is a linear combination of special coefficients called correlation weights (CW). A CW value is calculated for each SA type in the training set. The CW values for all the structural attributes are calculated via Monte Carlo (MC) simulation, searching for the highest correlation coefficient (r) between pIC_{50} and the DCW descriptor (Table 2S). The DCW depends upon the threshold value (T) and the number of epochs or iterations (N_{epochs}) used.[52] These parameters are positive integers from the MC method that must be specified in order to calculate the DCW values. T defines rare (noise) SMILES attributes that do not contribute to the predicted inhibitory activity, so that 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.

2.2 Model Validation

To verify the predictive capability of the proposed models the dataset was split into a training set (148 compounds) for model development, a validation set (148 compounds) for model validation, and a test set (148 compounds) for external validation. Randomly splitting a dataset may not lead to rational solutions unless the generated sets have similar structure-activity relationships. To this end the split of the dataset is carried out by the Balanced Subsets Method (BSM),[53–55] a technique based on k-Means Cluster Analysis (k-MCA). The procedure of BSM ensures that the training set is representative of the validation and test sets.

The Replacement Method (RM) technique[56–62] was applied to generate Multivariable Linear Regression (MLR) models on the training set. The algorithms used in our calculations were programmed in MATLAB software.[63] The MLR models were validated theoretically through the Leave-One-Out Cross Validation (*loo*) method to measure the stability of the QSAR model

Table 1. Descriptors identified for modeling inhibitory HER2 activity together with the squared correlation coefficient and the standard deviation for the training, validation, and test sets. The best model is in bold text.

#Des.	Descriptors	R^2_{Train}	S_{Train}	R^2_{Val}	S_{Val}	R^2_{Test}	S_{Test}
1	*PubchemFP539*	0.53	0.92	0.53	0.91	0.47	0.94
2	*PubchemFP192, PubchemFP539*	0.60	0.84	0.60	0.84	0.54	0.88
3	*GATS5c, PubchemFP192, PubchemFP539*	0.64	0.81	0.64	0.79	0.57	0.85
4	*minHdsCH, MIC2, PubchemFP192, PubchemFP539*	0.68	0.76	0.67	0.77	0.54	0.85
5	*ATSC5i, VE1_Dzs, MACCSFP28, PubchemFP192, PubchemFP539*	0.72	0.72	0.70	0.74	0.66	0.76
6	***ATSC5 i, VE1_Dzs, NaasC, MACCSFP28, PubchemFP539, APC2D9_N_O***	**0.76**	**0.67**	**0.72**	**0.71**	**0.64**	**0.79**
7	*SpMin1_Bhp, minHdsCH, MACCSFP28, VR2_D, PubchemFP192, PubchemFP435, PubchemFP539*	0.78	0.65	0.72	0.72	0.64	0.80

upon inclusion/exclusion of molecules. A general criterion is to validate the model if the *loo* variance (R^2_{loo}) is greater than 0.5. Since this is merely a necessary, not sufficient, condition for predictive power[64], more thorough validation was sought using the external test set of 148 compounds and the training set of 148. The Y-Randomization method[65] was used to scramble the experimental values such that they do not correspond to their respective compounds to check that the model does not result from happenstance.

Mean absolute error (MAE) was the criterion applied to evaluate the predictive error values relative to the training set response range.[66] A MAE of 10% of the training set range indicates that the model is adequately predictive. Comparison of error-based metrics values after eliminating a defined small fraction of compounds with high residual values of the test set avoids the effect of any rare prediction error that can diminish the quality of predictions for the whole test set. To evaluate the predictive capability of a model the percent of compounds with high residual values of the test set that need to be removed to accomplish the value of the MAE + 3σ to be less than 25% of the training set range was also assessed (σ the standard deviation of absolute error values for the test set).

The reliability of the predictions was confirmed by the determination of the error bias of the model.[67] This verifies the absence of systematic error, assuring that error values lie both above and below the null residual. The freeware XternalValidationPlus v.1.2[68] and Bias-Variance Estimator[69] were utilized for bias error determination of the models.

2.3 Applicability Domain

No QSAR model is expected to reliably predict studied activity for the universe of molecules. The applicability domain (AD) is a theoretically defined area that depends on the molecular descriptor values and the experimental activity analyzed.[70,71] Only molecules falling within this AD are not considered model extrapolations. The ADs for the proposed models were determined though two methodologies: the leverage approach[72] and a simple standardization method[73]. 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 regarded as substantially a model extrapolation and not reliable. In addition the recently proposed method for identifying compounds outside the applicability domain developed by Roy and coworkers was applied.[73] The standalone application named "Applicability domain using standardization approach" was used.[74]

3. Results and Discussion

Three different approaches were explored for the prediction of inhibitory activity of HER2: 1) Molecular Descriptors Model, 2) Flexible Molecular Descriptors Model, and 3) Hybrid Descriptors Model. The corresponding **s**tatistical parameters are provided as Supplementary Information (Tables 3S to 11S) and the results for each model are discussed in the following sections. As a general methodology, once the predictive ability of the molecular descriptors has been verified, the models are used for the experimental data in the test sets, in order to fully exploit the available structural and response information and to enlarge the applicability domain of the model.

