Development of QSAR Models to Identify *Mycobacterium tuberculosis* enoyl-ACP-reductase Enzyme Inhibitors

Aureo André Karolczak[a]; Luis Fernando S. M. Timmers[b]; Rafael Andrade Caceres[c]

[a]Graduate Program in Biosciences at UFCSPA - Universidade Federal de Ciências da Saúde de Porto Alegre (Federal University of Health Sciences of Porto Alegre). Porto Alegre/RS, CEP 90050-170, Brasil. Email contact: aureoa@ufcspa.edu.br

[b]Graduate Program in Biotechnology at UNIVATES - Universidade do Vale do Taquari (University of Vale do Taquari). Lageado/RS CEP 95914-014, Brasil. Email contact: luis.timmers@univates.br

[c] Graduate Program in Biosciences at UFCSPA - Universidade Federal de Ciências da Saúde de Porto Alegre (Federal University of Health Sciences of Porto Alegre). Porto Alegre/RS, CEP 90050-170, Brasil. Email contact: rafaelca@ufcspa.edu.br

ABSTRACT

Tuberculosis is a matter of global concern due to its prevalence in developing countries and the ability of mycobacterium to present resistance to existing therapeutic regimens. In this context, the present project proposes the QSAR (Quantitative Structure-Activity Relationships) modeling, as a way of identifying and evaluating, in silico, the estimated inhibitory activity of candidate molecules for the molecular improvement stages and/or in vitro assays, reducing research time and financial costs. For this purpose, the SAR (Structure-Activity Relationships) study conducted by He, Alian and Montellano (2007) was used on a series of arylamides, tested as inhibitors of the enzyme enoyl-ACP-reductase (InhA) of *Mycobacterium tuberculosis*. The Hansh-Fujita (classical) and CoMFA (Comparative Molecular Field Analysis) QSAR models were developed. The classic QSAR model obtained the best statistical result, using Multiple Linear Regression (MLR), with internal validation with correlation factor R^2 = 0.9012 and predictive quality, according to the Stone-Geisser indicator Q^2 = 0.8612. In the external validation, a correlation factor R^2 = 0.9298 and a Q^2= 0.720 was obtained, indicating a highly predictive mathematical model. The CoMFA Model managed to obtain a Q^2 = 0.6520 in the internal validation, which allowed the energy fields around the molecules used to be estimated, this is essential information to foster molecular improvement. A library of small molecules was built, with analogs to those used in the SAR study, which were subjected to classic QSAR function, resulting in a group of ten molecules with high estimated biological activity. The molecular docking results suggest that the ten analogs identified by the classical QSAR model presented favorable estimated free energy of binding. The conclusion points to the QSAR methodology as an efficient and effective tool in the search and identification of promising drug-like molecules.

Keywords: Tuberculosis. QSAR. CoMFA. InhA. Docking.

Introduction

According to the Global Tuberculosis Report 2020, published by the World Health Organization (WHO), the disease caused by *Mycobacterium tuberculosis* (MTB) is among the most lethal today, being classified as one of the 10 leading causes of death in the world and the leading cause of death having a single agent of infection. Another relevant data is the information that approximately a quarter of the world population is infected with MTB. The study conducted by the WHO points out a correlation between poverty and the incidence of the disease, suggesting that the higher the per capita income, the lower the incidence of tuberculosis (TB), making clear the social aspect that involves this disease [1].

A perverse feature of this social difference is the very low funding and production of new therapeutic products for the known neglected diseases, which includes TB [2]. The context of research underfunding becomes more serious when it is projected that multidrug-resistant tuberculosis (MDR-TB) will be responsible for 25% of all deaths from drug-resistant pathogens [3]. In countries such as Uzbekistan and Turkmenistan, the percentage of MDR-TB in relation to other strains reaches up to 30% of cases [4].

1.1. QSAR models

A useful approach to accelerate studies of new drugs and reduce research costs is the use of QSAR (Quantitative Structure-Activity Relationships) modeling to build mathematical models that relate the chemical structure and biological activity of a series of analogous compounds [5]. These models are generated from previous SAR (Structure-Activity Relationships) studies. The QSAR models selected in this survey were the following:

1.1.1. Hansh-Fujita or Classic Model

It seeks to relate biological activity with the so-called molecular descriptors (MD), through multiple linear regression (MLR) [6]. In this case, molecular descriptors (independent variables) seek to predict biological activity (dependent variable).

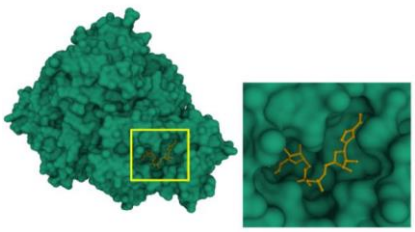
1.1.2. CoMFA

The CoMFA (Comparative Molecular Field Analysis) model, according to Cramer and his team [7], uses the representation of ligand molecules by their steric and electrostatic energetic fields of interaction, sampled at the intersections of a three-dimensional network (GRID). It also uses a "field fit" technique that allows for optimal mutual alignment with in a series, minimizing RMS field differences between molecules, and it performs partial least squares (PLS) data analysis. The model allows the graphical representation of the results, as graphs of coefficients and three-dimensional energy contours [7].

1.2. Pharmacological Target

This research aims to build mathematical models that can point out possible inhibitors of the enzyme enoyl-ACP-reductase (InhA), from *Mycobacterium tuberculosis*, which will be our pharmacological target.This enzyme (Picture 1) acts in the FAS II cycle (Fatty Acid Synthase) in which repetitive reactions of condensation, keto-reduction, dehydration and enoyl-reduction occur for the synthesis of fatty acids (FA) [8]. Studies indicate that the enzymatic action occurs around the aminoacid residues Tyr158, Lys165, Ser123 and Phe149, in the catalytic site [9].

Picture 1. The InhA enzyme showing, in detail, the cofactor NADH (yellow) in the catalytic site



Deposited in the Protein Data Bank (PDB) under code 6SQ5.

It is desirable to target an enzyme for which analogs are not found among mammals and that can avoid resistance mechanisms, such as those reported in studies with isoniazid and ethionamide [10]. In this case, InhA presents itself as a relevant research object [10-13].

2 Methodology

2.1 Flowchart of QSAR models

The classic QSAR and CoMFA models followed the same flowchart of actions (Picture 2).

Picture 2. Flowchart for building the classical QSAR and CoMFA mathematical models. In yellow, the QSAR function with predictive capability obtained after internal and external validation can be seen.

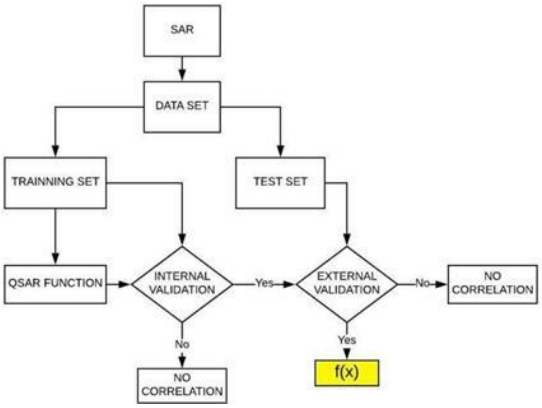
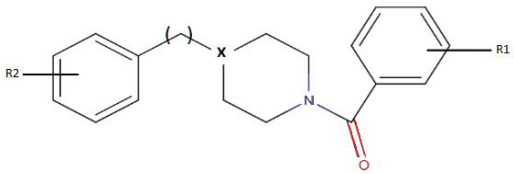


Image by the author.

