Hindawi
Journal of Chemistry
Volume 2019, Article ID 2954250, 11 pages
https://doi.org/10.1155/2019/2954250

Research Article

Computational Molecular Modeling of Pin1 Inhibition Activity of Quinazoline, Benzophenone, and Pyrimidine Derivatives

Nicolás Cabrera,[1] Jose R. Mora ,[2] and Edgar A. Marquez[3]

[1]*Universidad San Francisco de Quito, Departamento de Matemática, Diego de Robles y Vía Interoceánica, Quito 17-1200-841, Ecuador*
[2]*Universidad San Francisco de Quito, Grupo de Química Computacional y Teórica (QCT-USFQ), Departamento de Ingeniería Química, Diego de Robles y Vía Interoceánica, Quito 17-1200-841, Ecuador*
[3]*Departamento de Química y Biología, División de Ciencias Básicas, Universidad del Norte, Km 5 vía Puerto Colombia, Barranquilla, Colombia*

Correspondence should be addressed to Jose R. Mora; jrmora@usfq.edu.ec

Received 20 December 2018; Accepted 25 February 2019; Published 15 April 2019

Academic Editor: Fabio Polticelli

Pin1 (peptidyl-prolyl *cis-trans* isomerase NIMA-interacting 1) is directly involved in cancer cell-cycle regulation because it catalyses the cis-trans isomerization of prolyl amide bonds in proteins. In this sense, a modeling evaluation of the inhibition of Pin1 using quinazoline, benzophenone, and pyrimidine derivatives was performed by using multilinear, random forest, SMOreg, and IBK regression algorithms on a dataset of 51 molecules, which was divided randomly in 78% for the training and 22% for the test set. Topological descriptors were used as independent variables and the biological activity (pIC_{50}) as a dependent variable. The most robust individual model contained 9 features, and its predictive capability was statistically validated by the correlation coefficient for adjusting, 10-fold cross validation, test set, and bootstrapping with values of 0.910, 0.819, 0.841, and 0.803, respectively. In order to improve the prediction of the pIC_{50} values, the aggregation of the individual models was performed through the construction of an ensemble, and the most robust one was constructed by two individual models (LR3 and RF1) by applying the IBK algorithm, and a substantial improvement in predictive performance is reflected in the values of $R^2_{ADJ} = 0.982$, $Q^2_{CV} = 0.962$, and $Q^2_{EXT} = 0.918$. Mean square errors <0.165 and good fitting between calculated and experimental pIC_{50} values suggest a robustness on the prediction of pIC_{50}. Regarding the docking simulation, a binding affinity between the molecules and the active site for the Pin1 inhibition into the protein (3jyj) was estimated through the calculation of the binding free energy (BE), with values in the range of −5.55 to −8.00 kcal/mol, implying a stabilizing interaction molecule receptor. The ligand interaction diagrams between the drugs and amino acid in the binding site for the three most active compounds denoted a good wrapper of these organic compounds into the protein mainly by polar amino acids.

1. Introduction

Pin1 has been used as a target for treating cancer since its discovery [1] because it plays a critical role in cell-cycle regulation, it catalyses the cis-trans isomerization of prolyl amide bonds in its substrate proteins, and deregulated proteins are common human cancer cells [2]. Also, Pin1 induces apoptosis and mitotic arrest. Then, the inhibition of Pin1 presents new opportunities for the development of new anticancer treatments [3]. A potential prognostic marker in human cancers should be the overexpression of Pin1, as demonstrated for the breast [4], prostate [5], and lung [6]. Moreover, it has been reported that 38 of 60 tumours have more than 10% of Pin1 overexpression, compared with the corresponding normal controls [7].

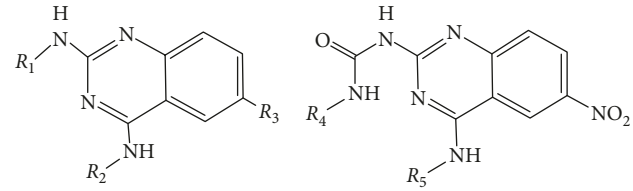
Focused on the importance of Pin1 in the cancer treatment, some inhibitors for Pin1 have been reported such as 2-{[4-(4-*tert*-butylbenzenesulfonamido)-1-*oxo*-1,4-dihydronaphthalen-2-yl]sulfanyl}acetic acid (KPT-6566) [8], all-trans retinoic acid (ATRA) [9], and inhibitors

based on aromatic compounds [10]. With respect to the last-mentioned compounds, the Bailing Xu's group had reported the synthesis of several compounds as potential Pin1 inhibitors. In their earlier efforts, in 2011, they synthesized 2,4-disubstituted quinazoline derivatives (Scheme 1). All quinazoline synthesized structures are available in Table S1 in the supplementary material. The most potential inhibitor, compound 13, with the 50% inhibitory concentration (IC_{50}) equal to 2.90 μM, has two chlorine atoms bonded in the position 3 of the aromatic ring on the substituent R_4, a carboxylic acid linked to the benzene in the position R_5, and an NO_2 group in the quinazoline nucleus. Clearly, the set of molecules depicted in Table S1 in the supplementary material have several heteroatoms present in their structures [11], which suggest a large dipole moment due to the electronegative differences between the atoms involved in the compound. These set of compounds were separated into two parts, the first containing an amine group to link the R_1, while in the second, an amide functional group is linked with the R_4 group.

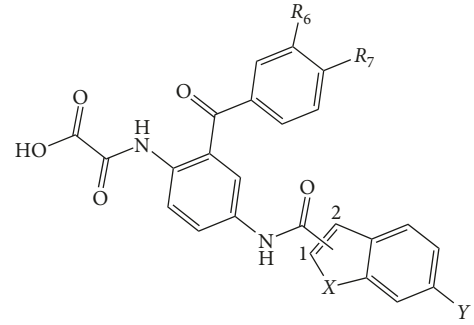
In the same context, in 2012, Liu et al. [12] prepared a series of Pin1 inhibitors with benzophenone skeleton (Scheme 2). The investigation on structure-activity relationships (SAR), varying the substituents on the position (R_6 and R_7), the binding (1 and 2), and the molecules (X and Y) was performed (complete structures are shown in Table S2 in the supplementary material). The author suggested that the Pin1 inhibition could be related to two molecular characteristics: an *oxo*-acetic group linked to the benzophenone moiety and the aromatic bicyclic ring, having different heteroatoms, linked to amide group moiety. In addition, the author suggested the methoxy group could enhance the activity substantially. The most active substrate of this set was compound 24 with an IC_{50} value of 5.99 μM, which contains a bicycle nitro-benzothiophene, which can increase substantially the polarity of the molecules in the corresponding evaluated set.