3.1 Molecular Descriptors Models

Table 1 shows the results obtained for the best models found using the molecular descriptors and fingerprints. Models involving from one to seven descriptors were explored and the best predictive performance is observed for the model involving six descriptors. Figure 1 shows the calculated pIC_{50} versus the experimental values for this model, represented by the equation:

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}
^3S_k	0.85	0.51	0.85	0.53	0.68	0.75
^2S_k, ^3S_k	0.84	0.53	0.83	0.55	0.71	0.70
^1S_k, ^2S_k, ^3S_k	0.85	0.52	0.84	0.54	0.71	0.70
$Pt2_k$	0.75	0.63	0.70	0.72	0.69	0.72
$Pt2_k$, ^2EC_j	0.84	0.53	0.80	0.58	0.69	0.72
$Pt2_k$, ^0EC_j, ^1EC_j, ^2EC_j	0.84	0.53	0.79	0.61	0.72	0.69
^1S_k, ^3S_k, $Pt2_k$	0.85	0.51	0.85	0.53	0.71	0.69
^1S_k, ^2S_k, ^3S_k, $Pt2_k$, ^0EC_j, ^1EC_j, ^2EC_j	**0.88**	**0.47**	**0.87**	**0.48**	**0.72**	**0.68**

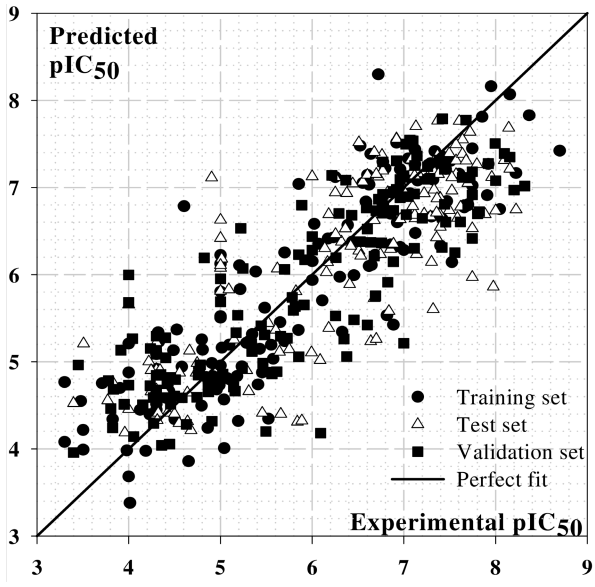


Figure 1. Predicted and experimental values for the training, validation and test sets for the six-descriptor model (Equation 1) for 444 HER2 inhibitors.

$$pIC_{50} = 3.70 + 0.0194\ ATSC5i + 0.921\ VE1\_Dzs + 0.223\ naasC + 2.12\ MACCSFP28 + 0.812\ PubchemFP539 - 0.230\ APC2D9\_N\_O \quad (1)$$

$N_{train} = 148$, $R^2_{train} = 0.76$, $S_{train} = 0.67$, $N_{val} = 148$, $R^2_{val} = 0.72$, $S_{val} = 0.71$, $F = 73$

$N_{test} = 148$, $R^2_{test} = 0.64$, $S_{test} = 0.79$, $o(2.5S) = 1$, $R^2_{Loo} = 0.73$ $S_{Loo} = 0.70$, $S_{rand} = 1.22$,

$h^* = 0.071$, MAE = 0.6, Train range = 5.40, Variance = 0.498, $Bias^2 = 0.59$

Here, F is the Fisher parameter and $o(2.5S)$[75] indicates the number of outlier compounds in the training set having a residual (difference between experimental and calculated pIC_{50}) greater than 2.5 times S_{train} and lower than 2.5 times S_{train}. The two applied methodologies found that two compounds (370 and 375) fall out of the applicability domain in this model. The MAE result including all test compounds is slightly over 10% of the training compounds' value range. To meet the established measure of $MAE + 3\sigma < 25\%$ it is necessary to eliminate 15% of the high residual values of the test set. Then, by the MAE criterion, this model is weak though it does not present systematic errors.[66] Figure 1S presents a dispersion plot of predicted values. Equation 1 satisfies the external validation conditions.[76]

