molecules

MDPI

Article

Cinnamic Derivatives as Antitubercular Agents: Characterization by Quantitative Structure–Activity Relationship Studies

Cátia Teixeira [1,*], Cristina Ventura [2], José R. B. Gomes [3], Paula Gomes [1] and Filomena Martins [4,*]

1 LAQV-REQUIMTE, Departamento de Química e Bioquímica da Faculdade de Ciências da Universidade do Porto, P-4169-007 Porto, Portugal; pgomes@fc.up.pt
2 Instituto Superior de Educação e Ciências, P-1750-142 Lisboa, Portugal; cventura@isec.universitas.pt
3 CICECO, Departamento de Química, Universidade de Aveiro, P-3810-193 Aveiro, Portugal; jrgomes@ua.pt
4 Centro de Química e Bioquímica (CQB), Centro de Química Estrutural (CQE), Faculdade de Ciências da Universidade de Lisboa, P-1749-016 Lisboa, Portugal
* Correspondence: catia.teixeira@fc.up.pt (C.T.); feleitao@fc.ul.pt (F.M.)

Received: 19 December 2019; Accepted: 17 January 2020; Published: 21 January 2020

Abstract: Tuberculosis, caused by *Mycobacterium tuberculosis* (*Mtb*), remains one of the top ten causes of death worldwide and the main cause of mortality from a single infectious agent. The upsurge of multi- and extensively-drug resistant tuberculosis cases calls for an urgent need to develop new and more effective antitubercular drugs. As the cinnamoyl scaffold is a privileged and important pharmacophore in medicinal chemistry, some studies were conducted to find novel cinnamic acid derivatives (CAD) potentially active against tuberculosis. In this context, we have engaged in the setting up of a quantitative structure–activity relationships (QSAR) strategy to: (i) derive through multiple linear regression analysis a statistically significant model to describe the antitubercular activity of CAD towards wild-type *Mtb*; and (ii) identify the most relevant properties with an impact on the antitubercular behavior of those derivatives. The best-found model involved only geometrical and electronic CAD related properties and was successfully challenged through strict internal and external validation procedures. The physicochemical information encoded by the identified descriptors can be used to propose specific structural modifications to design better CAD antitubercular compounds.

Keywords: antitubercular agents; cinnamic acids; multi-linear regression analysis; *Mycobacterium tuberculosis*; QSAR model

1. Introduction

Tuberculosis (TB), caused by *Mycobacterium tuberculosis* (*Mtb*), is one of the most devastating infectious diseases, which currently still has high mortality levels [1]. Despite the availability of treatments, 7 million new cases and 1.5 million deaths are the alarming figures recently reported by the World Health Organization (WHO) for 2018 [1]. The massive resurgence of multi- and extensively-drug resistant TB, together with the high susceptibility of HIV-infected persons to the disease, are current important concerns leading to an urgent demand for new and more effective antitubercular drugs [2–5]. One of the approaches being used for this purpose is the identification of new therapeutic uses (repurposing) for molecules that were already approved to treat a specific disease or were previously synthesized but were not found to have a clinical application [6]. This is the case of the fluoroquinolones gatifloxacin and moxifloxacin, marketed in 1999 for the treatment of respiratory tract infections, and which are presently the most valuable second-line anti-TB agents according to the WHO guidelines [7].

Cinnamic acid derivatives (CAD) have a century-old history as antitubercular agents [8]. However, this family of compounds was never fully explored for their antimicrobial activity against *Mtb*, and it was not until a decade ago that some studies were conducted to find novel CAD active against tuberculosis [8–16]. Noteworthy, trans-cinnamic acid was found to be bacteriostatic at 200 μg/mL against *Mycobacterium smegmatis* [9] and was reported to show synergism with some first-line antitubercular agents [9,10,12]. As the cinnamoyl scaffold is a privileged and important pharmacophore in medicinal chemistry, in recent years CAD have also attracted much attention due to their antitumoral, antioxidative, and antimalarial properties [17–19]. In this context, and following our experience in the establishment of biologically relevant quantitative structure–activity relationships (QSAR) [20–23], we have engaged in the setting up of a QSAR strategy to: (i) derive a statistically significant model to describe the antitubercular activity of CAD towards wild-type (*wt*) *Mtb*; and (ii) identify the most relevant properties that have a substantial effect on the antitubercular activity of those derivatives.

QSAR analysis is usually based on the assumption that compounds with similar structures are expected to exhibit similar properties and, therefore, changes in chemical structure are likely to be accompanied by proportional changes in biological activity. Although it is now widely known that this congeneric principle is not as universal as initially thought [24–26], it is still the basis behind many computational QSAR studies, from the time when Hansch established the very first QSAR model to predict chemical solubility up to these days [27]. Due to the current explosive growth of experimental data, mainly originated from high throughput screening campaigns, several QSAR methodologies have been called up to establish models involving large and complex data sets [28–30].