On the contrary, recently Cui et al. [13] reported the synthesis and Pin1 inhibitory activities of pyrimidine derivatives, and their core structures are shown in Scheme 3. A set of twenty-six compounds was prepared by the authors, and different aromatic substituents including the heteroatoms nitrogen, oxygen, sulphur, and some halogen were used as substituents, presenting a very interesting dataset with important variations between their structures (Table S3 in the supplementary material). The authors suggest that compounds 28, 33, 38, and 49 with IC_{50} values lower than 3 μM demonstrate potent inhibitory activities against Pin1, compound 38 being the most active with an IC_{50} value of 1.68 μM. This compound in analogy to the most active compounds of the other two sets is shown in Tables S1 and S2 in the supplementary material (13 and 24), and it also presents a bicyclic structure as substituent involving oxygen and nitrogen as heteroatoms in R_9 (4-benzoxazole). Additionally, the chlorine and nitro groups on 38 also suggest a large polarity of the molecule.

Based on the hypothesis of the existence of a relationship between the molecular structure and the biological activity, the drug design in general terms can be assisted by

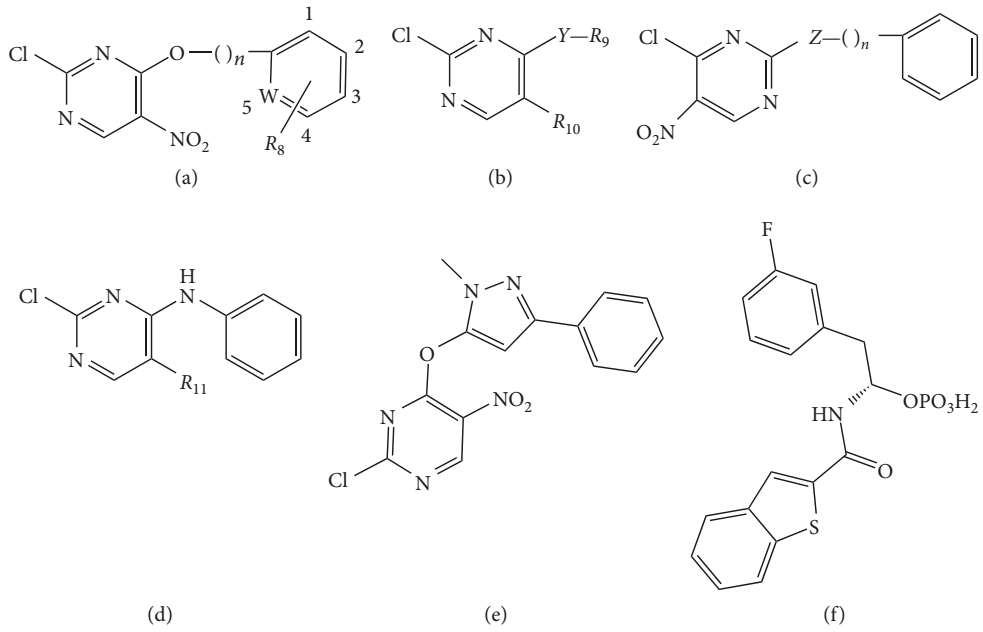


Scheme 1: Quinazoline derivatives as potential Pin1 inhibitors core structures for molecules (a) 1–10 and (b) 11–17.



Scheme 2: Series of Pin1 inhibitors with benzophenone skeleton core structure for molecules 18–26.

quantitative structure-activity relationship (QSAR) modeling, which has become a widely used tool in computer-aided drug design (CADD), fate modeling and predictive environmental risk assessment, property prediction, and toxicity of pharmaceuticals and chemicals [14]. Several molecular descriptors can be used to obtain QSAR models, and they can be achieved from quantum mechanics calculation and/or topological indexes for two- and three-dimensional structures [15–17]. In this sense, Ghalia et al. [18] reported a 3D-QSAR study of the TC_{50}, which was calculated by the Reed and Muench method and represents the concentration that inhibits 50% cellular growth compared to that untreated control, and IC_{50} (antiviral activity) by using quantum chemical descriptors, which were estimated on twenty-one molecules of novel *N*-phenyl benzamide and *N*-phenylacetophenone derivatives. Multilinear regression and non-multilinear regression models were obtained for the dataset, which was divided as the training and test set. The authors suggest that the results represent an excellent stability for high values based on correlation coefficients $R_{pIC50} = 0.91$ $R_{pTC50} = 0.96$ for the RNLM and $R_{pIC50} = 0.87$ and $R_{pTC50} = 0.95$ for MLR. In addition, a QSAR predictive analysis through an assembly of a regression model to predict the inhibition of aldose reductase for flavonoids was carried out on 55 molecular structures, including parameters of all types calculated using the software DRAGON 5.0 [19]. The predicted power of this model was measured with the following parameters $Q_{loo} = 0.934$ and l–n%–o $R_{l-30\%-o} = 0.803$ [20]. Furthermore, another study has developed QSAR models to predict the inhibitory activity of 88 organic bromodomain modulators. In this case, the descriptors were developed using QuBiLs-MIDAS and MAS



Scheme 3: Pyrimidine derivatives as potent Pin1 inhibitor core structure (a) for molecules 27–37, (b) for molecules 38–43, (c) for molecules 44–46, (d) for molecules 47 and 48, (e) Molecule 49, and (f) Molecule 52.

(Quadratic, Bilinear and *N*-Linear Maps based on *N*-tuple Spatial Metric [(Dis)-Similarity] Matrices and Atomic Weightings). One of the best models with 9 variables showed the following statistical parameters $R^2 = 0.794$, $Q^2_{\text{loo}} = 712$, $Q^2_{\text{Boot}} = 0.683$, $Q^2_{\text{EXT}} = 0.8563$, and 1 outlier [21].

On the contrary, a useful tool and widely used on drug design is the molecular docking, which is a computational procedure used to predict the binding affinity between a micromolecule (ligand) and a macromolecule (receptor), which has a particular importance in drug development [22]. In a recent study was reported a molecular docking study on the evaluation of potential anticancer agents, related with the half maximal inhibitory concentrations (IC_{50}) and their effect on microtubule assembly [23]. Docking programs have become popular to find the proper position ligands (orientation and conformation) into a protein-binding site [24]. Some scoring functions predict the ligand's biological and complementary activity; usually, docking scores are more important than having the correct position [25, 26]. One of the widely used docking programs is AutoDock Vina, which uses a sophisticated gradient optimization method in the optimization procedure [22]. By using AutoDock Vina docking algorithm, the platform Mcule's online app 1-click docking provides the highest quality purchasable molecular modeling and compound database tools, where the calculations are running on cloud machines [26].

In reference to the above description, this work is seeking a reasonable computational modeling for the inhibition of the Pin1 by six-membered aromatic derivative compounds involving the three set of molecules included in Tables S1–S3 in the supplementary material (quinazoline, benzophenone, and pyrimidine derivatives). Then, a total of fifty-two compounds with important variations into their structures, which imply a robust dataset, have been used for a molecular modeling simulation. Multilinear algebraic map descriptors were used in the modeling process, and different regression techniques were employed in the model construction as well as for the aggregation of these models through the construction of an ensemble.