$$1 - \frac{R_0^2}{R_{test}^2} < 0.1\ (0.00003)\ or\ 1 - \frac{R_0'^2}{R_{test}^2} < 0.1\ (0.21)\ and, 0.85 \le k \le 1.15\ (1.01)\ and,$$

$$0.85 \le k' \le 1.15\ (0.98)\ and, R_m^2 > 0.5\ (0.68)$$

The descriptors *ATSC5i*, and *VE1_Dzs* in the proposed models are 2D-autocorrelation descriptors originating in autocorrelation of topological structure of Broto-Moreau (ATS), and the last eigenvector from the Barysz matrix. The autocorrelation descriptors encode both the molecular structure and a physicochemical property as a vector. As a result these descriptors associate the topology of a structure with a selected physicochemical property. The two indices following the descriptor symbol represent the topological distance between pairs of atoms, or lag, and the physicochemical property considered in the weighting component for its computation. For example, the *ATSC5i* descriptor represents the Centered Broto-Moreau autocorrelation with lag 5 weighted by first ionization potential. The *VE1_Dzs* is a coefficient derived from the last eigenvector of the Barysz matrix weighted by I-state descriptors.

In the above model the *ATSC5i* and the *VE1_Dzs* descriptors both correlate positively to inhibitory activity. *NaasC* represents the number of aromatic carbons bonded with non-H atoms, and also presents a positive correlation. The model is completed by the positive contribution of two fingerprints, *MACCSFP28* (MACCS keys QCH2Q), *PubchemFP539* (N=C-C-[#1]), and the negative correlation of *APC2D9_N_O* (count of N–O at topological distance 9). *MACCSFP28* fingerprint accounts for a CH_2 moiety bonded to heteroatoms, for example nitrogen, whereas *PubchemFP539* is indicator of a nitrogen double-bonded to a carbon, as in an aromatic or tertiary amine. Recently secondary and tertiary amino groups have been 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.[77] The molecular descriptors and fingerprints appearing in the model of Eq. 1 suggest that the inhibitory activity of HER2 is affected by the aromatic character of the compounds, its ionization potential and the presence of nitrogen atoms in the structure, particularly those forming amino groups and not associated directly to oxygen.

3.2 Flexible Molecular Descriptors Model

The QSAR analysis was performed by searching the best linear regression models on the training set of 148 compounds. The most efficient structural attributes for each SR are searched by optimizing the DCW flexible descriptor by increasing R^2_{train} until the model begins to lose predictive capability in the validation set. The same procedure is followed when the most predictive model must be selected among several MRL, with the descriptors searched in a pool of thousands.[78] The test set was not involved in the development of the model. Table 3 contains a summary for the statistical quality of the best QSAR models found by trying different possible CORAL combinations. It reveals that the best choice is an approach that includes both graph and SMILES representations. The optimal descriptor involves seven variable types, and 775 active attributes are based on them (shown in Table 10S). The predicted and experimental values for the training, validation, and test sets follow a straight line (Figure 2), and the residuals are shown in the Figure 2S. The resultant equation for this model using one DCW is:

Table 3. Descriptors identified for modeling inhibitory HER2 activity together with the squared correlation coefficient and the standard deviation for training, validation, and test sets.

#Des.	Descriptors	R^2_{Train}	S_{Train}	R^2_{Val}	S_{Val}	R^2_{Test}	S_{Test}
1	*Coral*	0.88	0.47	0.87	0.48	0.72	0.68
2	*Coral, Ssl*	0.88	0.46	0.87	0.48	0.73	0.68
3	*Coral, ALogP, KRFP1811*	0.88	0.47	0.87	0.49	0.73	0.68
4	*Coral, ALogP2, AMR, KRFP2595*	0.89	0.45	0.86	0.52	0.71	0.71
5	***Coral, SpMax4_Bhv, SpMax3_Bhe, SpMax5_Bhi, APC2D3_C_N***	**0.90**	**0.43**	**0.87**	**0.48**	**0.74**	**0.66**
6	*Coral, ALogP2, SpMax4_Bhv, SpMax3_Bhe, SpMax5_Bhi, Ssl*	0.91	0.42	0.88	0.48	0.75	0.66
7	*Coral, SpMax3_Bhp, SpMax4_Bhp, MDEC-22, MLFER_L, SubFP17 KRFP142*	0.91	0.41	0.85	0.54	0.75	0.66