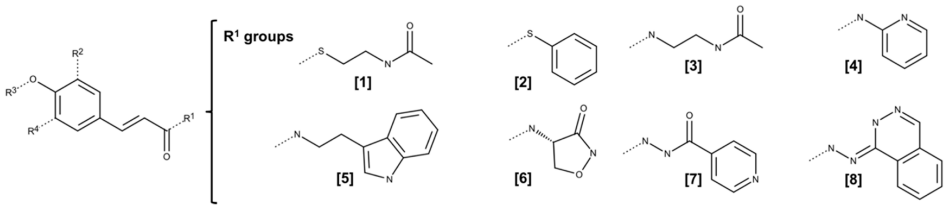
Differences among the various QSAR approaches depend mostly on the descriptors used to characterize the molecules and the methods used to establish relationships between input descriptor values and biological activities. The choice of a particular method depends mainly on the nature of the problem being addressed and on the final purpose of the analysis [22,23]. Linear methods are usually used if the main objective is to rationalize and/or interpret a given biological behavior, while nonlinear methods are more commonly employed if the main purpose is to accurately predict a property. However, non-linear methods are prone to overfitting, which occurs when the number of descriptors (ranging from hundreds to thousands) is much greater than the number of samples in the dataset (less than a hundred compounds is common). In this context, as we were handling a small modeling dataset, we chose a multiple linear regression (MLR) analysis, which is one of the most used linear methods to build up QSAR models and has been very profusely and successfully applied in the field of Medicinal Chemistry [31]. Additionally, MLR has the advantage of being easily interpretable, allowing a direct link between a given biological response and the set of molecular features, encoded by the descriptors, which are responsible for that response. In this work, we depict the details of the construction and validation of an MLR-based model to describe the antitubercular activity of a set of CAD and the analysis of the model's descriptors.

2. Results and Discussion

A dataset of 54 CAD with known MIC values for the *Mtb* H37Rv strain was retrieved from ChEMBL [32]. Two different research groups performed the in vitro experiments [10,11,14,16]. However, a QSAR dataset should include biological activity values for all compounds, preferably measured using the same experimental methodology. As this was not the case, and as the quality of the input data has a large influence in the QSAR model quality, the set with the higher number of compounds was selected in this study [11,16]. Thus, the final data set comprised 29 compounds covering a wide range of MIC values from 0.26 to 1560.59 μM. A pool of 33 molecular descriptors (energetic, geometrical, structural, physicochemical, electronic) was generated using MMPro+ or ChemDraw for cLogP (see experimental section for details). The data set was then split into a training set (22 compounds) and an independent test set (7 compounds) as indicated in Table 1. Compounds were selected in such a way that the chemical domain in the two sets was not too dissimilar and that both the training and the test sets spanned, separately, the entire descriptor space occupied by both sets (Supplementary Materials Table

S1). The training set was used to derive the model, whereas the test set was used to evaluate the predictive ability of the generated model.

Table 1. Structure of the cinnamic acid derivatives (CAD) used to derive the multiple linear regression (MLR)-based model. Compounds marked with an asterisk belong to the test set.

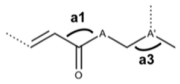


Cpd	R^1	R^2	R^3	R^4	MIC $H_{37}Rv$ (μM)[2]	Cpd	R^1	R^2	R^3	R^4	MIC $H_{37}Rv$ (μM)[2]
1	[1]	H	farnesyl	H	1.28	***16**	[5]	H	isopentenyl	H	168.23
2	[1]	H	isopentenyl	H	95.97	**17**[1]	[5]	OCH_3	H	H	23.78
3	[1]	H	methyl	H	225.52	***18**	[6]	H	methyl	H	950.00
***4**	[1]	OCH_3	H	OCH_3	384.16	**19**	[7]	H	isopentenyl	H	2.30
5	[1]	OCH_3	H	H	423.21	**20**	[7]	H	CF_3	H	1.10
6	[1]	H	H	H	237.44	**21**	[7]	H	CF_3CH_2	H	2.20
7	[2]	OCH_3	H	H	27.94	**22**	[7]	H	geranyl	H	1.90
8	[2]	H	H	H	31.21	***23**	[7]	H	ethyl	H	1.30
9	[3]	H	geranyl	H	0.26	***24**	[8]	H	isopentenyl	H	21.00
10[1]	[3]	H	isopentenyl	H	199.11	**25**	[8]	H	CF_3CH_2	H	20.00
11	[4]	H	isopentenyl	H	51.88	**26**	[8]	H	ethyl	H	12.00
***12**	[4]	H	methyl	H	247.75	**27**	[8]	H	CF_3	H	21.00
13	[4]	OCH_3	methyl	H	439.65	**28**[1]	[8]	H	geranyl	H	72.00
14	[5]	H	methyl	H	1560.59	**29**	[8]	H	methyl	H	50.00
***15**	[5]	H	geranyl	H	72.30						

[1] Compound identified as outlier. [2] MIC values were retrieved from references [11,16]. * Test set compounds.

