Available online at www.sciencedirect.com
ScienceDirect
Bioorganic & Medicinal Chemistry 15 (2007) 5322–5339
Bioorganic & Medicinal Chemistry

# Computational modeling tools for the design of potent antimalarial bisbenzamidines: Overcoming the antimalarial potential of pentamidine

Maykel Cruz-Monteagudo,[a,b,c,d] Fernanda Borges,[b] Maykel Perez González[c,e,f,*] and M. Natália Dias Soeiro Cordeiro[d,*]

[a]*Applied Chemistry Research Centre, Faculty of Chemistry and Pharmacy, Central University of "Las Villas", Santa Clara 54830, Cuba*
[b]*Physico-Chemical Molecular Research Unit, Department of Organic Chemistry, Faculty of Pharmacy, University of Porto, 4150-047 Porto, Portugal*
[c]*Chemical Bioactives Centre, Central University of "Las Villas", Santa Clara 54830, Cuba*
[d]*REQUIMTE, Department of Chemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal*
[e]*Service Unit, Experimental Sugar Cane Station "Villa Clara-Cienfuegos", Ranchuelo, Villa Clara C.P. 53100, Cuba*
[f]*Department of Organic Chemistry, Vigo University, C.P. 36200 Vigo, Spain*

Received 14 December 2006; revised 24 April 2007; accepted 2 May 2007
Available online 18 May 2007

**Abstract**—Malaria is nowadays a worldwide and serious problem with a significant social, economic, and human cost, mainly in developing countries. In addition, the emergence and spread of resistance to existing antimalarial therapies deteriorate the global malaria situation, and lead thus to an urgent need toward the design and discovery of new antimalarial drugs. In this work, a QSAR predictive model based on GETAWAY descriptors was developed which is able to explain with, only three variables, more than 77% of the variance in antimalarial potency and displays a good internal predictive ability (of 73.3% and 72.9% from leave-one-out cross-validation and bootstrapping analyses, respectively). The performance of the proposed model was judged against other five methodologies providing evidence of the superiority of GETAWAY descriptors in predicting the antimalarial potency of the bisbenzamidine family. Moreover, a desirability analysis based on the final QSAR model showed that to be a useful way of selecting the predictive variable level necessary to obtain potent bisbenzamidines. From the proposed model it is also possible to infer that elevated high atomic masses/polarizabilities/van der Waals volumes could play a negative/positive/positive role in the molecular interactions responsible for the desired drug conformation, which is required for the optimal binding to the macromolecular target. The results obtained point out that our final QSAR model is statistically significant and robust as well as possessing a high predictive effectiveness. Thus, the model provides a feasible and practical tool for looking for new and potent antimalarial bisbenzamidines.


## 1. Introduction

The social, economic, and human toll of parasitic diseases, such as malaria, is nowadays a worldwide and serious problem due to their high morbidity and mortality. Malaria, one of the 10 leading diseases for global disease burden,[1] occurs in around 500 million people and results in a number of deaths between 1 and 3 million each year. The majority of these deaths occur in young children who live in sub-Saharan Africa, and are caused by infection with *Plasmodium falciparum* and *Plasmodium vivax*.[2,3] It must be stressed that in a recent study published in *Nature* it was declared that the global extent of this disease is yet underestimated.[4] The full nature of the economic burden of malaria epidemics remains unclear, although it could be connected with substantial losses to households in the form of foregone income, treatment costs, missed schooling, and decreased agricultural production.[5,6] In fact, malaria is the fourth disease contributing to the loss of life quality (measured by its DALYs—disability adjusted life years),

*Keywords*: Antimalarial; Bisbenzamidines; GETAWAY descriptors; QSAR.
* Corresponding authors at present address. Departamento de Química, FCUP, Rua do Campo Alegre, 687, 4169-007 Porto, Portugal. Tel.: +351 226082802; fax: +351 226082959 (N.C); e-mail addresses: gmailkelcm@yahoo.es; ncordeir@fc.up.pt