2. Methodology

2.1. Dataset. A dataset was employed and separated randomly as the test set (22%) and training (78%). Compound 52 was considered as outlier based on the statistical parameters and adjusted on the training and test set; in addition, it is well-known in the literature that it failed to show cellular effects due to the poor permeability of the phosphate group [27]. Their biological activity expressed as IC_{50} was collected from the literature by the Bailing Xu research group, where 17 possess quinazoline structures (Table S1 in the supplementary material) [11], 9 are benzophenone structure (Table S2 in the supplementary material) [12], and 26 of the molecules possess pyrimidine and naphthalemic nucleus (Table S3 in the supplementary material) [13]. In Table S4 (Supplementary Materials) is shown all the IC_{50} values on μM, and a logarithmic transformation was applied (equation (1)), which was used as a dependent variable.

$$\text{pIC}_{50} = -\log\left(\frac{10^{-6}}{\text{IC}_{50}}\right). \quad (1)$$

2.2. Descriptors Calculation. All the molecules were drawn into the GaussView (Version 5.0) software, and the 3D structure was optimized with the semiempirical PM6 (parametric method 6) [28] by using the software Gaussian 16 suite [29], where the convergence criterion for the self-

consistent field (SCF) was set as default. The molecules were characterized as minimum stationary points, which were obtained by a frequency calculation on the optimized structures at 298.15 K [30]. 892 topological descriptors were calculated by using the free software QuBiLs-MIDAS and MAS, available on http://tomocomd.com/ [31], and that pool wad enlarged by the addition of 5 descriptors: *c*Log*P*, *c*Log*S*, druglikeness, total surface area, and polar surface area were calculated using the software OSIRIS DataWarrior [32]. The hydrophilicity of a drug is measured by the logarithm of the concentration of a drug in *n*-octanol over water (equation (2)); values of marked drugs are between −10 and 8. Another feature used to measure the drug effect is *c*Log*S*, which measures its distribution and absorption.

One aim for drug design is to avoid poorly soluble compounds; typical *c*log*S* values of traded drugs are greater than −8 and smaller than 2. It is calculated by applying a base 10 logarithm to solubility (*S*) in mol/liter. In addition, druglikeness is a qualitative concept for drug design and is estimated based on topological descriptors, *c*log*P*, molecular weighs, and other properties. Although positive values are recommended for traded drugs, it is not mandatory because it does not measure the biological activity or specific effect [33].

$$c\text{Log}P = \log\left(\frac{C_{\text{octanol}}}{C_{\text{water}}}\right). \quad (2)$$

Also, the total surface area and the polar surface area (psa) were also estimated. The total surface area, which considers all polar and nonpolar fragments of the molecule, as well as polar surface area (psa), which is a measure of the degree polarity of molecules, was estimated. psa is equal to the sum of surface contributions from polar fragments. "psa" of nontoxic compounds, which do not cause death or an adverse histological change, is greater than 75 $Å^2$, and compounds with psa <75 A^2 are more likely to be toxic drugs. [34].

The modeling process was done with the software Weka 3.8 [35] which offers several machine learning techniques, and the following regression techniques were used: multilinear regression (MLR), Smola and Scholkopf's algorithm for solving regression problem (SMOreg) [36], instance-based learning with parameter *k* (IBK) [37], and random forest (RF) [38, 39], which are described briefly.

2.2.1. MLR. It is a classical statistical method that calculates the "weights" or coefficients of the dependent variables of a linear expression, and the predicted value is the sum of the attributes multiplied by its weight and the Akaike criterion for model selection.

2.2.2. SMOreg. This method overcomes the sources of inefficiency and confusion caused by SMO, which maintains a single threshold value, while the SMOreg uses two criteria parameters, which significantly improve the adjusted value on regression as well as the model predictability.

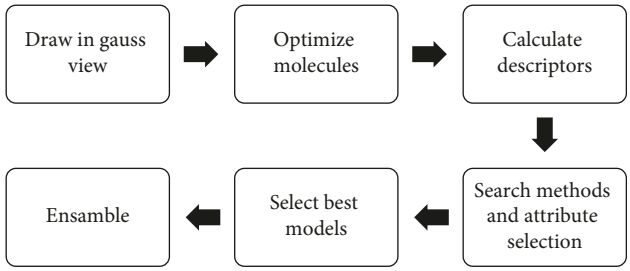
2.2.3. IBK. This method is amended in the lazy algorithms set to implement in Weka and widely used for classification and regression, which uses cross validation to select the best number of *k*, which is the same value for k nearest neighbour (KNN) approach. It measures simple distances to find the training instance closest to the test set. In case of same distance obtained in multiple instances, the first one found will be used. The parameter KNN specifies the number of the nearest neighbors to use when predicting a test instance and the outcome is determined by majority vote.

2.2.4. Random Forest. It consists of unpruned classification or regression trees by using a bootstrap sample of random feature and training data selection. The prediction values are made by the averaging or majority of votes of the ensemble. In addition, the relation with the dependent variable and the descriptors are hidden inside a "black box" and does not produce an explicit model. RF algorithm overcomes the instability of decision caused by its hierarchical nature applying subset selection and bagging techniques and reduces the bias due to class imbalance and overfitting.

2.3. Statistical Analysis. In order to determinate the robustness of a model, several statistical parameters must be calculated [40]. First, in case the coefficient of determination for adjusting (R^2) value is close to 1, the model is considered robust. Second, a model is considered suitable when the average bootstrapping (Q^2_{boot}), which provides information about the predictability of a particular model, is close to 10-fold cross validation (Q^2_{CV}). Third, the values of a(R^2) < 0.3 and $b(Q^2) < 0.05$ are accepted to validate the model. Also, the difference between the total correlation in the attributes (K_{xx}), which value is lower than 50, and the correlation in the set specified by attributes plus the dependent variable (K_{xy}) must be positive ($\Delta K = (K_{xy} - K_{xx}) > 0$). For the purpose of evaluating the internal predictability of each model, standard deviation error of prediction (SDEP), and standard deviation error of calculation (SDEC) values must be close to zero. Finally, models good fitting are corroborated by a high value of Fisher ratio (*F*) and a low-value standard deviation (*s*) [41].

Scheme 4 summarizes the process of this work step by step. To begin with, molecules are drawn in GaussView, followed by an optimization in Gaussian 16 at the PM6 level. Then, the following software is used to calculate features: QuBiLs-MIDAS, MAS, and DataWarrior. Subsequently, regression algorithms are applied in Weka 3.8 and the most robust models are selected. Finally, an ensemble by using IBK and/or RF regression techniques was assembled. The parameters to select the best assemble are coefficient of determination of adjust (R^2_{ADJ}), cross validation (Q^2_{CV}), and test set (Q^2_{EXT}), as well as the corresponding mean square errors (MAE).

2.4. Docking Analysis. The docking analysis was done by using the platform Mcule's online app 1-click docking, where the 3D structure of the receptor description file



Scheme 4: Summary of steps done in this study.

(RDF), Pin1, can be found as 3jyj on the PDB library. The protein 3jyj was selected because in the previous reports in the literature it has been identified as the most adequate receptor binding site for the evaluation and screening of possible active organic compound in pin1 inhibition [42]. The cartesian 3D coordinates were identified for the binding site as *X*: 1.1902, *Y*: 29.3651, and *Z*: 22.2862, which was established as default, and the size of the binding site was 22 Angstrom. The water molecules in Pin1 protein was removed, hydrogen was added, and incomplete residues were corrected. According to the binding free energy (BE) of the molecules, the 6 final docked conformations were ranked.