To establish a relationship between the MIC value and the molecular characteristics of the training set compounds, we derived standard MLR-based QSAR models, testing all combinations of the 33 descriptors and retaining or disregarding descriptors according to rigorous statistical criteria. The intercorrelation matrix among descriptors was always checked; descriptors were considered not intercorrelated, and therefore non-redundant, if r^2 between any two descriptors was below 0.5 and R^2 of one against all others was below 0.8 [21,22,33]. Suspicious points were initially spotted by inspection of a plot of Y_{calc} vs. Y_{exp} and then confirmed as outliers according to two criteria: the conventional measure $|Y_{calc} - Y_{exp}| > 2$ SD, where SD stands for standard deviation of the fit, and a more refined measure known as the Cook's distance (see experimental section for details) [22]. The identified outliers were compounds **10**, **17,** and **28**, as seen in Table 1. The occurrence of outliers can happen for many reasons, such as: i) an error in the reported MIC value or in one or several derived descriptors' values; ii) a mechanism of action different from that of the majority of the data set points; or iii) a non-representative sampling design, among others. Still, no plausible explanation could be assigned for the outlier behavior of compounds **10**, **17,** and **28**. The best model, found by a forward stepwise procedure, upon removal of these three compounds from the training set, is shown in Table 2.

Table 2. Best model found by MLR analysis: pMIC = $a_0 + a_1$**a3** + a_2**a1** + a_3**PSA** + a_4**HansPol**. Molecular descriptors are in bold, and values in parenthesis correspond to the significance level of each adjusted parameter. All descriptors were normalized.



$a_0 \pm s(a_0)$ (SL)	a_1**a3**[1] $\pm s(a_1)$ (SL)	a_2**a1**[1] $\pm s(a_2)$ (SL)	a_3**PSA**[2] $\pm s(a_3)$ (SL)	a_4**HansPol**[3] $\pm s(a_4)$ (SL)
−5.061 ± 0.442 (100.00%)	2.899 ± 0.332 (100.00%)	2.082 ± 0.268 (99.99%)	1.673 ± 0.311 (99.99%)	−1.800 ± 0.354 (99.98%)

[1] **a1** and **a3** correspond to the angles represented in picture of Table 2. [2] **PSA**: polar surface area. [3] **HansPol**: Hansen polarity.

The model's robustness (evaluated with training set compounds) was duly assessed, and the best found model fulfilled all the recommended criteria for internal validation (Table 3) such as a determination coefficient (R^2) higher than 0.6, a leave-one-out (LOO) cross validation correlation coefficient (Q^2_{LOO}) higher than 0.6, the *F*-test value ($F = 35$) significant at 99% with its corresponding tabulated value, a small value for the standard deviation of the fit ($SD = 0.357$), and a significance level (SL) of each adjusted parameter higher than 95% [21,29,34,35]. In order to remove any possibility of attributing the quality of the statistics of these models to a chance correlation between the response variable and the descriptors, a Y-randomization test was performed on the developed QSAR model (Supplementary Materials Table S2). We observed a significant decrease in the quality of the randomized models when compared to the original non-randomized one, and therefore it seemed there was no chance correlation, as corroborated by the value of $^{c}R^2_p$, quite above the 0.5 threshold value [36]. Chance correlation was also assessed by applying the Q under influence of K (QUIK) rule, a technique that measures the total correlation of a set of variables and that allows the rejection of models with high predictor collinearity, as proposed by Todeschini [37]. Thus, according to the QUIK rule, our model was not due to chance correlation, as the *xy* correlation was higher than the *x* correlation (Supplementary Materials Table S2). Finally, the absence of intercorrelation between descriptors in the best model (Supplementary Materials Table S3) also indicated that the quality of the statistics was not due to collinearity among descriptors.

Table 3. Summary of statistical results for the best-found quantitative structure–activity relationships (QSAR) model.

Set	*N*[1]	*SD*[2]	R^2[2,3]	*F*[4]	R^2_0[5]	*AE*[6]	*AAE*[7]	*RMSE*[8]	Q^2[9]	r^{-2}_m[10]	Δr^2_m[11]	*CCC*[12]
Training	19	0.357	0.909	35	-	-	-	-	0.930	-	-	-
Test	7	0.297	0.920	58	0.913	0.100	0.260	0.294	0.933	0.879	0.070	0.953

[1] Number of compounds. [2] Standard deviation of fit. [3] Determination coefficient. [4] The F statistics. [5] Determination coefficient of regression through the origin. [6] Average error. [7] Absolute average error. [8] Root-mean square error. [9] Cross-validation correlation coefficient. [10] Average value between observed vs. predicted and predicted vs. observed Roy's parameter, r^2_m, for the test set. [11] Absolute difference between observed vs. predicted and predicted vs. observed Roy's parameter, r^2_m, for the test set. [12] Concordance correlation coefficient.