2.2. SAR study

In 2007, He, Alian and Montellano [16] carried out an inhibition study, targeting the InhA of mycobacteria. For this purpose, a series of arylamides was used (Pic.5). The bioassay reported 24 active molecules (Appendix II) which are described in the PubChem public database [17] under AID code 301095.

Picture 5. Basic Structure of Arylamide.

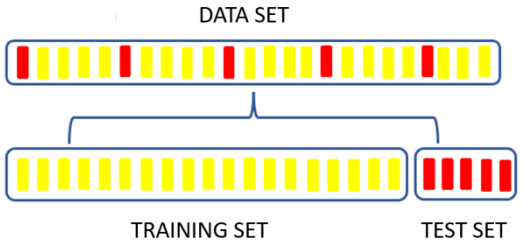


Source:He, Alian and Montellano (modified picture) [16]

2.3. Data Classification

According to Martin *et al.* [18] every set of active molecules, used for QSAR models, corresponds to the so-called data set and they should be divided into training set and test set in percentages of 80% for the first one and 20% for the second one (Picture 6) [19]. In this study, we use a proposal adapted from Damre *et al*. [20] that takes into account structural diversity and biological activity values.

Picture 6. Division of molecules into training set and test set, according to Martin *et al* [18] and Leonard and Roy [19].



Source: Image by the author.

The adapted proposal from Damre and his team [20] involves the following 2 steps:

a) rank the compounds in decreasing order of measured biological activity values (IC_{50}, EC_{50}, MIC or Ki) in logarithm form. In this study, the $LogIC_{50}$ index was chosen, which is the logarithm of half of the substance concentration needed to reach the maximum inhibitory activity.

b) transfer the first compound to the test set, separate the next 4 or 5 compounds for the training set [18], the subsequent compound is destined for the Test set. The process is repeated until all the compounds that make up the data set are divided.

2.4. Classic Model

2.4.1. Molecular Descriptors

For the classic QSAR, 18 molecular descriptors were evaluated. Most of them were calculated using the program MarvinView 17.24.0 [21], using SDF format files obtained from the Pubchem database [17]. The water solubility was obtained using the ALOGPS 2.1 web application [22] using smile codes from the PubChem database. The DMs used are specified in Appendix I.

2.4.2 QSAR Function

With the training set defined, all molecular descriptors calculated and the biological activity ($LogIC_{50}$) obtained from the SAR study, the data were processed using the BuildQSAR program [23] to obtain the QSAR equation. The method chosen was multiple linear regression (MLR), it is an efficient method when the objective is to use few molecular descriptors and a reduced number of compounds (≤ 30) [20]. The process involved evaluating the relationship of each molecular descriptor individually with the biological activity, then performing the same process using DM pairs, and finally using 3 molecular descriptors to correlate with the $LogIC_{50}$. Combinations with good results were obtained; however, it was decided to choose the descriptors that reached the highest correlation without generating outliers in the internal validation.

2.4.3 Internal Validation

The internal validation of the QSAR function used residual analysis to evaluate the fit of data and interpret the RLM model obtained with the training set. In a well-fitted model, that is, in which the model errors are independent and have a Normal distribution with the same variance, we expect the following:

- The points on the plot of adjusted values x standardized residuals are mostly between -2 and 2 on the vertical axis and do not have any pattern of behavior,
- that the QQ-plot points are close to the line,
- that the density of the residues be symmetrical in a bell-like shape around zero, ranging from -2 to 2,
- There are no points with high Cook Distance, which could be disturbing the model's fit.

The correlation graph to be generated must have a dotted line that represents the equality between the observed and predicted values, so that the closer a point is to it, the closer the observed and predicted values are. If the Prediction Interval intersects this line, then we have that the model predicted the observed value within the margin of error. In addition, we have the solid line that represents the line that best fits the observed (y) and predicted (x) results. In order to complete the statistical analysis, the respective standard deviations, Pearson's correlation factor R, coefficient of determination R^2 and the scatter plot [24] must also be calculated. However, as correlation does not mean causality, the results must be subjected to the Stone-Geisser indicator Q^2 to determine predictive quality [25].

2.4.4 External Validation

The test set composed of 5 molecules [20] was submitted to QSAR function where, the biological activity was estimated. The calculated results were compared with the biological activity observed in the AID 301095 bioassay, to obtain the respective standard deviations and the Pearson correlation factors R, coefficient of determination R^2 and the scatter plot [24], in addition to the predictive quality via the indicator Stone-Geisser Q^2 [25].

2.4.5 Virtual Screening

The process of searching for molecules with the same nucleus as the arylamides used the Tanimoto coefficient as a selection tool, available in the public database PubChem [17]. Once the analogous compounds were found, the respective standard SDF files were downloaded for the second step of the process, which is to calculate the molecular descriptors of the compounds, required by the QSAR function. For this, the

program MarvinView17.24.0 [21] is used again. Subsequently, the MD results are submitted to the QSAR function to calculate the biological activity, ordering the results in descending order of $LogIC_{50}$. At the end of the step, the 10 molecules with the highest estimated biological activity passed the molecular docking phase.

2.4.6. Molecular Docking

For this step, molecule 2 of the training set (CID 447767) was used due to its high in vitro biological activity observed in the AID 301095 bioassay. Next, a structure was searched in the Protein Data Bank (PDB) [26], a structure that has molecule 2 bound to InhA. Once the desired complex was found, the obtained PDB format file was edited to make it compatible with the PyRx Python Prescription 0.8 program [27]. Then, all 10 molecules selected in the virtual screening were submitted to Molecular docking via PyRx program with AutoDock Vina.The results are classified according to the binding energy EFEB (Estimated Free Energy of Binding) and the molecule with the lowest binding energy is compared to the complex deposited in the PDB database used as a model. The 3D image of the new complex obtained via molecular docking is generated with the application PyMol (The PyMOL Molecular Graphics System, Version1.8, 2015) and its molecular interactions at the catalytic site are obtained with the program BIOVIA Discovery Studio Visualizer [28]. Contributions to the global binding energy to the catalytic site were calculated with the Blind Docking Server, available at:<http://bio-hpc.eu/software/blind-docking-server/> [29].

2.5 CoMFA Model

For the CoMFA modeling, the tool developed by Prof. Rhino Ragno [30]. As in the classic QSAR model, the same distribution of molecules from the SAR study was used, divided into training set and test set. The internal and external validations were performed by Professor Ragno's tool, as well as the generation of 3D graphics of the energy fields. In the graph to be generated, the dotted line represents the equality between the observed and predicted values, so that the closer a point is to it, the closer the observed and predicted values are. Besides that, we have the solid line that represents the line that best fits the points of the graph, establishing a relationship between the values of the experiment (y) and the predicted values (x).Statistical analysis was performed to obtain the respective standard deviations, Pearson correlation factor R, coefficient of determination R^2 and the scatter plot [24], in addition to the predictive quality via the Stone-Geisser Q^2 indicator [25].

2.5.1 Alignment

For the alignment of the molecules, the RDkit method was used, having as reference the most active compound of the training set (CID 447767), in the conformation with the lowest global energy.