3. Result and Discussion

A total of 51 compounds were used in this study. In order to show the random distribution of the pIC_{50} values of the training and test datasets, a histogram is represented in Figure 1. It shows an adequate distribution taking values between the application domains defined by the training dataset.

A total of seventeen models were selected by applying the first condition, which implies that the models must contain a maximum of nine descriptors in order to obtain a ratio number of molecules/descriptors >5. In this sense, 5 models were obtained by applying the regression technique IBK, four with MLR, six with RF, and two with SMOreg. The adjusted and cross-validation correlation coefficient only for the training dataset is available in Supplementary Materials (Table S5). To select the most robust models, some criteria were applied as follows: for $R^2_{ADJ} > 0.78$, for 10-fold $Q^2_{CV} > 0.51$, and for the test set, a $Q^2_{EXT} > 0.64$. In this sense, in Table 1, the most robust models comply with all the conditions are shown with the corresponding correlation coefficients and MAE for adjusting, cross validation, and test set. Four of these models were obtained by using the technique MLR, and the remaining one was found by using random forest.

For the case of models obtained by MLR techniques, the equations to calculate pIC_{50} values are presented as follows:

$$\begin{aligned} pIC_{50-LR1} = {} & 8.26 + 0.0686\text{AMh} + 0.332\text{ACIch} - 4.82\text{TSch} \\ & - 0.202\text{N3h} + 0.0988\text{ACRch} - 0.0115\text{Q1ve} \\ & + 124\text{Q1ch} - 2.31\text{SICh}, \end{aligned} \tag{3}$$

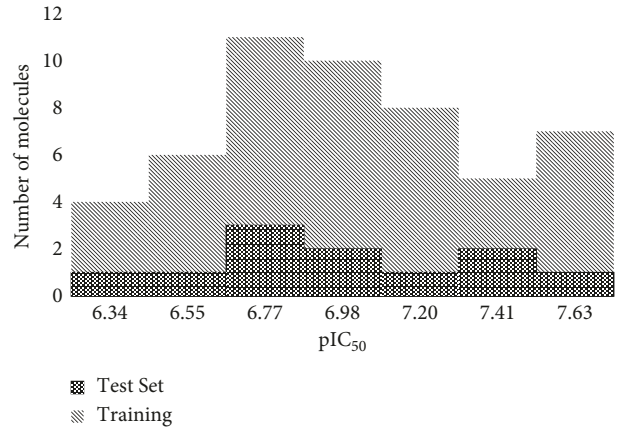


Figure 1: Distribution of experimental pIC_{50} values between training and test sets.

Table 1: The five most robust models based on regression coefficients and MAE values for adjusting, cross validation, and test set estimated by using Weka 3.8.

Model		LR 1	LR 2	LR 3	LR 4	RF1
Num. var.		8	7	9	9	8
Adjust	R^2_{ADJ}	0.781	0.843	0.889	0.910	0.978
	MAE_{ADJ}	0.157	0.131	0.108	0.099	0.089
Cross validation	Q^2_{CV}	0.684	0.773	0.755	0.819	0.514
	MAE_{CV}	0.190	0.162	0.159	0.143	0.249
Test set	Q^2_{EXT}	0.645	0.819	0.922	0.841	0.962
	MAE_{EXT}	0.211	0.129	0.098	0.156	0.189

$$\begin{aligned} pIC_{50-LR2} = {} & 11.6 - 12.4\text{TSh} - 0.412\text{GVpsa} - 266\text{TSs} \\ & - 1.45\text{SICpsas} - 0.0199\text{GVph} - 0.0226\text{Q1ve} \\ & + 0.247\text{TSms}, \end{aligned} \tag{4}$$

$$\begin{aligned} pIC_{50-LR3} = {} & 10.9 - 20.5\text{P2Ach} - 0.0248\text{Q1ve} - 0.0159\text{GVph} \\ & + 0.207\text{TSms} + 0.0601\text{Sre} - 229\text{TSs} \\ & - 0.615\text{GVpsa} + 0.771\text{Amme} - 0.435\text{RAa}, \end{aligned} \tag{5}$$

$$\begin{aligned} pIC_{50-LR4} = {} & 12 - 0.148\text{She} + 0.101\text{Sre} + 0.228\text{TSms} \\ & - 0.0237\text{Q1ve} - 0.0151\text{GVph} - 223\text{TSs} \\ & - 0.595\text{GVpsa} - 16.4\text{P2Tch} - 0.363\text{RAa}. \end{aligned} \tag{6}$$

The independent variables included in the robust models described above were named only by the invariants and the physical-chemistry (PC) properties and abbreviated in capital letters and lowercase letters, respectively. In case of two or more descriptors having the same invariant and PC property but differ at least in one characteristic, the name includes a capital letter in the middle. The description and abbreviations of the independent variables invariants used in equations (3)–(6) are presented in Table 2. The whole name

Table 2: Invariants used on the attributes calculations.

Complete description	Abbreviation
Autocorrelation with a lag value of #	AC[#]
Arithmetic mean (alfa = 1)	AM
Geometric mean	GM
Gravitational with a lag value of #	GV#
Minimum	MN
Minkowski distance	N3
Quadratic mean (alfa = 2)	P2
Range	RA
Skewness	S
Standarized information content	SIC
Percentile 25	Q1
Total sum with a lag value of #	TS[#]

for each feature is available in the supplementary materials (Table S6).

The PC parameters found in the models described above are shown in Table 3, and they have become popular in the description of many biological activities; for example, van der Waals volume (v) performs a key role in the interaction/orientation for biological activity of an organic compound. While "s" and "h" are linked with donor-acceptor properties of a molecule, "c" has a significant effect on the enzyme-substrate electrostatic interaction. Additionally, "psa" provides information about the capability of bond formation of a particular compound and has exhibited as an essential property on QSAR studies [41]. Furthermore, "r" is the refractivity calculated by Lorentz–Lorenz Formula, related not only to the London dispersive forces but also to the volume of the molecules [43]. Finally, polarizability (p) delineates the interactions between molecules, nonpolar atoms, and polar and ions molecules with dipole moments [44].

In order to validate LR models, all statistical parameters were calculated and reported in Table 4. In all the cases, the values of Q^2_{boot} are greater than 0.59; SDEP and SDEC values and the difference between Q^2_{loo} and Q^2_{boot} are all near to zero. The K values smaller than 50 suggest a noncollinearity between the selected attributes, and the ΔK values are positive except for those of the LR2 model, which was discarded for further analysis.

By taking into account the statistical parameters described in Table 4, the most robust model is LR4, which is composed by a total of nine attributes, with a Q^2_{boot} value of 0.803. It is important to highlight that all the physical-chemical parameters described above are present in the features of this model. In Figure 2 shows the graphical correlation between experimental and predicted pIC_{50} values for the training and test dataset. This result suggests a good fitting and predictability for this model. For example, for the three most potent inhibitor compounds, 13, 24, and 38, the predicted pIC_{50} values were 6.64, 6.78, and 6.33, correspondingly, while the experimental values were 6.46, 6.78, and 6.23. This result suggests an excellent prediction of the absolute values as well as the order of pIC_{50} ($24 > 13 > 38$), with a small error in the prediction of the Pin1 biological activity.