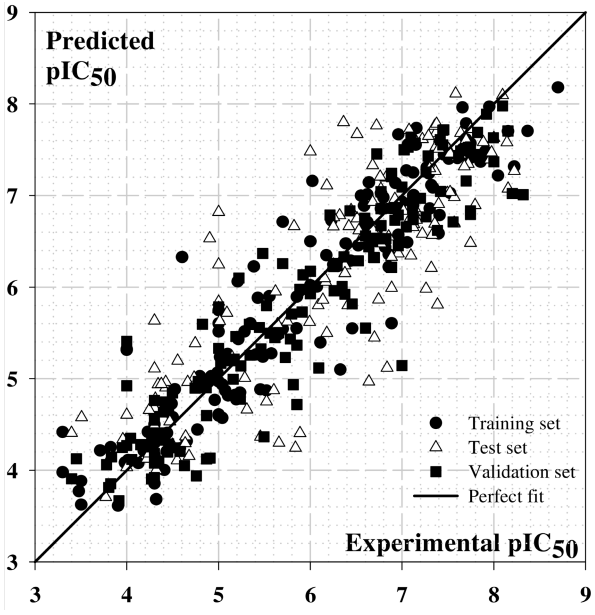


Figure 2. Predicted and experimental values for the training, validation and test sets for the one- flexible-descriptor model (Equation 2) for 444 HER2 inhibitors.

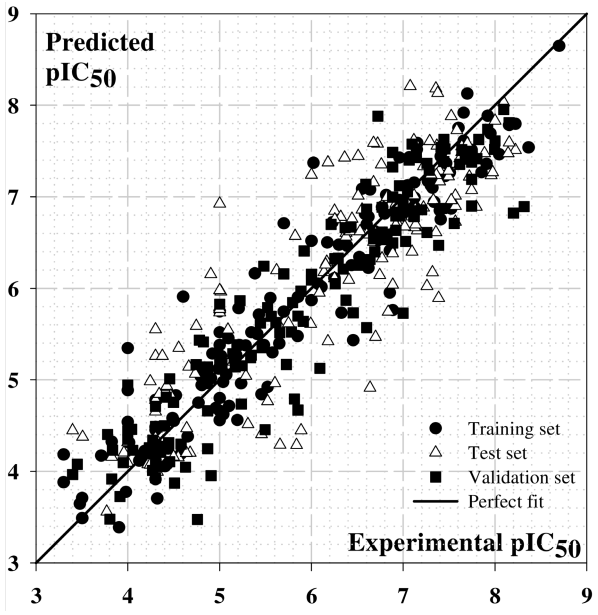


Figure 3. Predicted and experimental values for the training, validation and test set for the five-descriptors hybrid model (Equation 3) for 444 HER2 inhibitors.

$$pIC_{50} = 2.54 + 0.033 * DCW \quad (2)$$

$N_{train} = 148$, $R^2_{train} = 0.88$, $S_{train} = 0.47$, $N_{val} = 148$, $R^2_{val} = 0.87$, $S_{val} = 0.48$, $F = 1033$

$N_{test} = 148$, $R^2_{test} = 0.72$, $S_{test} = 0.68$, $o(2.5S) = 5$, $R^2_{Loo} = 0.87$ $S_{Loo} = 0.48$, $S_{rand} = 1.26$,

h* = 0.02, MAE = 0.53, Train range = 5.40, Variance = 0.002, $Bias^2 = 0.46$

All compounds are within the applicability domain according to both methods used, and systematic error is absent. The MAE calculated including all test compounds lower than 10% of the training compounds' value range. It is necessary to exclude 9% of the high residual compounds of the test set to accomplish the provision of MAE + 3σ < 25% of the training set range. We apply both Y-randomization to demonstrate that $S_{train} < S^{rand}$, and also the external validation criterion[76] to ensure that a valid structure-activity relationship is achieved:

$$1 - \frac{R_0^2}{R_{test}^2} < 0.1\ (0.07)\ or\ 1 - \frac{R_0'^2}{R_{test}^2} < 0.1\ (0.008)\ and, 0.85 \le k \le 1.15\ (0.98)\ and,$$

$$0.85 \le k' \le 1.15\ (1.00)\ and, R_m^2 > 0.5\ (0.57)$$

The parameters used for model building were $T = 5$ and $N_{epochs} = 50$. Table 11S includes an example of a DCW calculation for compound **1.** 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.

3.3 Hybrid Descriptors Model

Finally calculations combining PaDEL, EPI Suite, Mold2, and the flexible CORAL descriptors and fingerprints were explored. The combination of various flexible descriptors or flexible descriptors with molecular descriptors create models with better predictive quality, with however, significant increase in complexity. Best results were achieved with five descriptors (see Table 3). Figures 3 and 3S show the predicted and experimental

values for the training, validation, and test sets using the best model represented by the equation:

$$pIC_{50} = 1.111\text{-}1.877\ Coral + 1.344\ SpMax4\_Bhv + 0.798\ SpMax3\_Bhe + 0.012\ SpMax5\_Bhi + 0.0320\ APC2D3\_C\_N \quad (3)$$