2.5.2 Model Generation

The formatting to search for conformations and model generation followed the settings below:

- Probe carbon atom Sp3 charge +1,
- Grid with an extension of 8 Å,
- Dielectric constant of 18,
- Cutoff of 20 kcal/mol,
- External test: true.

3 Results and discussion

3.1 Data division for both models

The molecules classified as:2, 3, 4, 5, 7, 8, 9, 10, 12, 13, 14, 15, 17, 18, 19, 20, 22, 23 and 24, in order of decreasing biological inhibitory activity. For the test set, the molecules chosen were classified as:1, 6, 11, 16 and 21. The identification of each molecule and the inhibitory activity, described by the AID 301095 bioassay, can be checked in Appendix III.

3.2. Classic Model

3.2.1. Calculation of Molecular Descriptors

Altogether, 18 molecular descriptors (Appendix I) were calculated and evaluated regarding their correlation with the inhibitory activity of the molecules that compose the training set. The following DMs that obtained the maximum correlation without producing outliers were selected to compose the QSAR function:

- HA – number of hydrogen acceptors,
- RB – number of routable links,
- VWALLS V – Van der Walls volume.

Appendix IV contains the Table with the calculated values for the 3 chosen molecular descriptors related to the 19 molecules that make up the training set.

3.2.2 QSAR Function

With the values of the 3 DM (Appendix IV) plus the inhibitory activity ($LogIC_{50}$) of the molecules that make up the training set, the predictive function was obtained, which can be seen in the equation (Equation 1) that represents the QSAR function for the series of molecules evaluated by He, Alian and Montellano [16].

Equation 1. QSAR function.

$LogIC_{50}$ = 0.1727 HA + 0.3886 RB - 0.0204 VWALLS V + 5.4255

3.2.3 Internal Validation

After obtaining the biological activity calculated using the QSAR function (Table 3), the results were statistically evaluated regarding the adequacy of the data used and it was proven that the parameters of the 3 independent variables are statistically meaningful at a significance of 1% ($p < 0.01$), so there is a significant relationship between molecular descriptors and biological activity. Table 4 shows the result of the good fit of the model. Appendix V presents an analysis of the residuals of the model obtained, which are the difference between the observed value and the predicted value in the training sample. The results observed in the graphs in appendix V show a behavior that evidences the good fit of the model in relation to the training set data.

The graph in Appendix VI presents the observed values of the $logIC_{50}$ and the values predicted by the model for each observation in the training sample, along with a Prediction Interval of 95%. The equation of the line and its R^2, a measure of the quality of its fit, are presented in the graph, together with the Pearson Correlation Coefficient (R) between the observed and predicted values, and its p-value (p).We observed a high R^2 (0.9012) and a high and statistically significant correlation between the observed and predicted values in the training sample, so that the model predicts the training data well, a fact also evidenced by the adjusted line being close to the y = x line. Finally, we have that, for the training set, Q^2 was equal to 0.8612, which is further evidence of the model's prediction quality.

3.2.4 External Validation

Table 6 presents the observed and calculated results for the biological activity of the test set. Graph 1 shows the dispersion of results, with Pearson's correlation being R = 0.9643 and the coefficient of determination R^2 being 0.9298. Despite the presence of a compound with very high biological activity (compound 1) with a tolerance band that did not cross the dotted line. The predictive quality of the classic model was so robust that the Stone-Geisser Q^2 indicator was calculated at 0.720.

3.2.5 Virtual Screening

The search for compounds with the same piperazine nucleus (Picture 2) resulted in a set of 135 analogs. The molecular descriptors HA, RB and VVWALLS of the screened molecules were calculated using the program MarvinView 17.24.0 [21]. The calculated DM values were applied in the QSAR function (Equation 1). And the 10 best calculated results were ordered in descending values as shown in table 7.

Table 3 Observed and calculated values of the biological activity of the training set molecules.

Molecule[a]	Log $IC_{50Obs.}$[b]	$LogIC_{50Calc.}$[c]	SD[d]
2	-0.699	-0.712	0.0092
3	-0.398	-0.513	0.0813
4	-0.004	0.413	0.2949
5	0.017	-0.049	0.0467
7	0.276	0.150	0.0891
8	0.310	0.497	0.1322
9	0.487	0.760	0.1930
10	0.713	0.813	0.0707
12	0.797	0.911	0.0806
13	0.828	0.821	0.0049
14	0.869	0.810	0.0417
15	0.889	0.867	0.0156
17	0.989	1.177	0.1329
18	1.142	1.179	0.0262
19	1.150	1.230	0.0566
20	1.189	0.921	0.1895
22	1.246	0.816	0.3041
23	1.498	1.409	0.0629
24	1.590	1.388	0.1428

[a] molecules ordered in decreasing order of molecular biological activity ($LogIC_{50}$).
[b]inhibitory biological activity observed by He, Alian and Montellano [16].
[c]inhibitory biological activity calculated using the BuildQSAR program[23]
[d] calculated standard deviation.
Source: Prepared by the author.

Table 4 Model Parameters.

Parameter	Estimate	Standard Error	Value -t	p
Intercepto	5.426	0.428	12.689	<0.001
HA	0.173	0.048	3.629	0.002
RB	0.389	0.079	4.901	<0.001
VWallsV	-0.020	0.002	-11.567	<0.001

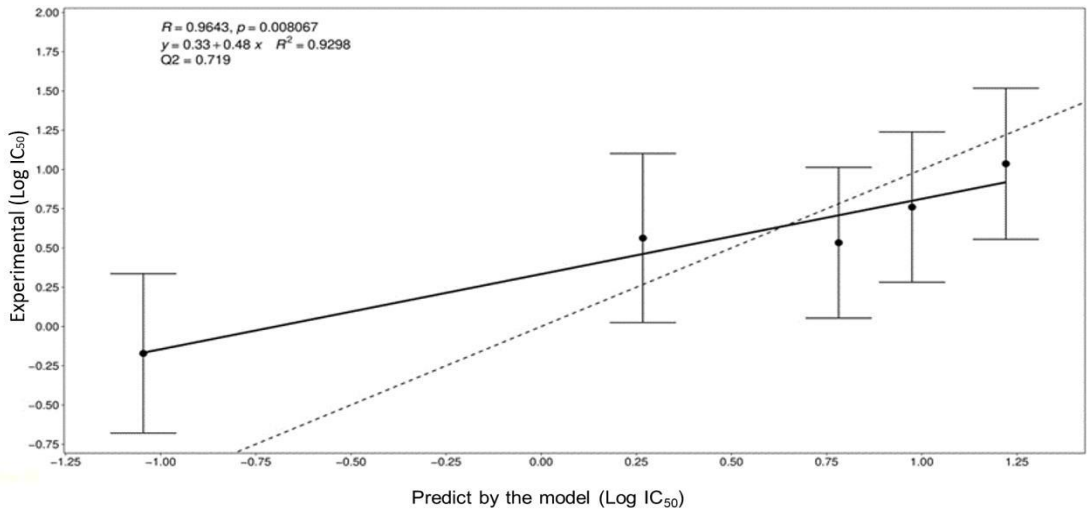
Source:Prepared by the author.

Table 5 Values calculated for the molecular descriptors of the test set compounds

Molecule[a]	HA[b]	RB[c]	VWALSS V[d]
1	2	2	329.56
6	5	2	318.93
11	2	2	294.98
16	2	2	283.85
21	2	2	270.31

[a]classification by decreasing inhibitory activity of the molecule.
[b] number of hydrogen acceptors in the molecule.
[c] number of rotatable bonds in the molecule.
[d] Van der Walls volume to cubic angstrom (Å3).
Source: Prepared by the author.