Table 3: Physical-chemistry properties found in models LR1–4 and RF1.

Physical-chemistry properties	Abr.
*Alog*P	a
Charge	c
Electronegativity	e
Hardness	h
Mass	m
Polarizability	p
Topological polar surface area	psa
Refractivity	r
Softness	s
Van der Waals volume	v

Table 4: Statistics parameters used for robustness evaluation of the MLR selected models by using MATLAB.

Model	Q^2_{boot}	K_{xx}	K_{xy}	ΔK	F
LR 1	0.594	33.1	36.2	3.12	13.9
LR 2	0.729	40.5	39.9	−0.62	24.6
LR 3	0.754	39.5	39.6	0.10	26.5
LR 4	0.803	36.6	37.1	0.55	33.7
Model	$a(R^2)$	$b(Q^2)$	SDEP	SDEC	s
LR 1	0.179	−0.399	0.233	0.185	0.211
LR 2	0.121	−0.383	0.191	0.157	0.176
LR 3	0.188	−0.474	0.179	0.133	0.153
LR 4	0.18	−0.499	0.159	0.119	0.138

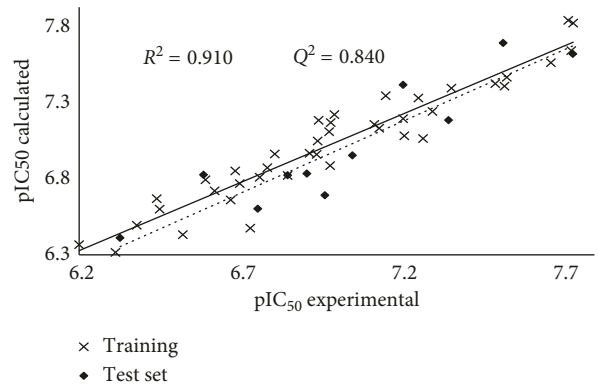


Figure 2: Graphical plot between experimental and calculated pIC_{50} by LR4.

It is important to note that individual models normally present highly sensitive to a small perturbation in the training set. Then, to tackle this problem, the construction of an ensemble modeling, which has become popular in recent years, aggregates results from different individual models [45]. IBK and RF machine learning techniques were used to construct the ensemble model where predicted values from individual models were taken as independent variables and experimental pIC_{50} as the dependent variable. Because R^2_{ADJ}, Q^2_{CV}, and Q^2_{EXT} are all over 0.89 for obtained ensembles (Table 5), the criterion of selection for the ensemble was the smaller number of variables.

Table 5: The ensemble models with the corresponding regression coefficients and MAE values.

Model		IBK	RF
Num. var.		2	4
Adjust	R^2_{ADJ}	0.982	0.994
	MAE_{ADJ}	0.041	0.024
Cross validation	Q^2_{CV}	0.962	0.960
	MAE_{CV}	0.060	0.064
Test set	Q^2_{EXT}	0.918	0.891
	MAE_{EXT}	0.164	0.146

Consequently, the ensemble model IBK was chosen as the most robust one, and its independent variables, which are individual models, are LR3 and RF1. Similar to the results found with the most robust individual model (LR4), the predicted pIC_{50} by the ensemble for the molecules 13, 24, and 38 were 6.462, 6.739, and 6.313 correspondingly, which are also in agreement with the experimental values and the order of the inhibitory activity. The graphical plot for the experimental and predicted values obtained by the ensemble is depicted in Figure 3, where an excellent fitting was obtained.

OSIRIS DataWarrior descriptors: *c*log*P*, *c*log*S*, druglikeness, and polar surface area values are shown in Figure 4 as histograms, and they were used in order to corroborate if the compounds used in this study can be considered as drugs. *c*log*P* and *c*log*S* values calculated are in the range for being declared as drugs from −9.9 to 3.6 and between −5.5 and 0, respectively. With respect to druglikeness values, most of the molecules are in the range from −17 to 2, which also support that these compounds can be considered as drugs. However, molecules 9 and 10 cannot be considered as drug trades because of present values <−17. Lastly, psa values showed a uniform distribution from 54 to 194.4, where more than 92% of molecules have values greater than 75, and consequently can be considered nontoxic drugs. The larger psa, as well as the lower *c*log*P*, is in agreement with the high polarity of these molecules due to the presence of heteroatoms in their structures. The most active compound (38) has values of *c*log*P*, *c*log*S*, and psa of −4.54, −2.74, and 100.7 A^2, respectively; consequently, this compound complies with all the necessary requirements to be considered as a drug.

3.1. Docking Simulation. The docking simulation represents a powerful tool in the drug design; thus, in the present study, all structures were docked into the binding site described for the 3jyj protein, which is reported to be related with the most common action mechanism for the Pin1 inhibitor biological activity. In Figure 5 is presented, as a histogram, the distribution of values for the binding free energy (BE), which suggest a strong-affinity protein drugs with negative values from the range −5.2 to −8.2 kcal/mol. Also, a good distribution for the test set into the training set was found, which suggests a good representation of the data by the selected training and test.

In order to gain more insight into the binding affinity on these series of compounds, three compounds from the total (13, 24, and 38) were selected and are presented in Figure 6. These molecules were selected because they are representative of the three subsets described in Tables S1–S3 in the supplementary material and are the most active compounds in each subset. The compounds 13, 24, and 38 have values of pIC_{50} of 6.5, 6.8, and 6.2, respectively; compound 38 is the most active compound in the total dataset. In contrast, values on the binding free energy of −7.5, −7.9, and −6.2 kcal/mol were found for the docked 13, 24, and 38, respectively, which suggest a good affinity interaction between the receptor and the organic compound. They have in common a bicyclic compound in their structures, where compound thirteen is a quinazoline derivative (two six-membered merged rings), twenty-four is a benzophenone derivative with a bicyclic as substituent (a five and six merged rings), and thirty-eight is a pyrimidine derivative with a bicyclic compound (five and six-membered rings). With respect to the ligand interaction diagram of these three compounds with the different present amino acids (right on Figure 6), the interaction with the residues lysine (LYS), arginine (ARG), serine (SER), leucine (LEU), aspartame (ASP), methionine (MET), glutamine (GLN), histidine (HIS), glycine (GLY), and phenylalanine (PHE) can be observed. The interaction of the three evaluated compounds with the mentioned amino acids are almost the same in each one and for the smaller one, which is the most active compound and presents a psa value of 100.7 A^2, also a good wrapper of these amino acids in the ligand is observed. All the amino acids mentioned are polar in nature, and as expected, they can have a strong interaction with the drugs considered in

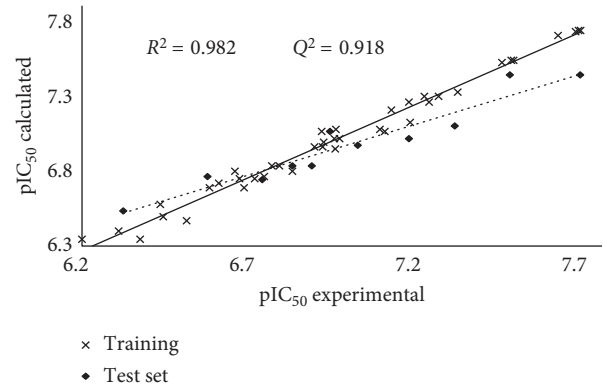


Figure 3: Graphical plot between experimental and calculated pIC_{50} by IBK ensemble.