Internal validation methods, as the ones mentioned above, are very useful to assess whether a model is stable and robust, and whether overfitting occurs. However, it is more and more commonly accepted that the predictive power of a QSAR model should be evaluated by verifying if the model is able to predict the behavior of chemicals not used on the training set. For that purpose, the predictive ability of the best-found model was analyzed using the test set. The external predictivity was confirmed as the established QSAR model fulfilled all the following recommended "classic" criteria for external validation (Table 3): $Q^2_{ext} > 0.5$, $R^2 > 0.6$; $(R^2 - R_0^2)/R^2 < 0.1$; $0.85 < m < 1.15$, where R_0^2 is the test set's regression determination coefficient that goes through the origin, and *m* is the slope of the

regression between the predicted and the experimental values. The parameter r_m^2, proposed by Roy and Paul [29,38,39], was also used to assess external validation of the model (Table 3). This stricter parameter penalizes a model for large differences between predicted and experimental values of the test set compounds not accounted for by Q^2_{ext}, being an indicator of good external predictivity if greater than 0.65. External validation was also performed by using the very demanding concordance correlation coefficient (CCC) [40,41]. This coefficient measures both accuracy (how far the regression line deviates from the concordance line) and precision (how far the observations are from the fitting line) between experimental and predicted values (Table 3), and a minimum value of 0.85 is required as an indicator of good predictive ability. Finally, a scatter plot of predicted vs. experimental values was also obtained, as recent studies have recommended the visual inspection of these plots as important complementary indicators of model predictivity [22,41]. The scatter plot for the best-found model is represented in Figure 1 showing that no systematic deviations from the ideal line were observed.

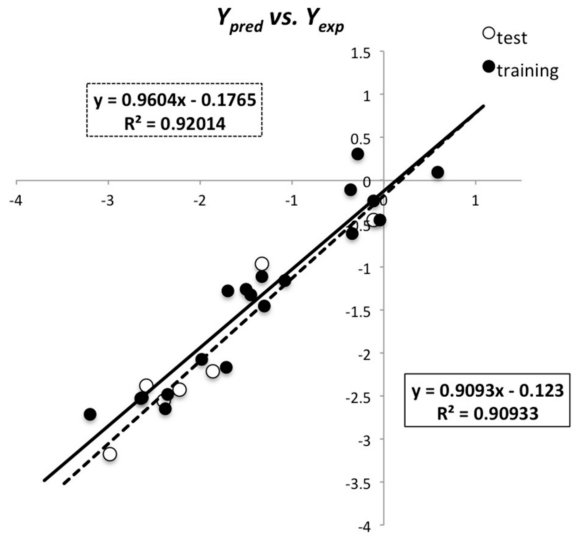


Figure 1. $Log(1/MIC)_{pred}$ vs. $log(1/MIC)_{exp}$ according to the best built QSAR model.

A close analysis of Table 2 reveals that the activity of the studied CAD against *wt Mtb* H37Rv strain did not depend on their energetic, steric, or physicochemical features. Descriptors belonging to these classes were found not to contribute to model log (1/MIC) values. Additionally, the lipophilicity, as measured by cLogP, did not seem significant to explain the antitubercular activity of this family of compounds. On the other hand, geometrical and electronic properties came out as very effective in explaining the biological activity of these derivatives. Indeed, the best-found model included two geometrical descriptors (angles a_1 and a_3 as depicted in Table 2), which both favored activity, suggesting that the sp^2 hybridization for these sets of atoms is preferred over the sp^3. The model also comprises two properties related to the ability of permeation through membranes, polar surface area (PSA) and Hanse polarity parameter (HansPol). This last parameter represents the energy from dipolar intermolecular interactions and contributes negatively to the antitubercular activity of the CAD. Conversely, PSA, which corresponds to the surface sum over all polar atoms (primarily oxygen and nitrogen, including their attached hydrogens), contributes to enhanced activity with an average of 71.3 $Å^2$ for training set and 73.2 $Å^2$ for test set compounds, thus indicating that compounds that are good at permeating membranes are preferred. However, both geometrical descriptors have a relatively higher impact than the two electronic descriptors in the activity of these compounds. Although the cinnamic skeleton has been considered an interesting scaffold for the development of novel antimicrobials, little is known about its mechanism of antimicrobial action. Therefore, no clear relation between a possible mode of action of this family of compounds and the model's descriptors can be made. Still, the results clearly suggest that both penetration through cell membrane and adequate geometrical properties to bind to their target are crucial for the biological activity of CAD.

An additional important aspect to take into consideration is the applicability domain (AD) of the built MLR model, which is crucial to ensure that its predictions are reliable. Thus, to assess the AD for the QSAR model obtained, two different methods were considered: i) the leverage approach (Figure 2) [42], and ii) the range of the individual descriptors (Supplementary Materials Table S1). The leverage value, *h*, provides a measure of the distance of a molecule from the training set's centroid. A "cautionary leverage", *h**, is usually set to $3p/N$, where *N* is the number of molecules in the training set and *p* the number of model descriptors plus one [42]. Thus, plotting the standardized residuals, *SR*, as a function of the leverage values (Williams plot) allows for a graphical assessment of the AD, enabling the detection of influential points, i.e., compounds structurally distant from training set compounds ($h > h^*$) and of response outliers ($SR > \pm 3$ *SR* units). In a Williams plot, the AD is defined by the squared area between ± 3 SR and the threshold *h** value. By analyzing Figure 2, we can observe that the built MLR model performed well in terms of AD. In fact, training and test set compounds lie within ± 3 *SR* units, indicating the absence of any outlier. Additionally, there are also no significant influential points in the training set since the leverage values of all compounds are smaller than the cut-off value *h**. Three compounds from the test set (**12**, **23,** and **24**) had *h* values higher than the warning value *h**, thus falling somewhat outside the model AD. Still, the model accurately predicted these compounds (*SR* within ± 1.2 units), being such points being called "good influential points" [43].