$N_{train} = 148$, $R^2_{train} = 0.90$, $S_{train} = 0.43$, $N_{val} = 148$, $R^2_{val} = 0.87$, $S_{val} = 0.48$, $F = 254$

$N_{test} = 148$, $R^2_{test} = 0.74$, $S_{test} = 0.66$, $o(2.5S) = 4$, $R^2_{Loo} = 0.89$ $S_{Loo} = 0.45$, $S_{rand} = 1.22$,

$h^* = 0.061$, MAE = 0.52, Train range = 5.40, Variance = 0.006, $Bias^2 = 0.42$

According to the two methods used all the compounds are within the applicability domain and no systematic error was observed. The MAE result including all test compounds is lower than 10% of the training compounds' value range and 7% of the test set compounds had to be removed to achieve a MAE + 3σ value lower than 25%.

The flexible descriptor of the best model found for *CORAL* also appears in the most predictive hybrid model. The other four descriptors for equation 3 are: *SpMax4_Bhv* (Largest absolute eigenvalue of Burden modified matrix - n 4 weighted by relative van der Waals volumes), *SpMax3_Bhe* (Largest absolute eigenvalue of Burden modified matrix - n 3 weighted by relative Sanderson electronegativities), *SpMax5_Bhi* (Largest absolute eigenvalue of Burden modified matrix - n 5 weighted by relative first ionization potential), and *APC2D3_C_N* (Count of atom pairs C–N at topological distance 3). Analysis of the descriptors involved in Eq 3, as in the Eq 1 model, shows that ionization potential, electronegativity, and steric factor influence HER2 inhibitory activity. This finding agrees with those of Chen and Chen in CoMFA analysis of 48 purine-based molecules.[38] They propose that steric impediment and hydrogen bonding should be taken into account in designing efficient HER2 inhibitors. Equation 3 also satisfies the external validation conditions.[76]

$$1 - \frac{R_0^2}{R_{test}^2} < 0.1\ (0.006)\ or\ 1 - \frac{R_0'^2}{R_{test}^2} < 0.1\ (0.06)\ and, 0.85 \le k \le 1.15\ (1.00)\ and,$$

$$0.85 \le k' \le 1.15\ (0.99)\ and, R_m^2 > 0.5\ (0.69)$$

4. Conclusions

In this work, we have developed a structure-inhibitory activity relationship for HER2 inhibitors through a computational technique that does not require knowing the molecular conformation as part of the structural representation. A model that uses traditional molecular descriptors gives slightly better results than one involving one flexible descriptor from the CORALSEA program, however, the former model is but poorly validated by MAE metrics. Better performance is obtained with a hybrid approach that combines the optimal CORAL descriptor, three PaDel descriptors, and one fingerprint. This model was validated through Y-randomization, cross-validation and MAE criteria, and satisfies Applicability Domain analysis.

The descriptors involved in the models here proposed suggest that ionization potential, aromatic character of the molecules studied and the presence of amino groups seem to have influence in HER2 inhibitory activity. This could be useful in development of new anticancer drugs that efficiently inhibit HER2 expression and function.

Acknowledgments

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. EAC, DEB, SEF and PRD are members of the scientific researcher career of CONICET.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: Cancer · HER2 · QSAR · Tyrosine kinase protein