0968-0896/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2007.05.034

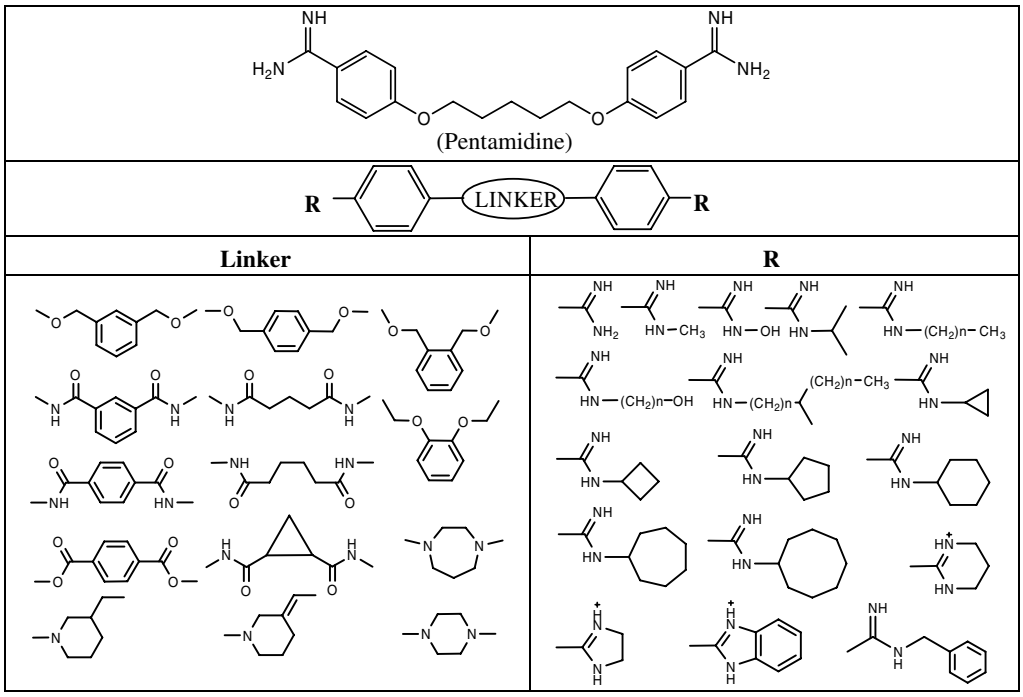



**Figure 1.** General structure of bisbenzamidines and pentamidine molecular structure.

only preceded by respiratory infections, HIV/AIDS, and diarrhea.[3]

Actually, the only malaria therapeutic options are drug combination of the elderly agents (chloroquine plus sulfadoxine-pyrimethamine, artesunate plus sulfadoxine-pyrimethamine, artesunate plus mefloquine, chlorproguanil plus dapsone), or new ones (artemether plus lumefantrine). However, the efficacy of drug combinations is short-lived due to the selection and spread of mutant drug-resistant parasites and to multidrug resistance emergence.[7] It is well known that strains of *P. falciparum* parasites have developed resistance to most of the commonly used antimalarials including sulfadoxine-pyrimethamine (Fansidar) and mefloquine.[8–10]

Consequently, multidrug resistance has led to the need of the development of new antimalarial drugs. For this purpose, Quantitative Structure–Activity Relationship (QSAR) methods have proved their utility into design and drug development.[11–13] Particularly, several QSAR studies have been conducted with success in the field of antimalarial drugs.[14,15]

Three-dimensional (3D) representations are particularly suitable for QSAR purposes due to their ability to simultaneously encode topological and conformational information, being the last one determinant for drug–target interaction.[16] Specifically, *GE*ometry, *T*opology, and *A*tom-*W*eights *A*ssembl*Y* (GETAWAY) descriptors encode geometrical (given by the influence of the molecular matrix) and topological information (given by the molecular graph) weighted by chemical information encoded in selected atomic weightings.[17] The success of GETAWAY descriptors is extensively reported in the literature.[18–21]

In the present work, we intend to build up a predictive model based on GETAWAY descriptors for assessing the antimalarial potential of bisbenzamidines, a promissory chemical family of new antimicrobial drugs.[22–25] Figure 1 depicts the general molecular structure of the present bisbenzamidines, which are pentamidine congeners in which the flexible pentyldioxy linker in pentamidine was replaced with various restricted linkers namely 1,3-, 1,4- or 1,2-phenylenedimethyleneoxy analogues, 1,2-phenylenedioxymethylene, 1,3- or 1,4-phenylenedicarboxamides, 1,4-phenylenedicarboxylate, 1,4-homopiperazinedyl, and 1,4-piperazinedyl.