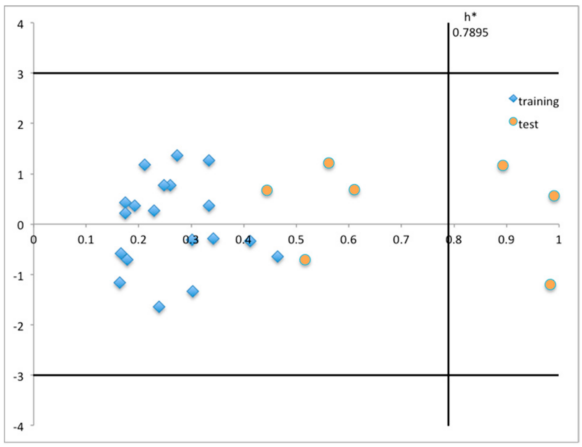


Figure 2. Williams plot for the built model representing the leverage values for the training and test set compounds.

In summary, a set of 29 CAD was investigated to relate their antitubercular activity values to their molecular structure. Using MLR analysis, a stable and predictive QSAR model with good statistical results was developed. The main descriptors involved in the model were related to geometrical and electronic CAD properties. However, more in-depth studies regarding the mechanism of action of the antitubercular activity of CAD should be performed in order to further explore the relationship between QSAR model descriptors and the physicochemical properties of the surface of bacterial cells. Still, the physicochemical meaning of the descriptors of the proposed model will be helpful for rational structural modifications on this class of compounds in order to design better antitubercular agents.

3. Materials and Methods

3.1. Data Set Preparation and Descriptors Calculation

The data set consisted of 29 cinnamic acid derivatives retrieved from ChEMBL [32] with known MIC values for *Mtb* H37Rv strain (Supplementary Materials Table S4) [11,16]. MIC values were converted to the pMIC scale (–log MIC). MarvinSketch was used for molecules' construction and the dominant protonation states of molecules were calculated using the Major Microspecies Plugin, MarvinSketch 16.2.1, ChemAxon [44–47].

A pool of 33 molecular descriptors (Supplementary Materials Table S4) was generated using Molecular Modeling Pro Plus software [44]. Each compound was first submitted to a molecular structure optimization by MM2, a molecular mechanics method incorporated in the software. ChemDraw was used to calculate cLogP values [48].

To determine a relationship between the molecular descriptors of the selected compounds and their respective biological activity, we performed a standard MLR of the type

$$\mathrm{Y} = \mathrm{AX} + \zeta, \tag{1}$$

where ζ is an $n \times 1$ residuals vector whose elements are assumed to be independent normal random variables with mean zero and known variance σ^2, X is a known $n \times k$ matrix of molecular descriptors, A is a $k \times 1$ vector of adjusted parameters, and Y is an $n \times 1$ vector of the response variable related, in this case, to the biological activity. For this purpose, we used the Microsoft Excel Data Analysis add-in and several statistical validation tests to ensure the trustworthiness of the analyses.

3.2. Outlier Search

The decision to consider a given point as an outlier was made according to two criteria: Cook's distance and the more conventional measure $|Y_{calc} - Y_{exp}| > 2$ SD, where SD stands for standard deviation of the fit.

Cook's distance, *D*i, is a measure of the influence of a suspicious point (outlier) in the results of a certain regression and is given by [45,46]

$$D_i = \frac{\sum_i \left(\hat{Y} - \hat{Y}_i\right)^2}{k\sigma^2}, \tag{2}$$

where $\hat{Y}$ and $\hat{Y}_i$ are the $n \times 1$ vectors of the predicted observations for the entire data set and for the data set without the ith observation, respectively, and k is the number of parameters adjusted by the linear model with a variance σ^2. The specific criterion used to exclude a supposed outlier was $D_i > 4/(n - k - 1)$, where n is the number of experimental points.

3.3. Internal Validation

The data set was divided into training (22 compounds) and test (7 compounds) sets with similar degrees of variability. In order to make an internal validation of the data, we applied the leave-one-out (LOO) approach to the training set [22,34,35] as follows:

$$Q^2 = 1 - \frac{\sum_{i=1}^{training} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{training} \left(y_i - \overline{y}_i\right)^2}, \tag{3}$$

where y_i, $\hat{y}_i$ and $\overline{y}_i$ are the measured, predicted, and averaged (over the whole data set) values of the dependent variable, respectively, and Q^2 is a cross-validated correlation coefficient.

We also considered traditional statistical criteria such as the determination coefficient, R^2, the standard deviation, SD, the F statistic, and the significance level, SL, of each adjusted parameter (parameters were kept if SL > 95%) and tested the intercorrelations among all descriptors included in each regression.

3.4. External Validation

The test set was used for external validation, and the predictive ability of the model was assessed by an external Q^2_{ext} parameter defined as [22,34,35]:

$$Q^2_{ext} = 1 - \frac{\sum_{i=1}^{test}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{test}\left(y_i - \overline{y}_{training}\right)^2}, \quad (4)$$

where y_i and $\hat{y}_i$ are the experimental and predicted (over the test set) values, respectively, and $\overline{y}_{training}$ is the averaged value of the dependent variable for the training set.