[1] C. L. Arteaga, J. A. Engelman, *Cancer Cell* **2014**, *25*, 282–303.
[2] J. Schlessinger, *Cell* **2000**, *103*, 211–225.
[3] R. H. Engel, V. G. Kaklamani, *Drugs* **2007**, *67*, 1329–1341.
[4] S. Signoretti, R. Montironi, J. Manola, A. Altimari, C. Tam, G. Bubley, S. Balk, G. Thomas, I. Kaplan, L. Hlatky, P. Hahnfeldt, *J. Natl. Cancer Inst.* **2000**, *92*, 1918–1925.
[5] K. D. Steffensen, M. Waldstrom, U. Jeppesen, E. Jakobsen, I. Brandslund, A. Jakobsen, *Int. J. Gynecol. Cancer* **2007**, *17*, 798–807.
[6] K. Sakai, S. Mori, T. Kawamoto, S. Taniguchi, O. Kobori, Y. Morioka, T. Kuroki, K. Kano, *J. Natl. Cancer Inst.* **1986**, *77*, 1047–1052.
[7] J. S. Orman, C. M. Perry, *Drugs* **2007**, *67*, 2781–2789.
[8] K. Liang, F. J. Esteva, C. Albarracin, K. Stemke-Hale, Y. Lu, G. Bianchini, C. Y. Yang, Y. Li, X. Li, C. T. Chen, G. B. Mills, *Cancer Cell* **2010**, *18*, 423–435.
[9] M. P. Curran, *Drugs* **2010**, *70*, 1411–1422.
[10] J. R. Kroep, S. C. Linn, E. Boven, H. J. Bloemendal, J. Baas, I. A. Mandjes, J. van den Bosch, W. M. Smit, H. de Graaf, C. P. Schröder, G. J. Vermeulen, *Neth. J. Med.* **2010**, *68*, 371–376.
[11] A. Gazit, N. Osherov, I. Posner, P. Yaish, E. Poradosu, C. Gilon, A. Levitzki, *J. Med. Chem.* **1991**, *34*, 1896–1907.
[12] A. Gazit, N. Osherov, I. Posner, A. Bar-Sinai, C. Gilon, A. Levitzki, *J. Med. Chem.* **1993**, *36*, 3556–3564.
[13] A. Levitzki, A. Gazit, *Science* **1995**, *267*, 1782–1788.
[14] Y. Yarden, *Eur. J. Cancer* **2001**, *37*, S3-S8.
[15] K. M. Ferguson, M. B. Berger, J. M. Mendrola, H. S. Cho, D. J. Leahy, M. A. Lemmon, *Molecular Cell* **2003**, *11*, 507–517.
[16] B. E. Fink, G. D. Vite, H. Mastalerz, J. F. Kado, S. H. Kim, K. J. Leavitt, K. Du, D. Crews, T. Mitt, T. W. Wong, J. T. Hunt, D. M. Vyas, J. Tokarski, *Bioorg. Med. Chem. Lett.* **2005**, *15*, 4774–4779.
[17] L. Llauger, H. He, J. Kim, J. Aguirre, N. Rosen, U. Peters, P. Davies, G. Chiosis, *J. Med. Chem.* **2005**, *48*, 2892–2905.
[18] H. He, D. Zatorska, J. Kim, J. Aguirre, L. Llauger, Y. She, N. Wu, R. M. Immormino, D. T. Gewirth, G. Chiosis, *J. Med. Chem.* **2006**, *49*, 381–390.
[19] H. Mastalerz, M. Chang, P. Chen, P. Dextraze, B. E. Fink, A. Gavai, B. Goyal, W. Han, W. Johnson, D. Langley, F. Y. Lee, P. Marathe, A. Mathur, S. Oppenheimer, E. Ruediger, J. Tarrant, J. S. Tokarski, G. D. Vite, D. M. Vyas, H. Wong, T. W. Wong, H. Zhang, G. Zhang, *Bioorg. Med. Chem. Lett.* **2007**, *17*, 2036–2042.
[20] H. Mastalerz, M. Chang, A. Gavai, W. Johnson, D. Langley, F. Y. Lee, P. Marathe, A. Mathur, S. Oppenheimer, J. Tarrant, J. S. Tokarski, G. D. Vite,