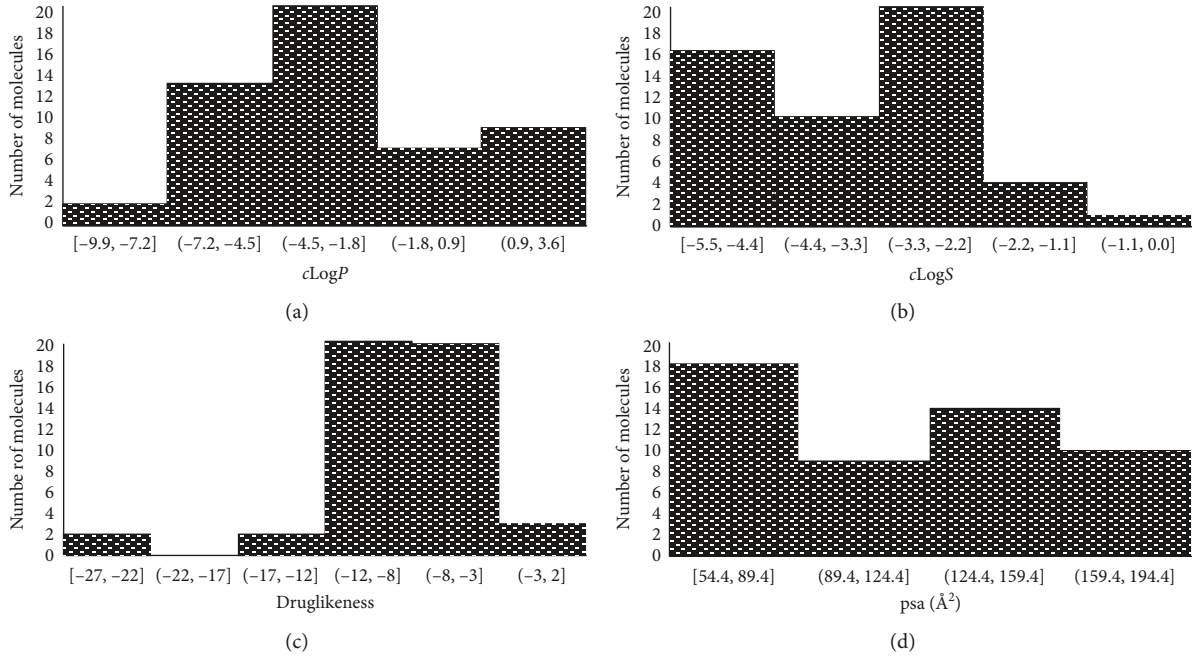


Figure 4: Distribution of (a) *c*Log*P*, (b) *c*Log*S*, (c) druglikeness, and (d) polar surface area values of the whole compounds.

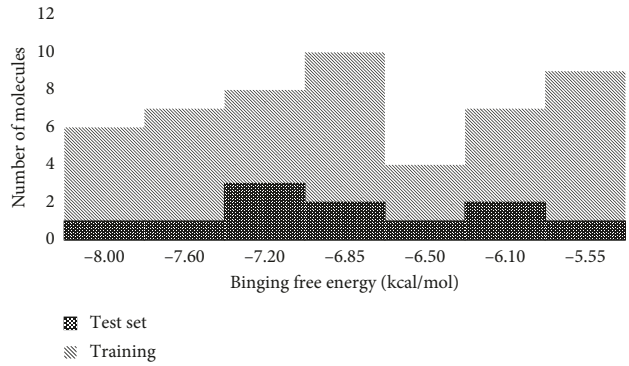


Figure 5: Distribution of binding free energy values of the whole compounds.

this study due to their polar nature because in the structures can be found some heteroatoms, which increase reasonably their polarity.

4. Conclusions

A molecular modeling simulation for the Pin1 inhibition by an organic compound containing an aromatic ring in their structure was evaluated by using QSAR approach. A total of 51 compounds, divided randomly as training (78%) and test set (22%), were used in the calculations and topological descriptor was employed for the models construction. Models were obtained by different regression techniques such as MLR, SMOreg, IBF, and RF. Five individual models were selected based on the statistical parameters, and the most robust one was constructed by using MLR approach with a total of 9 descriptors, which are weighted by the physical-chemical properties, which affect significantly the biological activity, such as *a*Log*P*, charge, electronegativity, hardness, mass, polarizability, topological polar surface area, refractivity, softness, and van der Waals volume. These properties are closely related to the biological activity of organic compounds. Regression coefficients of 0.910, 0.819, and 0.841 were obtained for adjusting, 10-fold cross validation, and test set, respectively, while the MAE values are less than 0.156. In order to improve the predictability of these models, an ensemble was constructed by using the five obtained employed IBK and RF techniques. A significant improvement was obtained in the predictability by using a multiclassifier constructed with IBk involving only two individual models (LR3 and RF1), with values of $R^2_{ADJ} = 0.982$, $Q^2_{CV} = 0.962$, and $Q^2_{EXT} = 0.918$. Consequently, this ensemble can be used for the prediction of the Pin1 inhibition activity of analogs compounds to the series used in this study. With respect to the druglikeness, *c*log*P*, *c*log*S*, and psa values, it is possible to conclude that the majority of