This approach corresponds to an invaluable tool to contribute to fight malaria infections, throughout the development of new potent drugs. It has great applicability in medicinal chemistry studies, specifically in early stages of drug development, providing rationality to the process, and appears as an interesting approach to the design of new drugs with potential antimalarial activity.

## 2. Results and discussion

### 2.1. Prediction model for antimalarial potential of bisbenzamidines

After reduction of the initial set of 197 GETAWAY molecular descriptors, generated for the set of $N = 42$ conformationally optimized bisbenzamidines, the 115 descriptors selected by GA were used to build up a regression model suitable to predict the antimalarial potential (expressed here as $\log IC_{50}$) of bisbenzamidines. At least nine variables were significantly selected to describe the mentioned activity. In this sense, the ratio between cases and variables ($4 < \rho < 5$) as well as

the significant increments in goodness of fit and predictivity parameters were considered in order to establish the optimum number of variables to be included in the model (see details in Table 1 and Fig. 2). From this study a model which includes eight parameters was found to be the best selection:

$$\begin{aligned}\log IC_{50} = {} & 2.299\ (\pm 0.264) \cdot \mathrm{HTm} \\ & - 1.541\ (\pm 0.179) \cdot \mathrm{H1p} \\ & + 1.029\ (\pm 0.129) \cdot \mathrm{R6v} \\ & - 0.595\ (\pm 0.100) \cdot \mathrm{R8v} \\ & - 0.936\ (\pm 0.168) \cdot \mathrm{H4m} \\ & + 0.689\ (\pm 0.156) \cdot \mathrm{R1e^{+}} \\ & - 0.317\ (\pm 0.113) \cdot \mathrm{R7u^{+}} \\ & + 0.348\ (\pm 0.130) \cdot \mathrm{R8p^{+}} \\ & - 0.678\ (\pm 0.063) \end{aligned} \quad (1)$$

| $N$ | $R^2$ | $R^2_{Adj.}$ | $s$ | $F$ | $p$ | $\rho$ | $Q^2_{LOO}$ | $Q^2_{Boots.}$ | AIC | FIT |
|---|---|---|---|---|---|---|---|---|---|---|
| 42 | 0.783 | 0.730 | 0.407 | 14.879 | $<10^{-8}$ | 4.667 | 0.689 | 0.613 | 0.256 | 1.118 |

As can be seen this model is good in both statistical significance and predictive ability, judging by the attained statistics. Good overall quality of the model is in fact indicated by the large $F$ and small $p$ values, large FIT and small $\rho$ values (model significance), small AIC value (overfitting), along with $R^2$ and $R^2_{Adj.}$ (goodness of fit) as well as $Q^2_{LOO}$ and $Q^2_{Boots.}$ (predictivity) values close to one.

Another aspect deserving special attention is degree of multicollinearity among variables. Highly collinear variables may be identified by examining the cross-correlation matrix (Table 2). One can notice that many variables are correlated, suggesting that multicollinearity may be a problem. As previously mentioned, the primary concern with multicollinearity is that uncertainty is introduced into the estimated coefficients. Thus, to obtain more efficient coefficients orthogonal complements were determined, by Randić's technique, for all variables in Eq. 1.

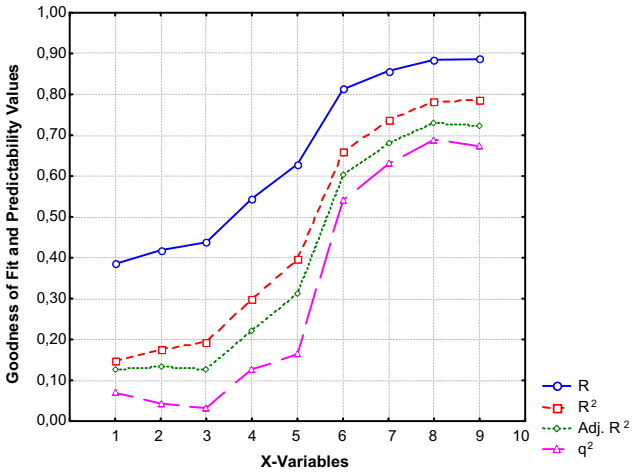



**Figure 2.** Plot of goodness of fit and predictivity parameters for models obtained by successive including the first nine variables selected by the genetic algorithm.