Graph 1. Dispersion of calculated and observed results of biological activity ($LogIC_{50}$) of the Classic test set



Source: Prepared by the author.

Table 6 Observed and calculated values of the biological activity of the test set molecules.

Molecule[a]	$LogIC_{50Obs}$.[b]	$LogIC_{50Calc}$.[c]	SD[d]
1	-1.046	-0.175	0.616
6	0.267	0.560	0.207
11	0.782	0.531	0.178
16	0.975	0.758	0.153
21	1.221	1.034	0.132

[a]molecules ordered in decreasing order of biological activity ($LogIC_{50}$).
[b] inhibitory biological activity observed by He, Alian and Montellano [16].
[c] inhibitory biological activity calculated using the QSAR function.
[d] calculated standard deviation.
Source: Prepared by the author.

3.2.6. Molecular Docking

The 10 compounds selected by virtual screening were submitted to molecular docking, using the enzyme InhA deposited in the PDB database under code 1P44 as a ligand. The results of this process are shown in Table 8. It is noted that the energy of binding of the compound CID 4005008 to the InhA enzyme was lower than that obtained for the model ligand (CID 447767) of the complex deposited in the PDB database (1P44). The EFEB of the sorted compound was -13.3 kcal/mol, while the compound deposited in the PDB database had an EFEB of -11.1 kcal/mol. Which indicates a higher theoretical affinity of the CID 4005008 molecule with InhA. The other molecules tested also showed low binding energies, which supports the indication of in vitro tests of the entire set selected by the QSAR model. In image 7, we can see the compound CID 4005008 complexed to InhA at its catalytic site. The analysis of the energetic interactions of the CID 4005008 molecule with the most favorable EFEB, confirmed its location and interaction with the catalytic aminoacid residues, especially Phe149 and Lys165 as shown in Picture 9. We also noticed that the repulsion energy, as well as rotational impediments were minimal when compared with the total energy of affinity as it can be seen in picture 10, corroborating the potential of the compound CID 4005008 for eventual in vitro test.

Table 7 Compounds located by virtual search and $LogIC_{50}$ values calculated with the QSAR function obtained in this study.

Molecule[a]	CID[b]	$LogIC_{50}$[c]
1	4005008	-1.345
2	70154773	-0.771
3	10179644	-0.770
4	16764927	-0.707
5	5274981	-0.681
6	110399797	-0.570
7	78897937	-0.523
8	118725820	-0.517
9	118725821	-0.516
10	9517736	-0.456

[a]molecules ranked in descending order of biological activity.
[b] identifier code of the molecule deposited in the PubChem database [17].
[c] biological activity measured in logarithm of a micromolar substance required to reach 50% of the maximum inhibitory concentration (IC_{50}).
Source: Prepared by the author.

Table 8. Estimated binding free energy obtained during molecular docking.

Molecule (CID)	Energy of binding [a]
4005008	-13.3
5274981	-12.0
70154773	-11.9
10179644	-11.7
16764927	-11.6
9517736	-11.5
118725820	-11.5
118725821	-11.0
7887937	-10.6
110399797	-9.4

[a]Estimated energy of binding measured in kcal/mol.

3.3. CoMFA

3.3.1 Internal Validation

The training set was submitted to CoMFA modeling with the configurations mentioned in the methodology, obtaining as estimated biological activity the values shown in table 9. Appendix VII presents the graph of the experiment and the values predicted by the model for each observation of the training sample. We observed a high and statistically significant R^2 and R correlation, at a statistical significance of 1% ($p<0.01$), between the observed and predicted values by cross-validation in the training sample. The obtained Q^2 was equal to 0.6517, which is evidence of the good quality of the model.

3.3.2 External Validation

With the test set we found, in this case, there is a high correlation and R^2 between the values observed in the SAR experiment, and predicted by the model. However, Q^2 is not as high as in the training set, being equal to 0.5728, which represents only a reasonable predictive quality of the model in the test sample. Also, the fitted line deviates a lot from the y = x line, mainly due to the point that had an experimental value of 7.046 (Table 10). The predicted and calculated values can be seen in table 10 and the scatter plot can be seen in graph 2.

Table 9. Experimental and calculated results of the CoMFA training set.

Molecule	Exp[b]	Calc[c]
116119	4.502	4.680
258240	4.410	4.616
447767	6.699	6.720
667749	4.850	4.797
679013	4.754	5.153
681767	5.011	4.823
702506	5.172	4.825
739897	4.858	4.859
749123	5.513	5.473
788974	6.004	5.697
807901	5.203	5.172
814705	5.111	4.985
1378917	5.131	5.183
1415427	5.287	5.357
2838469	4.811	5.103
9003925	6.398	6.433
9378917	5.690	5.334
17312855	5.724	5.824
25093354	5.983	6.069

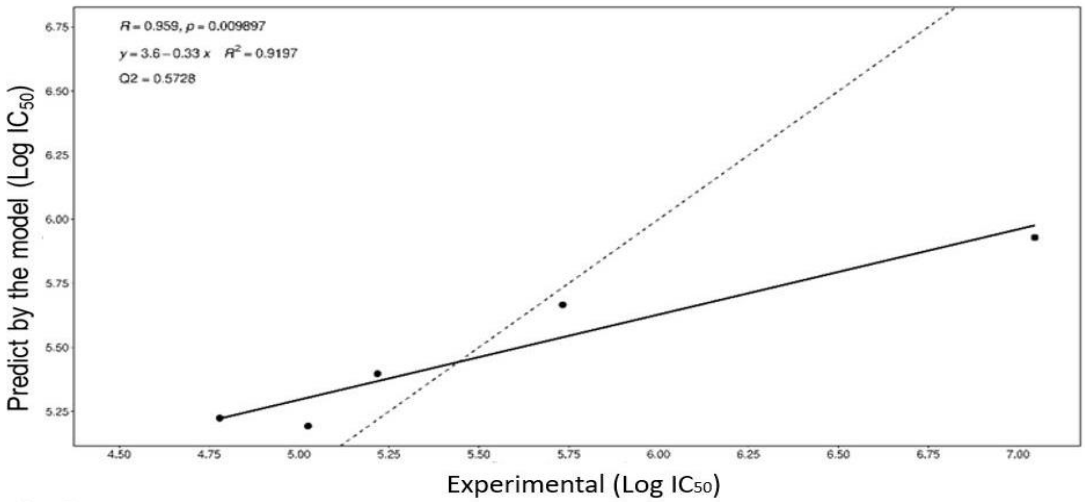
[a] molecules identified by the CID code (PubChem).
[b] experimental biological activity converted by function "$p = -\log(x)$(logistics)", the range of these logistic values spans over 2 log values [30].
[c] biological activity calculated by the CoMFA model.

Table 10. Experimental and calculated results of the CoMFA test set.

Molecule (CID)[a]	**Exp[b]**	**Calc[c]**
749125	4.779	5.224
1012371	5.733	5.666
1070143	5.218	5.397
4006554	5.025	5.193
8366407	7.046	5.929

[a]molecules identified by the CID code (PubChem).
[b] experimental biological activity converted by function"$p = -\log(x)$ (logistics)", the range of these logistic values spans over 2 log values.
[c] biological activity calculated by the CoMFA model.