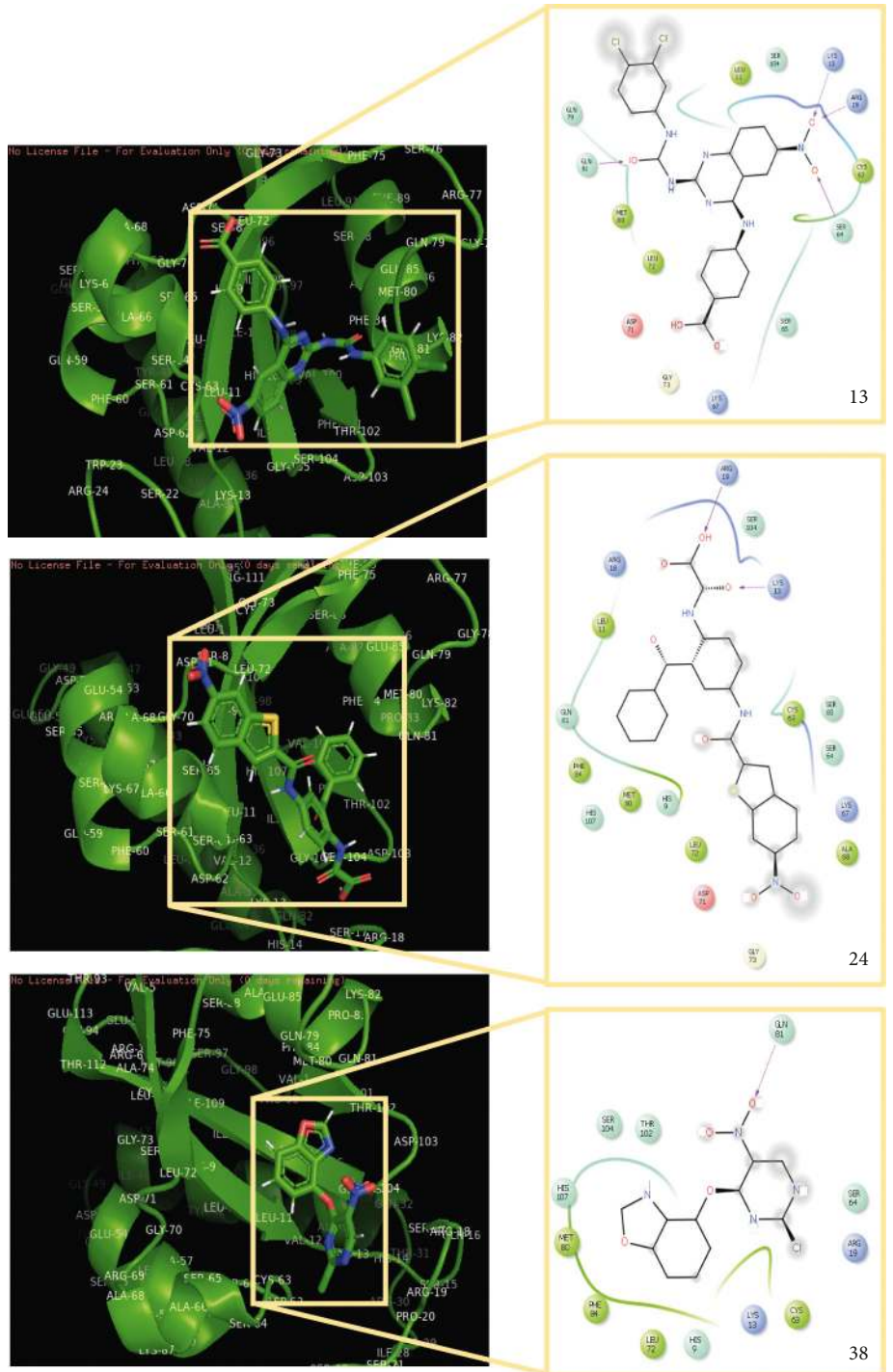


Figure 6: Docking illustration for compound 13, 24, and 38, which are most active compounds in the series.

the dataset comply with the criteria to be drugs. Finally, the docking simulation suggests a good affinity between the molecules and the Pin1 receptors with BE values in the range −5.55 to −8.00 kcal/mol.

Data Availability

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This investigation was supported by USFQ Poligrants 2018–2019. The authors have used the high-performance

computing (HPC) system available in the USFQ for the development of this project.

Supplementary Materials

Quinazoline, benzophenone, and pyrimidine derivatives as potential Pin1 inhibitor structures, dependent variables pIC_{50} of all the dataset, models for 17 models with 2–9 variables, and descriptors of the 5 most robust models. (*Supplementary Materials*)

References

[1] T. H. Lee, L. Pastorino, and K. P. Lu, "Peptidyl-prolyl*cis*-trans isomerase Pin1 in ageing, cancer and Alzheimer disease," *Expert Reviews in Molecular Medicine*, vol. 13, p. e21, 2011.

[2] E. S. Yeh and A. R. Means, "PIN1, the cell cycle and cancer," *Nature Reviews Cancer*, vol. 7, no. 5, pp. 381–388, 2007.

[3] K. P. Lu, "Phosphorylation-dependent prolyl isomerization: a novel cell cycle regulatory mechanism," *Progress in Cell Cycle Research*, vol. 4, pp. 83–96, 2000.

[4] G. M. Wulf, "Pin1 is overexpressed in breast cancer and cooperates with Ras signaling in increasing the transcriptional activity of c-Jun towards cyclin D1," *EMBO Journal*, vol. 20, no. 13, pp. 3459–3472, 2001.

[5] G. Ayala, "The prolyl isomerase Pin1 is a novel prognostic marker in human prostate cancer," *Cancer Research*, vol. 63, no. 19, pp. 6244–6251, 2003.

[6] X. Tan, F. Zhou, J. Wan et al., "Pin1 expression contributes to lung cancer prognosis and carcinogenesis," *Cancer Biology & Therapy*, vol. 9, no. 2, pp. 111–119, 2010.

[7] L. Bao, A. Kimzey, G. Sauter, J. M. Sowadski, K. P. Lu, and D.-G. Wang, "Prevalent overexpression of prolyl isomerase Pin1 in human cancers," *American Journal of Pathology*, vol. 164, no. 5, pp. 1727–1737, 2004.

[8] E. Campaner, A. Rustighi, A. Zannini et al., "A covalent Pin1 inhibitor selectively targets cancer cells by a dual mechanism of action," *Nature Communications*, vol. 8, article 15772, 2017.

[9] D. Yang, W. Luo, J. Wang et al., "A novel controlled release formulation of the Pin1 inhibitor ATRA to improve liver cancer therapy by simultaneously blocking multiple cancer pathways," *Journal of Controlled Release*, vol. 269, pp. 405–422, 2018.

[10] T. Uchida, M. Takamiya, M. Takahashi et al., "Pin1 and Par14 peptidyl prolyl isomerase inhibitors block cell proliferation," *Chemistry & Biology*, vol. 10, no. 1, pp. 15–24, 2003.

[11] L. Zhu, J. Jin, C. Liu et al., "Synthesis and biological evaluation of novel quinazoline-derived human Pin1 inhibitors," *Bioorganic & Medicinal Chemistry*, vol. 19, no. 9, pp. 2797–2807, 2011.

[12] C. Liu, J. Jin, L. Chen et al., "Synthesis and biological evaluation of novel human Pin1 inhibitors with benzophenone skeleton," *Bioorganic & Medicinal Chemistry*, vol. 20, no. 9, pp. 2992–2999, 2012.

[13] G. Cui, J. Jin, H. Chen, R. Cao, X. Chen, and B. Xu, "Synthesis and biological evaluation of pyrimidine derivatives as novel human Pin1 inhibitors," *Bioorganic & Medicinal Chemistry*, vol. 26, no. 8, pp. 2186–2197, 2018.

[14] K. Roy, S. Kar, and P. Ambure, "On a simple approach for determining applicability domain of QSAR models," *Chemometrics and Intelligent Laboratory Systems*, vol. 145, pp. 22–29, 2015.

[15] A. Gupta, V. Kumar, and P. Aparoy, "Role of topological, electronic, geometrical, constitutional and quantum chemical based descriptors in QSAR: mPGES-1 as a case study," *Current Topics in Medicinal Chemistry*, vol. 18, no. 13, pp. 1075–1090, 2018.