To further assess the predictive capability of the established QSAR model, we also computed three measures of fit, namely the average error (*AE*), the absolute average error (*AAE*), and the root-mean square error (*RMSE*). Additionally, we determined Roy's parameters [38–41], r^2_m and $\overline{r^2_m}$ and the concordance correlation coefficient (*CCC*) [40,41]. The former two criteria were calculated according to the following formula:

$$r^2_m = R^2\left(1 - \sqrt{R^2 - R^2_0}\right), \overline{r^2_m} = \frac{\left(r^2_m + r'^2_m\right)}{2}, \quad (5)$$

where R^2 and R^2_0 are, respectively, the determination coefficients of the regression function, calculated using the experimental and the predicted data of the prediction set, forcing the regression to pass, respectively, through the origin of the axis (R^2_0) or not (R^2). r^2_m is calculated using the experimental values on the ordinate axis, and r'^2_m using them on the abscissa. The latter criterion *CCC* is obtained by

$$CCC = \frac{2\sum_{i=1}^{n_{test}}\left(Y_i - \overline{Y}\right)\left(\hat{Y}_i - \overline{\hat{Y}}\right)}{\sum_{i=1}^{n_{test}}\left(Y_i - \overline{Y}\right)^2 + \sum_{i=1}^{n_{test}}\left(\hat{Y}_i - \overline{\hat{Y}}\right)^2 + n_{test}\left(\hat{Y}_i - \overline{\hat{Y}}\right)^2}, \quad (6)$$

where Y_i and $\hat{Y}_i$ stand for the abscissa and ordinate values of the plot of experimental vs. predicted values (or, similarly the opposite, which causes no difference), n is the number of compounds, and $\overline{Y}$ and $\overline{\hat{Y}}$ correspond to the averages of experimental and predicted values, respectively.

Supplementary Materials: The following are available online at http://www.mdpi.com/1420-3049/25/3/456/s1, Table S1: Range of variability of descriptors for the training and test sets, Table S2: Results of the Y-randomization (30 shuffles) and the QUIK rule for the best model, Table S3: Intercorrelation matrix between any two descriptors and between one descriptor and a linear combination of all other descriptors, Table S4: Dataset compounds in SMILES format and respective MIC values (μM) and values of descriptors (values not normalized).

Author Contributions: Conceptualization, C.T., C.V., J.R.B.G., P.G., and F.M.; methodology, C.T., C.V., and F.M.; validation, C.T., C.V., and F.M.; investigation, C.T.; data curation, C.T.; writing—original draft preparation, C.T., J.R.B.G., P.G., and F.M.; writing—review and editing, C.T., P.G, and F.M.; visualization, C.T., C.V., and F.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Fundação para a Ciência e Tecnologia (FCT), Portugal, grants UID/QUI/50006/2019, PTDC/BTM-SAL/29786/2017, PTDC/QUI/67933/2006, and PTDC/MED-QUI/29036/2017.

Acknowledgments: The authors thank Fundação para a Ciência e Tecnologia (FCT, Portugal) for funding through grants UID/QUI/50006/2019, PTDC/BTM-SAL/29786/2017, PTDC/QUI/67933/2006, and PTDC/MED-QUI/29036/2017.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization. Global Tuberculosis Report. 2019. Available online: https://www.who.Int/tb/global-report-2019 (accessed on 3 December 2019).
2. Brigden, G.; Hewison, C.; Varaine, F. New developments in the treatment of drug-resistant tuberculosis: Clinical utility of bedaquiline and delamanid. *Infect. Drug. Resist.* **2015**, *8*, 367–378. [CrossRef] [PubMed]