Graph 2. Scatter of calculated and observed biological activity results ($LogIC_{50}$) from the CoMFA test set.



Source: Prepared by the author.

3.3.4 Force Fields Used

It is suggested that with the set of SAR molecules plus the modeling configurations used, the interaction energy based on stereochemistry was more relevant to correlate with biological activity than electrostatic energy, or both energies together. This conclusion can be understood by looking at panel 1, where we can assess the predictive quality of the CoMFA model using only stereochemical interaction energy (STE) values, only electrostatic interaction energy values (ELE) or both energies (STE-ELE).Table 1 led us to consider only the STE energy to model the training set and the test set, and all CoMFA results cited in this research were obtained using STE interaction energy data.

Panel 1.

	STE[a]	ELE[b]	STE-ELE[c]
Q^2d	0.6517	0.0720	0.3250

[a] stereochemical energy.
[b] electrostatic energy.
[c] both energies.
[d] Stone-GeisserQ^2 indicator

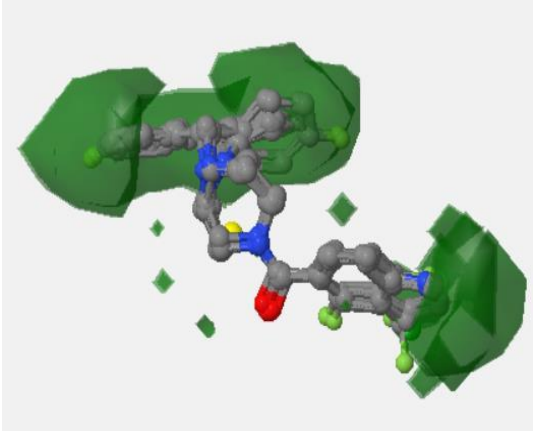
3.3.5 Generation of STE 3D fields

With the CoMFA modeling completed, three-dimensional graphs were generated indicating the location where the STE interaction energy is more favorable and less favorable. This is relevant information for molecular improvement. In Picture 6 we can identify where the STE interactions are more favorable, in Picture 7 we can see where the STE interactions are less favorable. According to Ankur Vaidya *et al* [31] bulky STE energy fields (picture 6) favor biological activity, so the study of bulky hydrophobic substitution in these regions could theoretically increase the desired inhibitory activity.

3.3.6 Discussion

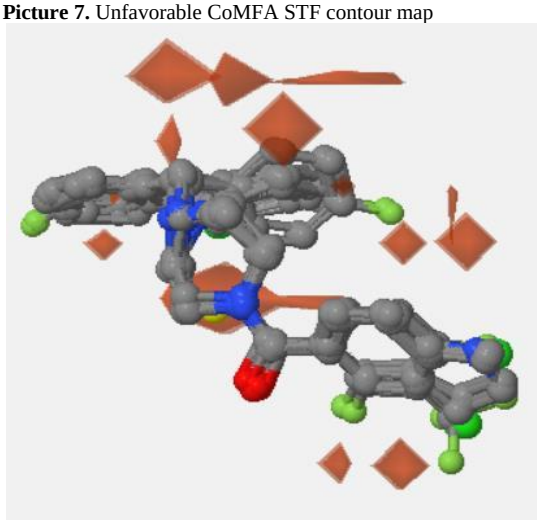
We note that the classic QSAR model obtained a more favorable statistical analysis than the CoMFA model. This result may be related to the fact that classical modeling, using RLM, was able to better reproduce an extreme result, referring to molecule 1 of the test set, than CoMFA using PLS. However, this research decided not to perform an unwanted statistical treatment of outliers, as a way to more clearly evaluate the 2 models in question. The discussion of whether or not to remove outliers is valid, as both CoMFA and classical QSAR would obtain exceptional results without the “extreme” point, however, it is desirable to seek a model capable of reasonably reproducing or predicting reality data, even with those values that stray from the average. Although an outlier eventually creates a negative bias in an analysis, this same outlier may be the answer to the problem under study.

Picture 6. Favorable CoMFA STE contour map



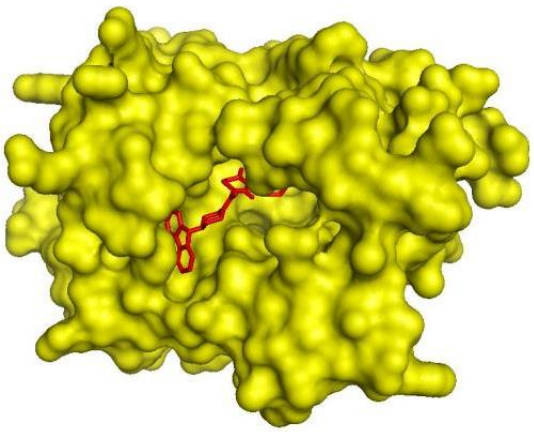
■ Favorable stereochemical interactions.

Picture 7. Unfavorable CoMFA STF contour map



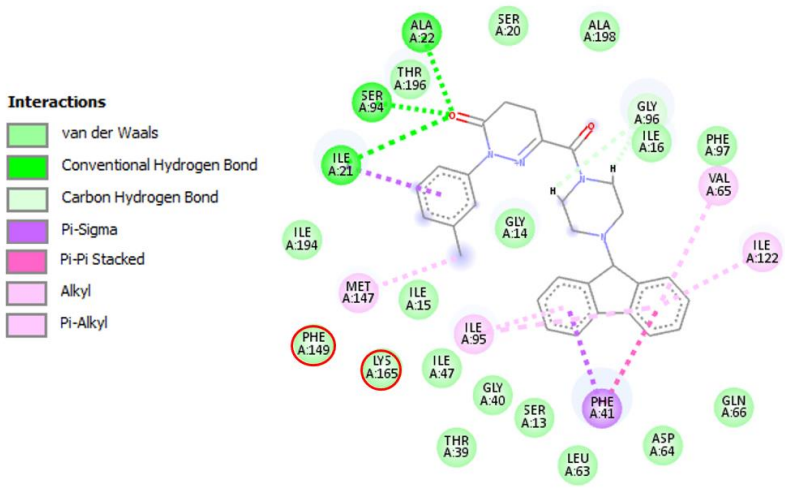
■ Unfavorable stereochemical interactions.

Picture 8 Compound CID 4005008 (red) complexed to InhA (yellow).



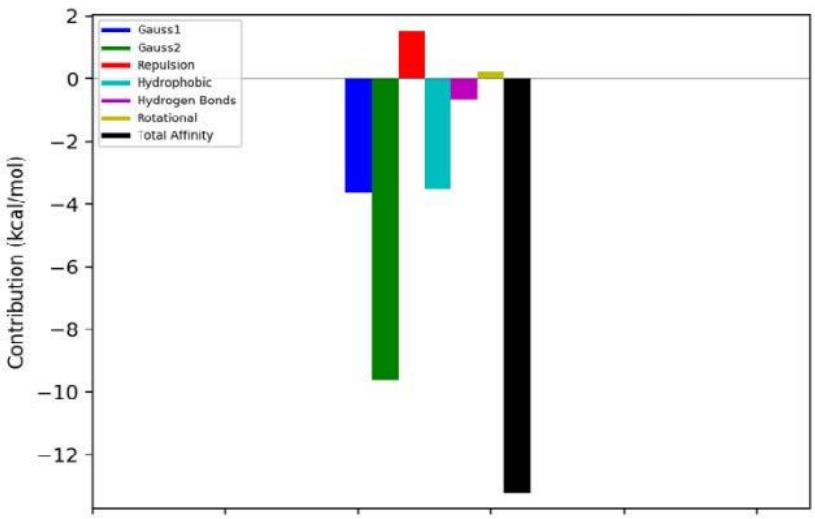
Source:Image by the author.