[16] H. Kubinyi, "Book review: molecular descriptors in QSAR/QSPR by Mati Karelson," *Angewandte Chemie International Edition*, vol. 40, no. 6, pp. 1136-1137, 2001.

[17] R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors*, John Wiley & Sons, Hoboken, NJ, USA, 2011.

[18] H. E. Ghalia, M. Bourass, and A. Ouammou, "3D-QSAR models to predict the antiviral activities of a series of novel *N*-phenylbenzamideand *N*-phenylacetophenone compounds based on density functional theory using statistical methods," *Moroccan Journal of Chemistry*, vol. 4, pp. 204–214, 2016.

[19] C. Tebby, E. Mombelli, P. Pandard, and A. R. R. Péry, "Exploring an ecotoxicity database with the OECD (Q)SAR Toolbox and DRAGON descriptors in order to prioritise testing on algae, daphnids, and fish," *Science of the Total Environment*, vol. 409, no. 18, pp. 3334–3343, 2011.

[20] A. G. Mercader, P. R. Duchowicz, F. M. Fernández et al., "QSAR prediction of inhibition of aldose reductase for flavonoids," *Bioorganic & Medicinal Chemistry*, vol. 16, no. 15, pp. 7470–7476, 2008.

[21] C. R. García-Jacas, K. Martinez-Mayorga, Y. Marrero-Ponce, and J. L. Medina-Franco, "Conformation-dependent QSAR approach for the prediction of inhibitory activity of bromodomain modulators," *SAR and QSAR in Environmental Research*, vol. 28, no. 1, pp. 41–58, 2017.

[22] O. Trott and A. J. Olson, "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading," *Journal of Computational Chemistry*, vol. 31, no. 2, pp. 455–461, 2009.

[23] H.-L. Qin, Z.-P. Shang, I. Jantan et al., "Molecular docking studies and biological evaluation of chalcone based pyrazolines as tyrosinase inhibitors and potential anticancer agents," *RSC Advances*, vol. 5, no. 57, pp. 46330–46338, 2015.

[24] P. A. Greenidge, C. Kramer, J.-C. Mozziconacci, and W. Sherman, "Improving docking results via reranking of ensembles of ligand poses in multiple X-ray protein conformations with MM-GBSA," *Journal of Chemical Information and Modeling*, vol. 54, no. 10, pp. 2697–2717, 2014.

[25] I. Wallach, *Improving posing and ranking of molecular docking*, Ph.D. Thesis, ACS, Washington, DC, USA, 2013.

[26] R. Kiss, M. Sandor, and F. A. Szalai, ": a public web service for drug discovery," *Journal of Cheminformatics*, vol. 4, no. 1, p. P17, 2012, http://Mcule.com.

[27] C. Guo, X. Hou, L. Dong et al., "Structure-based design of novel human Pin1 inhibitors (I)," *Bioorganic & Medicinal Chemistry Letters*, vol. 19, no. 19, pp. 5613–5616, 2009.

[28] J. J. P. Stewart, "Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements," *Journal of Molecular Modeling*, vol. 13, no. 12, pp. 1173–1213, 2007.

[29] M. Frisch, *Gaussian 09, Revision B.01, Gaussian 16 Revis, A03*, Gaussian Inc., Wallingford, CT, USA, 2016.

[30] L. L. Julio, J. R. Mora, A. Maldonado, and G. Chuchani, "Gas-phase elimination kinetics of selected aliphatic α,β-unsaturated aldehydes catalyzed by hydrogen chloride," *Journal of Physical Organic Chemistry*, vol. 28, no. 4, pp. 261–265, 2015.

[31] C. R. García-Jacas, Y. Marrero-Ponce, L. Acevedo-Martínez, S. J. Barigye, J. R. Valdés-Martiní, and E. Contreras-Torres,

"QuBiLS-MIDAS: a parallel free-software for molecular descriptors computation based on multilinear algebraic maps," *Journal of Computational Chemistry*, vol. 35, no. 18, pp. 1395–1409, 2014.

[32] M. Flores-Sumoza, J. Alcázar, E. Márquez, J. Mora, J. Lezama, and E. Puello, "Classical QSAR and docking simulation of 4-pyridone derivatives for their antimalarial activity," *Molecules*, vol. 23, no. 12, p. 3166, 2018.

[33] T. Sander, J. Freyss, M. von Korff, and C. Rufener, "DataWarrior: an open-source program for chemistry aware data visualization and analysis," *Journal of Chemical Information and Modeling*, vol. 55, no. 2, pp. 460–473, 2015.

[34] M. J. Waring, J. Arrowsmith, A. R. Leach et al., "An analysis of the attrition of drug candidates from four major pharmaceutical companies," *Nature Reviews Drug Discovery*, vol. 14, no. 7, pp. 475–486, 2015.

[35] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, p. 10, 2009.

[36] C. Li and L. Jiang, "Using locally weighted learning to improve SMOreg for regression," in *Lecture Notes in Computer Science*, pp. 375–384, PRICAI 2006: Trends in Artificial Intelligence, Guilin, China, 2006.

[37] S. Vijayasarathy and J. Chatterjee, "Comparison of MLR, isotonic regression and KNN based QSAR models for the prediction of inhibitory activity of HDAC6 inhibitors," *International Journal of Life Sciences Biotechnology and Pharma Research*, vol. 4, no. 2, pp. 127–131, 2015.

[38] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and QSAR modeling," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.

[39] K. Lee, M. Lee, and D. Kim, "Utilizing random Forest QSAR models with optimized parameters for target identification and its application to target-fishing server," *BMC Bioinformatics*, vol. 18, no. 16, p. 567, 2017.

[40] L. D. S. Veras, M. Arakawa, K. Funatsu, and Y. Takahata, "2D and 3D QSAR studies of the receptor binding affinity of progestins," *Journal of the Brazilian Chemical Society*, vol. 21, no. 5, pp. 872–881, 2010.

[41] J. R. Mora, E. A. Márquez, and L. Calle, "Computational molecular modelling of *N*-cinnamoyl and hydroxycinnamoyl amides as potential *α*-glucosidase inhibitors," *Medicinal Chemistry Research*, vol. 27, no. 9, pp. 2214–2223, 2018.

[42] H. Zhao, G. Cui, J. Jin, X. Chen, and B. Xu, "Synthesis and Pin1 inhibitory activity of thiazole derivatives," *Bioorganic & Medicinal Chemistry*, vol. 24, no. 22, pp. 5911–5920, 2016.

[43] M. A. F. Afzal and J. Hachmann, "Benchmarking DFT approaches for the calculation of polarizability inputs for refractive index predictions in organic polymers," *Physical Chemistry*, vol. 21, no. 8, pp. 4452–4460, 2019.

[44] C. Hansch, W. E. Steinmetz, A. J. Leo, S. B. Mekapati, A. Kurup, and D. Hoekman, "On the role of polarizability in chemical–biological interactions," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 1, pp. 120–125, 2003.

[45] Q. Zhang, J. M. Hughes-Oliver, and R. T. Ng, "A model-based ensembling approach for developing QSARs," *Journal of Chemical Information and Modeling*, vol. 49, no. 8, pp. 1857–1865, 2009.