3. Gunther, G. Multidrug-resistant and extensively drug-resistant tuberculosis: A review of current concepts and future challenges. *Clin. Med.* **2014**, *14*, 279–285. [CrossRef] [PubMed]
4. Pawlowski, A.; Jansson, M.; Skold, M.; Rottenberg, M.E.; Kallenius, G. Tuberculosis and hiv co-infection. *PLoS Pathog.* **2012**, *8*, e1002464. [CrossRef] [PubMed]
5. Zumla, A.; Chakaya, J.; Centis, R.; D'Ambrosio, L.; Mwaba, P.; Bates, M.; Kapata, N.; Nyirenda, T.; Chanda, D.; Mfinanga, S.; et al. Tuberculosis treatment and management–an update on treatment regimens, trials, new drugs, and adjunct therapies. *Lancet Respir. Med.* **2015**, *3*, 220–234. [CrossRef]
6. Maitra, A.; Bates, S.; Kolvekar, T.; Devarajan, P.V.; Guzman, J.D.; Bhakta, S. Repurposing-a ray of hope in tackling extensively drug resistance in tuberculosis. *Int. J. Infect. Dis.* **2015**, *32*, 50–55. [CrossRef]
7. Pranger, A.D.; van der Werf, T.S.; Kosterink, J.G.W.; Alffenaar, J.W.C. The role of fluoroquinolones in the treatment of tuberculosis in 2019. *Drugs* **2019**, *79*, 161–171. [CrossRef]
8. De, P.; Bedos-Belval, F.; Vanucci-Bacque, C.; Baltas, M. Cinnamic acid derivatives in tuberculosis, malaria and cardiovascular diseases - a review. *Curr. Org. Chem.* **2012**, *16*, 747–768.
9. Asif, M.; Mohd, I. Synthetic methods and pharmacological potential of some cinnamic acid analogues particularly against convulsions. *Prog. Chem. Biochem. Res.* **2019**, *2*, 192–210.
10. Bairwa, R.; Kakwani, M.; Tawari, N.R.; Lalchandani, J.; Ray, M.K.; Rajan, M.G.; Degani, M.S. Novel molecular hybrids of cinnamic acids and guanylhydrazones as potential antitubercular agents. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 1623–1625. [CrossRef]
11. De, P.; Koumba Yoya, G.; Constant, P.; Bedos-Belval, F.; Duran, H.; Saffon, N.; Daffe, M.; Baltas, M. Design, synthesis, and biological evaluation of new cinnamic derivatives as antituberculosis agents. *J. Med. Chem.* **2011**, *54*, 1449–1461. [CrossRef]
12. Eedara, B.B.; Tucker, I.G.; Zujovic, Z.D.; Rades, T.; Price, J.R.; Das, S.C. Crystalline adduct of moxifloxacin with trans-cinnamic acid to reduce the aqueous solubility and dissolution rate for improved residence time in the lungs. *Eur. J. Pharm. Sci.* **2019**. [CrossRef] [PubMed]
13. Guzman, J.D. Natural cinnamic acids, synthetic derivatives and hybrids with antimicrobial activity. *Molecules* **2014**, *19*, 19292–19349. [CrossRef] [PubMed]
14. Kakwani, M.D.; Suryavanshi, P.; Ray, M.; Rajan, M.G.; Majee, S.; Samad, A.; Devarajan, P.; Degani, M.S. Design, synthesis and antimycobacterial activity of cinnamide derivatives: A molecular hybridization approach. *Bioorg. Med. Chem. Lett.* **2011**, *21*, 1997–1999. [CrossRef] [PubMed]
15. Liu, Q.; Liu, Z.; Sun, C.; Shao, M.; Ma, J.; Wei, X.; Zhang, T.; Li, W.; Ju, J. Discovery and biosynthesis of atrovimycin, an antitubercular and antifungal cyclodepsipeptide featuring vicinal-dihydroxylated cinnamic acyl chain. *Org. Lett.* **2019**, *21*, 2634–2638. [CrossRef] [PubMed]
16. Yoya, G.K.; Bedos-Belval, F.; Constant, P.; Duran, H.; Daffe, M.; Baltas, M. Synthesis and evaluation of a novel series of pseudo-cinnamic derivatives as antituberculosis agents. *Bioorg. Med. Chem. Lett.* **2009**, *19*, 341–343. [CrossRef]
17. Chung, H.S.; Shin, J.C. Characterization of antioxidant alkaloids and phenolic acids from anthocyanin-pigmented rice (oryza sativa cv. Heugjinjubyeo). *Food Chem.* **2007**, *104*, 1670–1677. [CrossRef]
18. De, P.; Baltas, M.; Bedos-Belval, F. Cinnamic acid derivatives as anticancer agents-a review. *Curr. Med. Chem.* **2011**, *18*, 1672–1703. [CrossRef]
19. Teixeira, C.; Vale, N.; Perez, B.; Gomes, A.; Gomes, J.R.; Gomes, P. "Recycling" classical drugs for malaria. *Chem. Rev.* **2014**, *114*, 11164–11220. [CrossRef]
20. Kovalishyn, V.; Aires-de-Sousa, J.; Ventura, C.; Elvas Leitão, R.; Martins, F. Qsar modeling of antitubercular activity of diverse organic compounds. *Chemom. Intell. Lab. Syst.* **2011**, *107*, 69–74. [CrossRef]
21. Martins, F.; Santos, S.; Ventura, C.; Elvas-Leitao, R.; Santos, L.; Vitorino, S.; Reis, M.; Miranda, V.; Correia, H.F.; Aires-de-Sousa, J.; et al. Design, synthesis and biological evaluation of novel isoniazid derivatives with potent antitubercular activity. *Eur. J. Med. Chem.* **2014**, *81*, 119–138. [CrossRef]
22. Martins, F.; Ventura, C.; Santos, S.; Viveiros, M. Qsar based design of new antitubercular compounds: Improved isoniazid derivatives against multidrug-resistant tb. *Curr. Pharm. Des.* **2014**, *20*, 4427–4454. [CrossRef] [PubMed]
23. Ventura, C.; Latino, D.A.; Martins, F. Comparison of multiple linear regressions and neural networks based qsar models for the design of new antitubercular compounds. *Eur. J. Med. Chem.* **2013**, *70*, 831–845. [CrossRef] [PubMed]