Picture 9. Energetic interactions of the CID 4005008 molecule at the catalytic site of InhA. Highlighting with a red circle, the aminoacid residues Phe149 and Lys165, important for biological activity.



Source: Prepared by the author with BIOVIA [28].

Picture 10. Energetic contributions to the overall binding energy between the InhA receptor and the CID 4005008 ligand.



Source: obtained from Achilles blind docking server [29].

4 Conclusion

In this research, we observed the ability of classical QSAR modeling to generate a function that estimates the inhibitory activity of a series of arylamides analogs, in terms of their molecular descriptors. It was possible to obtain a function that, both in the internal and external validation, generated excellent results. Such results allowed the use of the classic QSAR function to perform a screening of molecules structurally similar to those used by He and collaborators [16]. Of the 135 molecules obtained in virtual screening and submitted to the QSAR function, we used the 10 compounds with the highest calculated biological activity, for a molecular docking step that not only pointed to the appropriate classical QSAR modeling to locate molecules with estimated low energy of binding, but also found compounds theoretically with more affinity for the InhA enzyme than the ligand/bound complex originally deposited in the PDB database used as a model.

The application of the CoMFA mathematical model obtained good results using the data of the STE energetic interactions, allowing the generation of the three-dimensional graph of the field of interaction of the stereochemical energy in the surroundings of the molecules of the SAR study. This information made it possible to propose bulky hydrophobic substitutions that could theoretically increase the biological activity of drug-like compounds. The use of the 2 QSAR models provided a broader picture of the factors that favor or disfavor the biological activity at the catalytic site of InhA, in addition to suggesting new candidate molecules for in vitro assays. The research found that using the 2 mathematical models gives us more comprehensive answers for

the search for new compounds of interest, which may be a proposal not to limit future research in this area to just 1 predictive model.

DECLARATION OF CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

BIBLIOGRAPHY

[1] World Health Organization (WHO). Global tuberculosis report 2020: executive summary. Geneva; 2020.available in: <https://www.who.int/publications-detail-redirect/978924001313>.

[2] Pedrique B, Strub-Wourgaft N, Some C, Olliaro P, Trouiller P, Ford N, *et al*. The drug and vaccine landscape for neglected diseases (2000–11): a systematic assessment. The Lancet Global Health. 2013 Dec;1(6):e371–9. available in: <http://www.thelancet .com/journals/langlo/article/PIIS2214-109X(13)70078-0/abstract>. Acesso em: 10 de maio de 2021. doi: 10.1016/S2214-109X(13)70078-0

[3] Brigden G, Castro J, Ditiu L, Gray G, Hanna D, Low M, *et al*. Tuberculosis and antimicrobial resistance – new models of research and development needed. Bulletin of the World Health Organization [Internet]. 2017 May 1;95(5):315–5. doi:10.2471/blt.17.194837.

[4] Golan DE. Princípios de farmacologia: a base fisiopatológica da farmacoterapia. Rio De Janeiro: Guanabara Koogan; 2009.

[5] Ferreira MMC, Montanari CA, Gaudio AC. Seleção de variáveis em QSAR. Química Nova. 2002 May; 25(3):439–48.

[6] Narasimhan B. Qsarbyhanschanalysis. J PharmBioSci. 2014;2(4):2-6

[7] Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. Journal of the American Chemical Society. 1988 Aug;110(18):5959–67. doi: 10.1021/ja00226a005.

[8] Nelson DL, Cox MM. Princípios de bioquímica de Lehninger.5th ed. Porto Alegre: Artmed; 2011.

[9] Chollet A, Maveyraud L, Lherbet C, Bernardes-Génisson V. An overview on crystal structures of InhA protein: Apo-form, in complex with its natural ligands and inhibitors. European Journal of Medicinal Chemistry. 2018 Feb;146:318–43.

[10] Morlock GP, Metchock B, Sikes D, Crawford JT, Cooksey RC. ethA,inhA , and katG Loci of Ethionamide-Resistant Clinical Mycobacterium tuberculosis Isolates. Antimicrobial Agents and Chemotherapy. 2003 Dec;47(12):3799–805.doi: 10.1128/AAC.47.12.3799-3805.2003.

[11] Freundlich Joel S, Wang F, Vilchèze C, Gulten G, Langley R, Schiehser Guy A, *et al*. Triclosan Derivatives: Towards Potent Inhibitors of Drug-Sensitive and Drug-Resistant Mycobacterium tuberculosis. ChemMedChem. 2009 Feb 13;4(2):241–8.doi: 10.1002/cmdc.200800261

[12] Khan R, Zeb A, Roy N, Magar RT, Kim HJ, Lee KW, *et al*. Biochemical and Structural Basis of Triclosan Resistance in a Novel Enoyl-Acyl Carrier Protein Reductase. Antimicrobial Agents and Chemotherapy [Internet]. 2018 Aug 1;62(8). doi:10.1128/AAC.00648-1

[13] Khade AB, Boshoff HIM, Arora K, Vandana KE, Verma R, Shenoy GG. Design, synthesis, evaluation, and molecular dynamic simulation of triclosan mimic diphenyl ether derivatives as antitubercular and antibacterial agents. Structural Chemistry. 2020 Jan 6;31(3):983–98. doi:10.1007/s11224-019-0147

[14] Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, *et al*. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research. 2019 Jan 8;47(Database issue):D607–13.

[15] Vilchèze C, Jacobs JR. WR. Resistance to Isoniazid and Ethionamide in Mycobacterium tuberculosis: Genes, Mutations, and Causalities. Microbiology Spectrum. 2014 Aug 22;2(4).

[16] He X, Alian A, Ortiz de Montellano PR. Inhibition of the Mycobacterium tuberculosis enoyl acyl carrier protein reductaseInhA by arylamides. Bioorganic & Medicinal Chemistry. 2007 Nov;15(21):6649–58. doi:10.1016/j.bmc.2007.08.013

[17] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, *et al*. PubChem Substance and Compound databases. Nucleic Acids Research [Internet]. 2015 Sep 22;44(D1):D1202–13.

[18] Martin TM, Harten P, Young DM, Muratov EN, Golbraikh A, Zhu H, *et al*. Does Rational Selection of Training and Test Sets Improve the Outcome of QSAR Modeling? Journal of Chemical Information and Modeling. 2012 Oct 3;52(10):2570–8. doi: 10.1021/ci300338w

[19] Leonard JT, Roy K. On Selection of Training and Test Sets for the Development of Predictive QSAR models. QSAR & Combinatorial Science. 2006 Mar;25(3):235–51. doi: 10.1002/qsar.200510161

[20] Damre MV, Gangwal RP, Dhoke GV, Lalit M, Sharma D, Khandelwal K, *et al*. 3D-QSAR and molecular docking studies of amino-pyrimidine derivatives as PknB inhibitors. Journal of the Taiwan Institute of Chemical Engineers. 2014 Mar;45(2):354–64. doi: 10.1016/j.jtice.2013.05.016

[21] Chemaxon LTD., 2017. MarvinView 17.24.0para Windows.Hungria: ChemAxon. Conjunto de dados. Ambiente operacional.