D. M. Vyas, H. Wong, T. W. Wong, H. Zhang, G. Zhang, *Bioorg. Med. Chem. Lett.* **2007**, *17*, 2828–2833.
[21] H. Mastalerz, M. Chang, P. Chen, B. E. Fink, A. Gavai, W. C. Han, W. Johnson, D. Langley, F. Y. Lee, K. Leavitt, P. Marathe, D. Norris, S. Oppenheimer, B. Sleczka, J. Tarrant, J. S. Tokarski, G. D. Vite, D. M. Vyas, H. Wong, T. W. Wong, H. Zhang, G. Zhang, *Bioorg. Med. Chem. Lett.* **2007**, *17*, 4947–4954.
[22] Z. Y. Cheng, W. J. Li, F. He, J. M. Zhou, X. F. Zhu, *Bioorg. Med. Chem.* **2007**, *15*, 1533–1538.
[23] B. Lippa, G. S. Kauffman, J. Arcari, T. Kwan, J. Chen, W. Hungerford, S. Bhattacharya, X. Zhao, C. Williams, J. Xiao, L. Pustilnik, *Bioorg. Med. Chem. Lett.* **2007**, *7*, 3081–3086.
[24] T. V. Hughes, G. Xu, S. K. Wetter, P. J. Connolly, S. L. Emanuel, P. Karnachi, S. R. Pollack, N. Pandey, M. Adams, S. Moreno-Mazza, S. A. Middleton, L. M. Greenberger, *Bioorg. Med. Chem. Lett.* **2008**, *18*, 4896–4899.
[25] G. Xu, M. C. Abad, P. J. Connoll, M. P. Neeper, G. T. Struble, B. A. Springer, S. L. Emanuel, N. Pandey, R. H. Gruninger, M. Adams, S. Moreno-Mazza, *Bioorg. Med. Chem. Lett.* **2008**, *18*, 4615–4619.
[26] G. Xu, L. L. Searle, T. V. Hughes, A. K. Beck, P. J. Connolly, M. C. Abad, M. P. Neeper, G. T. Struble, B. A. Springer, S. L. Emanuel, R. H. Gruninger, *Bioorg. Med. Chem. Lett.* **2008**, *18*, 3495–3499.
[27] X. Cai, H. X. Zhai, J. Wang, J. Forrester, H. Qu, L. Yin, C. J. Lai, R. Bao, C. Qian, *J. Med. Chem.* **2010**, *53*, 2000–2009.
[28] H. Q. Li, T. Yan, Y. Yang, L. Shi, C. F. Zhou, H. L. Zhu, *Bioorg. Med. Chem.* **2010**, *18*, 305–313.
[29] S. D. Fidanze, S. A. Erickson, G. T. Wang, R. Mantei, R. F. Clark, B. K. Sorensen, N. Y. Bamaung, P. Kovar, E. F. Johnson, K. K. Swinger, K. D. Stewart, Q. Zhang, L. A. Tucker, W. N. Pappano, J. L. Wilsbacher, J. Wang, G. S. Sheppard, R. L. Bell, S. K. Davidsen, R. D. Hubbard, *Bioorg. Med. Chem. Lett.* **2010**, *20*, 2452–2455.
[30] P. C. Lv, C. F. Zhou, J. Chen, P. G. Liu, K. R. Wang, W. J. Mao, H. Q. Li, Y. Yang, J. Xiong, H. L. Zhu, *Bioorg. Med. Chem.* **2010**, *18*, 314–319.
[31] B. E. Fink, D. Norris, H. Mastalerz, P. Chen, B. Goyal, Y. Zhao, S. Kim, G. D. Vite, F. Y. Lee, H. Zhang, S. Oppenheimer, J. S. Tokarski, T. W. Wong, A. Gavai, *Bioorg. Med. Chem. Lett.* **2011**, *21*, 781–785.
[32] J. C. Dearden, *Inter. J. Quantitative Structure-Property Relationships* **2016**, *1*, 1–44.
[33] K. Roy, S. Kar, R. N. Das, *A Primer on QSAR/QSPR Modeling: Fundamental Concepts*, Springer, New York, **2015**.
[34] K. Roy, S. Kar, R. N. Das, *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*, Academic Press, USA, **2015**.
[35] L. L. Zhu, T. J. Hou, L. R. Chen, X. J. Xu, *J. Chem. Inf. Comp. Sci.* **2001**, *41*, 1032–1040.
[36] S. Kamath, J. K. Buolamwini, *J. Med. Chem.* **2003**, *46*, 4657–4668.
[37] M. Awale, C. G. Mohan, *J. Mol. Graph. Model.* **2008**, *267*, 1169–1178.
[38] C. Y. Chen, C. Y. Chen, *J. Mol. Graph. Model.* **2010**, *291*, 21–31.
[39] M. F. Sun, S. C. Yang, K. W. Chang, T. Y. Tsai, H. Y. Chen, F. J. Tsai, J. G. Lin, C. Y. Chen, *Molecular Simulation* **2011**, *37*, 884–892.
[40] S. C. Yang, S. S. Chang, C. Y. Chen , *PLoS One* **2011**, *6*, pe28793.
[41] S. Mirzaie, M. Monajjemi, M. S. Hakhamaneshi, F. Fathi, M. Jamalan, *EXCLI J.* **2013**, *12*, 130.
[42] H. Zalloum, R. Tayyem, Y. Bustanji, M. Zihlif, M. Mohammad, T. A. Rjai, M. S. Mubarak, *J. Mol. Graph. Model.* **2015**, *61*, 61–84.
[43] P .R. Duchowicz, N. C. Comelli, E. V. Ortiz, E. A. Castro, *Curr. Drug. Saf.* **2012**, *7*, 282–288.
[44] A. Talevi, C. L. Bellera, M. D. Ianni, P. R. Duchowicz, L. E. Bruno-Blanch, E. A. Castro, *Curr. Comput. Aided Drug. Des.* **2012**, *8*, 172–181.
[45] C. W. Yap, *J. Comput. Chem.* **2011**, *32*, 1466–1474.