24. Dimova, D.; Stumpfe, D.; Bajorath, J. Method for the evaluation of structure-activity relationship information associated with coordinated activity cliffs. *J. Med. Chem.* **2014**, *57*, 6553–6563. [CrossRef] [PubMed]
25. Maggiora, G.M. On outliers and activity cliffs–why qsar often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535. [CrossRef] [PubMed]
26. Nikolova, N.; Jaworska, J. Approaches to measure chemical similarity – a review. *QSAR Comb. Sci.* **2003**, *22*, 1006–1026. [CrossRef]
27. Hansch, C.; Fujita, T. P-σ-π analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626. [CrossRef]
28. Butkiewicz, M.; Lowe, E.W., Jr.; Mueller, R.; Mendenhall, J.L.; Teixeira, P.L.; Weaver, C.D.; Meiler, J. Benchmarking ligand-based virtual high-throughput screening with the pubchem database. *Molecules* **2013**, *18*, 735–756. [CrossRef]
29. Cherkasov, A.; Muratov, E.N.; Fourches, D.; Varnek, A.; Baskin, II; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y.C.; Todeschini, R.; et al. Qsar modeling: Where have you been? Where are you going to? *J. Med. Chem.* **2014**, *57*, 4977–5010. [CrossRef]
30. Ekins, S.; Freundlich, J.S.; Reynolds, R.C. Are bigger data sets better for machine learning? Fusing single-point and dual-event dose response data for mycobacterium tuberculosis. *J. Chem. Inf. Model.* **2014**, *54*, 2157–2165. [CrossRef]
31. van de Waterbeemd, H.; Rose, S. Chapter 23 - quantitative approaches to structure–activity relationships a2 - wermuth, camille georges. In *The Practice of Medicinal Chemistry*, 3rd ed.; Academic Press: New York, NY, USA, 2008; pp. 491–513.
32. Bento, A.P.; Gaulton, A.; Hersey, A.; Bellis, L.J.; Chambers, J.; Davies, M.; Kruger, F.A.; Light, Y.; Mak, L.; McGlinchey, S.; et al. The chembl bioactivity database: An update. *Nucleic Acids Res.* **2014**, *42*, D1083–D1090. [CrossRef]
33. Livingstone, D. Data pre-treatment and variable selection. In *A Practical Guide to Scientific Data Analysis*; Livingstone, D., Ed.; Wiley: Chichester, UK, 2009; pp. 57–73.
34. Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graph. Model.* **2002**, *20*, 269–276. [CrossRef]
35. Tropsha, A.; Gramatica, P.; Gombar, V.K. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of qspr models. *QSAR Comb. Sci.* **2003**, *22*, 69–77. [CrossRef]
36. Mitra, I.; Saha, A.; Roy, K. Exploring quantitative structure–activity relationship studies of antioxidant phenolic compounds obtained from traditional chinese medicinal plants. *Mol. Simul.* **2010**, *36*, 1067–1079. [CrossRef]
37. Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. Detecting "bad" regression models: Multicriteria fitness functions in regression analysis. *Anal. Chim. Acta* **2004**, *515*, 199–208. [CrossRef]
38. Pratim Roy, P.; Paul, S.; Mitra, I.; Roy, K. On two novel parameters for validation of predictive qsar models. *Molecules* **2009**, *14*, 1660–1701. [CrossRef] [PubMed]
39. Roy, K.; Mitra, I.; Kar, S.; Ojha, P.K.; Das, R.N.; Kabir, H. Comparative studies on some metrics for external validation of qspr models. *J. Chem. Inf. Model.* **2012**, *52*, 396–408. [CrossRef]
40. Chirico, N.; Gramatica, P. Real external predictivity of qsar models: How to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *J. Chem. Inf. Model.* **2011**, *51*, 2320–2335. [CrossRef]
41. Chirico, N.; Gramatica, P. Real external predictivity of qsar models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. *J. Chem. Inf. Model.* **2012**, *52*, 2044–2058. [CrossRef]
42. Netzeva, T.I.; Worth, A.; Aldenberg, T.; Benigni, R.; Cronin, M.T.; Gramatica, P.; Jaworska, J.S.; Kahn, S.; Klopman, G.; Marchant, C.A.; et al. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ecvam workshop 52. *Altern. Lab. Anim.* **2005**, *33*, 155–173. [CrossRef]
43. Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. Qsar applicabilty domain estimation by projection of the training set descriptor space: A review. *Altern. Lab. Anim.* **2005**, *33*, 445–459. [CrossRef]
44. Molecular modeling pro plus, version 6.2.5. Available online: www.chemistry-software.com.
45. Dìaz-Garcìa, J.A.; González-Farìas, G. A note on the cook's distance. *J. Stat. Plan. Inference* **2004**, *120*, 119–136. [CrossRef]

46. Militino, A.F.; Palacios, M.B.; Ugarte, M.D. Outliers detection in multivariate spatial linear models. *J. Stat. Plan. Inference* **2006**, *136*, 125–146. [CrossRef]
47. ChemAxon - Software Solutions and Services for Chemistry & Biology. Available online: https://www.chemaxon.com (accessed on 17 January 2020).
48. ChemDraw – Chemical Communication Software. Available online: https://www.perkinelmer.com/category/chemdraw (accessed on 17 January 2020).