[22] Tetko IV, Gasteiger J, Todeschini R, Mauri A, Livingstone D, Ertl P, *et al*. Virtual Computational Chemistry Laboratory – Design and Description. Journal of Computer-Aided Molecular Design. 2005 Jun;19(6):453–63.

[23] Oliveira DB, Gaudio AC. BuildQSAR: A New Computer Program for QSAR Analysis. Quantitative Structure-Activity Relationships. 2000 Dec;19(6):599–601. doi:10.1002/1521-3838(200012)19:6<599::AID-QSAR599>3.0.CO;2-B

[24] Neter J, Kutner M, Wasserman W, Nachtsheim C. Applied linear statistical models. 4th ed. Boston: McGraw-Hill/Irwin, 1996. 1415p.

[25] Chin WW. How to Write Up and Report PLS Analyses. Handbook of Partial Least Squares [Internet]. 2009 Nov 16;655–90. doi: 10.1007/978-3-540-32827-8_29

[26] Berman HM. The Protein Data Bank. Nucleic Acids Research. 2000 Jan 1;28(1):235–42. doi: 10.1093/nar/28.1.235

[27] Dallakyan S, Olson AJ. Small-molecule library screening by docking with PyRx. Methods Mol Biol. 2015;1263:243-250. doi:10.1007/978-1-4939-2269-7_19

[28] Biovia. Dassault Systèmes. Discovery Studio Visualizer, V21.1.0.20298, San Diego: 2021.

[29] Sánchez-Linares I, Pérez-Sánchez H, Cecilia JM, García JM. High-Throughput parallel blind Virtual Screening using BINDSURF. BMC Bioinformatics. 2012;13(Suppl 14):S13.doi: 10.1186/1471-2105-13-S14-S13

[30] Ragno R. www.3d-qsar.com: a web portal that brings 3-D QSAR to all electronic devices—the Py-CoMFA web application as tool to build models from pre-aligned datasets. Journal of Computer-Aided Molecular Design. 2019 Sep;33(9):855–64.

[31] Vaidya A, Jain AK, Prashantha Kumar BR, Sastry GN, Kashaw SK, Agrawal RK. CoMFA, CoMSIA, kNN MFA and docking studies of 1,2,4-oxadiazole derivatives as potent caspase-3 activators. Arabian Journal of Chemistry. 2017 May;10:S3936–46.

[32] Todeschini R, Consonni V. Handbook of molecular descriptors. Weinheim; New York: Wiley-Vch; 2000

Appendix I

Table 11 Molecular descriptors evaluated in this study.

Molecular Descriptors	Description*
MW	Average molecular weight calculated from standard atomic weights.
MW Exact	Monoisotopic mass calculated from the weights of the most abundant natural isotopes of the elements.
Atomcount	Number of atoms in the molecule
Complexity	It takes into account both the variety of connectivity types and atom types of a molecular graph without H^+.
HA	Hydrogen acceptors
RB	Rotatable bonds in the molecule
XLogP3	Optimized classification with scheme of 87 types of atoms/groups as well as two factors responsible for internal H bonds and aminoacids
Log P HLB	Octanol/water partition coefficient. Hydrophilic-lipophilic balance, hydrophilic or lipophilic degree of the molecule.
MPLZ	Molecular polarizability, the relative tendency of an electron cloud (a charge distribution) of a molecule to be distorted by an external electric field.
Weiner Polarizability	The number of 3 bodies away from the molecule.
Dreiding Energy	Energy of the 3D structure (conformation) of the molecule using the Dreiding force field.
MMFF94	Energy of the 3D structure (conformation) of the molecule using the MMFF94 force field.
Van der Walls Volume PSA	Van der Waals conformer volume.
MSA 3D	The polar surface area (PSA) is formed by polarized atoms of the molecule.3D molecular surface area.
RF	Refractivity, related to the volume of molecules and London dispersive forces
Log S	Solubility in water.

Source:MarvinView 17.24.0 program [21,32] and Todeschini and Consonni [32]

Appendix II

Table 12 Series of arylamides tested in vitro by He *et al* [16].

Molecule*	CID**	IC_{50} pm***
1	8366407	0.09
2	447767	0.2
3	9003925	0.4
4	788974	0.99
5	25093354	1.04
6	1012371	1.85
7	17312855	1.89
8	9378917	2.04
9	749123	3.07
10	1415427	5.16
11	1070143	6.05
12	807901	6.26
13	702506	6.73
14	1378917	7.39
15	814705	7.74
16	4006554	9.43
17	681767	9.74
18	739897	13.87
19	667749	14.11
20	2838469	15.47
21	749125	16.64
22	679013	17.62
23	116119	31.5
24	258240	38.86

*molecules ranked in descending order of inhibitory activity. **Identifier code of the molecule deposited in the PubChem database [17]. ***biological activity measured in micromolar of substance necessary to reach 50% of the maximum inhibitory concentration (IC_{50}).
Source:He; Alliance; Montellano, 2007 [16].

Appendix III

Table 13 Composition of the Test set and Training set.

Molecule*	Test set (CID)	Training set (CID)
1	8366407	
2		447767
3		9003925
4		788974
5		25093354
6	1012371	
7		17312855
8		9378917
9		749123
10		1415427
11	1070143	
12		807901
13		702506
14		1378917
15		814705
16	4006554	
17		681767
18		739897
19		667749
20		2838469
21	749125	
22		679013
23		116119
24		258240

* Molecules that make up the SAR study, ordered in descending order of biological activity. The compounds are divided into test set and training set columns, and identified by the CID code used in the PubChem database. Source: Prepared by the author.