[46] PaDEL, http, //wwwyapcwsoftcom Accessed 2 May 2016
[47] US EPA 2016 Estimation Programs Interface Suite™ for Microsoft® Windows, v 411 United States Environmental Protection Agency, Washington, DC, USA.
[48] H. Hong, Q. Xie, W. Ge, F. Qian, F. Fang, L. Shi, Z. Su, R. Perkins, W. Tong, *J. Chem. Info. Mod.* **2008**, *48*, 1337–1344.
[49] Coral, http, //wwwinsilicoeu/coral Accessed 2 May 2016
[50] A. A. Toropov, A. P. Toropova, E. Benfenati, G. Gini, *Curr Comput Aided Drug Des.* **2013**, *9*, 226–232.
[51] Dassault Systèmes BIOVIA, Discovery Studio Modeling Environment, Release 2017, San Diego, Dassault Systèmes, 2016.
[52] A. P. Toropova, A. A. Toropov, S. E. Martyanov, E. Benfenati, G. Gini, D. Leszczynska, J. Leszczynski, *Chemometr. Intell. Lab. Syst.* **2012**, *110*, 177–181.
[53] C. Rojas, P. R. Duchowicz, P. Tripaldi, R. Pis Diez, *Chemometr. Intell. Lab. Syst.* **2015**, *140*, 126–132.
[54] C. Rojas, P. R. Duchowicz, P. Tripaldi, R. Pis Diez, *J. Chromatography A* **2015**, *1422*, 277–288.
[55] C. Rojas, P. Tripaldi, P. R. Duchowicz, *Int. J. Quant. Struct. Prop. Relationsh.* **2016**, *1*, 76–90.
[56] P. R. Duchowicz, E. A. Castro, F. M. Fernández, *MATCH Commun. Math. Comput. Chem.* **2006**, *55*, 179–719.
[57] P. R. Duchowicz, E. A. Castro, F. M. Fernández, M. González, *Chem. Phys. Lett.* **2005**, *412*, 376–380.
[58] P. R. Duchowicz, A. Talevi, L. E. Bruno-Blanch, E. A. Castro, *Bioorg. Med. Chem. Lett.* **2008**, *16*, 7944–7955.
[59] M. Goodarzi, P. R. Duchowicz, C. H. Wu, F. M. Fernández, E. A. Castro, *J. Chem. Inf. Model.* **2009**, *49*, 1475–1485.
[60] A. B. Pomilio, M. A. Giraudo, P. R. Duchowicz, E. A. Castro, *Food Chem.* **2010**, *123*, 917–927.
[61] A. Talevi, M. Goodarzi, E. V. Ortiz, P. R. Duchowicz, C. L. Bellera, G. Pesce, E. A. Castro, L. E. Bruno-Blanch, *Eur. J. Med. Chem.* **2011**, *46*, 218–228.
[62] G. Pasquale, G. P. Romanelli, J. C. Autino, J. García, E. V. Ortiz, P. R. Duchowicz, *J. Agric. Food Chem.* **2012**, *60*, 692–697.
[63] Matlab 70 Available online, http, //wwwmathworkscom (accessed on 29 July 2016)
[64] A. Golbraikh, A. Tropsha, *J. Mol. Graphics Model.* **2002**, *20*, 269–276.
[65] S. Wold, L. Eriksson, *Statistical validation of QSAR results, Chemometrics Methods in Molecular Design* (Eds.:vdW H), VCH, Weinheim, **1995**, p 309–318.
[66] K. Roy, R. N. Das, P. Ambure, R. B. Aher, *Chemometr. Intell. Lab. Syst.* **2016**, *152*, 18–33.
[67] K. Roy, P. Ambure, R. B. Aher, *Chemometr. Intell. Lab. Syst.* **2017**, *162,* 44–54.
[68] XternalValidationPlus v.1.2. Available online, *http://dtclab.webs.com/software-tools* (accessed on 15 February 2017)
[69] Bias-Variance Estimator. Available online, *http://dtclab.webs.com/software-tools* (accessed on 15 February 2017)
[70] P. Gramatica, *QSAR Comb. Sci.* **2007**, *26*, 694–701.
[71] D. Gadaleta, G. F. Mangiatordi, M. Catto, A. Carotti, O. Nicolotti, *Inter. J. Quantitative Structure-Property Relationships* **2016**, *1*, 45–63.
[72] L. Eriksson, J. Jaworska, A. P. Worth, M. T. Cronin, R. M. McDowell, P. Gramatica, *Environ. Health Perspect.* **2003**, *111*, 1361–1375.
[73] K. Roy, S. Kar, P. Ambure, *Chemometr. Intell. Lab. Syst.* **2015**, *145*, 22–29.
[74] Applicability domain 1.0. Available online, *http://dtclab.webs.com/software-tools* (accessed on 15 February 2017)
[75] R. P. Verma, C. Hansch, *Bioorg. Med. Chem.* **2005**, *13*, 4597–4621.
[76] K. Roy, *Expert Opinion on Drug Discovery* **2007**, *2*, 1567–1577.
[77] H. Singh, R. Kumar, S. Singh, K. Chaudhary, A. Gautam, G. P. Raghava, *BMC Cancer* **2016**, *16*, 1.
[78] A. G. Mercader, P. R. Duchowicz, F. M. Fernández, E. A. Castro, *J. Chem. Inf. Model.* **2011**, *51*, 1575–1581.

Submitted: February 28, 2017
Accepted: April 24, 2017