Appendix IV

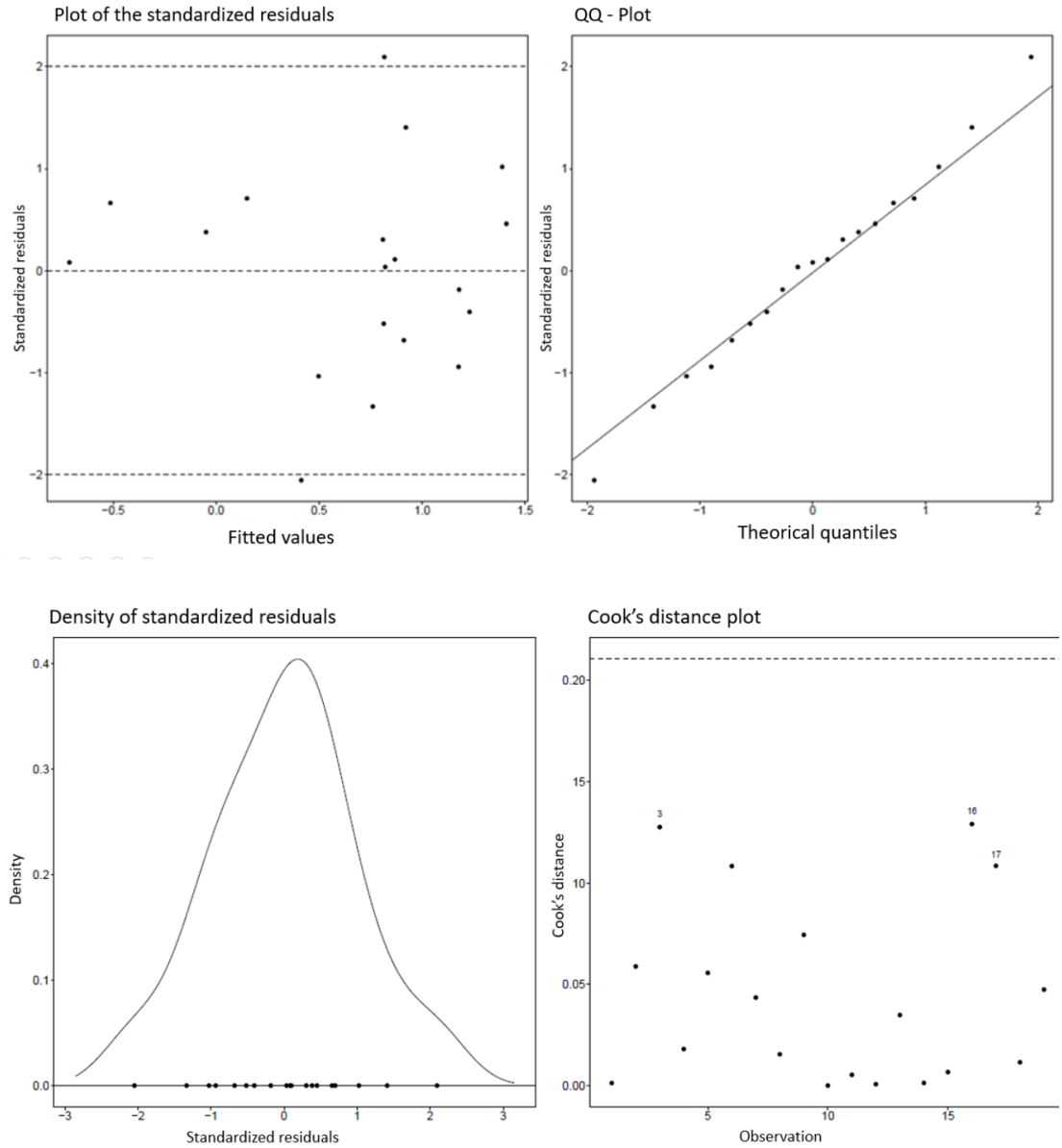
Table 14 Values calculated for the training set molecular descriptors.

Molecule[a]	HA[b]	RB[c]	VWALSS V[d]
2	2	2	356.067
3	2	2	346.311
4	2	2	300.899
5	4	4	378.613
7	4	4	368.857
8	4	4	351.813
9	2	2	283.854
10	1	3	291.845
12	5	2	301.864
13	2	2	280.899
14	1	3	292.032
15	1	3	289.196
17	3	2	271.895
18	3	2	271.782
19	2	3	279.875
20	4	2	292.928
22	2	2	281.106
23	2	3	271.115
24	2	2	253.055

[a] classification by decreasing inhibitory molecular activity.
[b] number of hydrogen acceptors in the molecule.
[c] number of rotatable bonds in the molecule.
[d] Van der Walls volume to cubic angstrom (Å3).
Source:Prepared by the author.

Appendix V

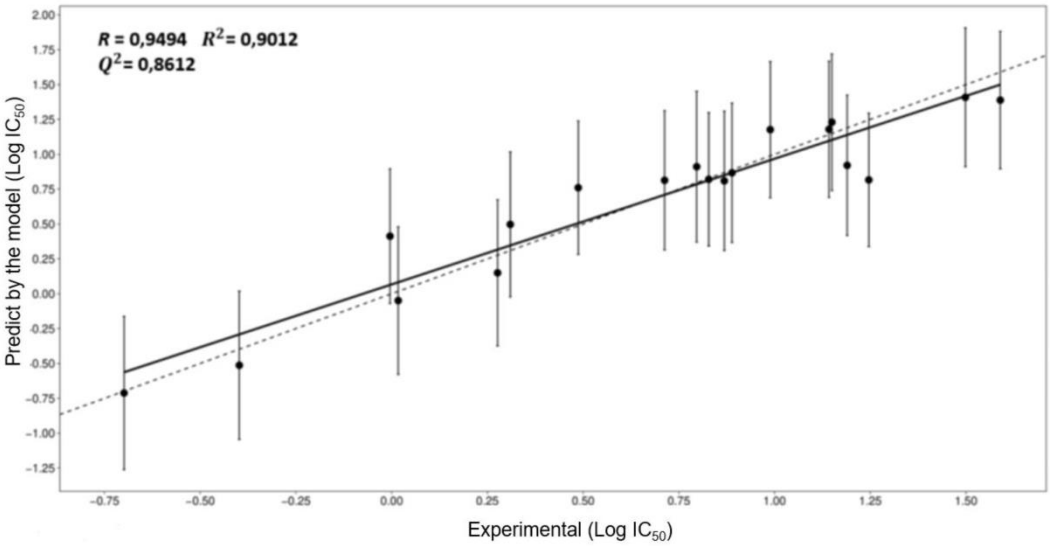
Pic.11 Residual analysis of the adjusted model.



Source: Prepared by the author.

Appendix VI

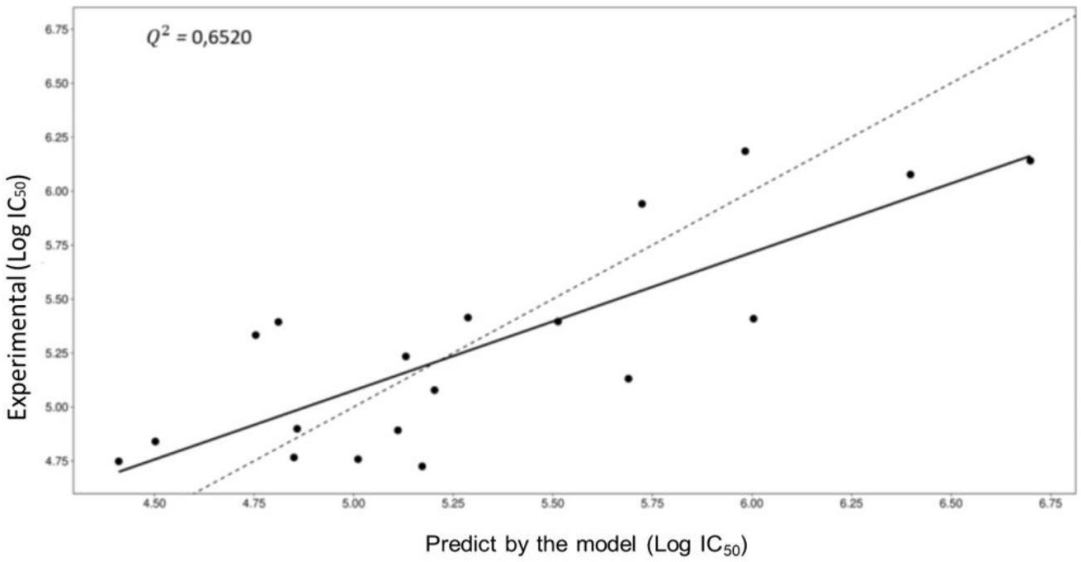
Graph 4. Dispersion of the calculated and observed results of biological activity (Log IC50) from the training set in classical QSAR modeling.



Source: Prepared by the author.

Appendix VII

Graph 5 Dispersion of the calculated and observed results of the biological activity ($LogIC_{50}$) of the training set in the QSAR CoMFA modeling.



Source: Prepared by the author.