The relating equation given below was obtained:

$$\begin{aligned}\log IC_{50} = {} & 0.028\ (\pm 0.064) \cdot {}^{1}\Omega\mathrm{R1e^{+}} \\ & + 0.314\ (\pm 0.064) \cdot {}^{2}\Omega\mathrm{HTm} \\ & - 0.221\ (\pm 0.064) \cdot {}^{3}\Omega\mathrm{R8v} \\ & - 0.060\ (\pm 0.064) \cdot {}^{4}\Omega H4m \\ & - 0.067\ (\pm 0.064) \cdot {}^{5}\Omega\mathrm{R7u^{+}} \\ & + 0.104\ (\pm 0.064) \cdot {}^{6}\Omega\mathrm{R8p^{+}} \\ & + 0.123\ (\pm 0.064) \cdot {}^{7}\Omega\mathrm{R6v} \\ & - 0.546\ (\pm 0.064) \cdot {}^{8}\Omega\mathrm{H1p} \\ & - 0.678\ (\pm 0.063) \end{aligned} \quad (2)$$

However, inspection of the $t$-values for the coefficient variables included in Eq. 2 by means of a paretto chart (see Fig. 3) revealed that after Randić's orthogonalization some variables were no longer statistically significant. So, they were removed from the equation. This finding is not surprising since after variables' orthogonalization some of them may lose the information content (consequently their significance) encoded in other

**Table 1.** Goodness of fit and predictivity parameters for models obtained by successively-including the first nine variables selected by the genetic algorithm[a]

| Variables included in the model | $R$ | $R^2$ | Adj. $R^2$ | $Q^2$ | $\Delta R^2$ |
|---|---|---|---|---|---|
| HTm | 0.3850 | 0.1482 | 0.1269 | 0.0690 | 0.1482 |
| HTm, H1p | 0.4188 | 0.1754 | 0.1331 | 0.0430 | 0.0272 |
| HTm, H1p, R6v | 0.4378 | 0.1916 | 0.1278 | 0.0320 | 0.0162 |
| HTm, H1p, R6v, R8v | 0.5446 | 0.2966 | 0.2205 | 0.1270 | 0.1050 |
| HTm, H1p, R6v, R8v, H4m | 0.6289 | 0.3956 | 0.3116 | 0.1640 | 0.0990 |
| HTm, H1p, R6v, R8v, H4m, R1e$^+$ | 0.8125 | 0.6601 | 0.6018 | 0.5400 | 0.2645 |
| HTm, H1p, R6v, R8v, H4m, R1e$^+$, R7u$^+$ | 0.8576 | 0.7355 | 0.6810 | 0.6300 | 0.0754 |
| **HTm, H1p, R6v, R8v, H4m, R1e$^+$, R7u$^+$, R8p$^+$** | **0.8848** | **0.7829** | **0.7303** | **0.6890** | **0.0474** |
| HTm, H1p, R6v, R8v, H4m, R1e$^+$, R7u$^+$, R8p$^+$, H0e | 0.8861 | 0.7851 | 0.7247 | 0.6730 | 0.0022 |

[a] The best model found is highlighted in bold.

**Table 2.** Correlation matrix for the original set of descriptors included in Eq. 1

| | $R1e^+$ | HTm | R8v | H4m | $R7u^+$ | $R8p^+$ | R6v | H1p |
|---|---|---|---|---|---|---|---|---|
| $R1e^+$ | 1.00 | | | | | | | |
| HTm | −0.20 | 1.00 | | | | | | |
| R8v | −0.22 | 0.44 | 1.00 | | | | | |
| H4m | −0.34 | 0.84 | 0.36 | 1.00 | | | | |
| $R7u^+$ | 0.60 | −0.09 | −0.44 | −0.24 | 1.00 | | | |
| $R8p^+$ | 0.55 | 0.33 | 0.41 | 0.07 | 0.08 | 1.00 | | |
| R6v | −0.05 | −0.10 | 0.01 | −0.03 | 0.31 | −0.31 | 1.00 | |
| H1p | 0.38 | 0.55 | 0.19 | 0.26 | 0.42 | 0.44 | 0.39 | 1.00 |

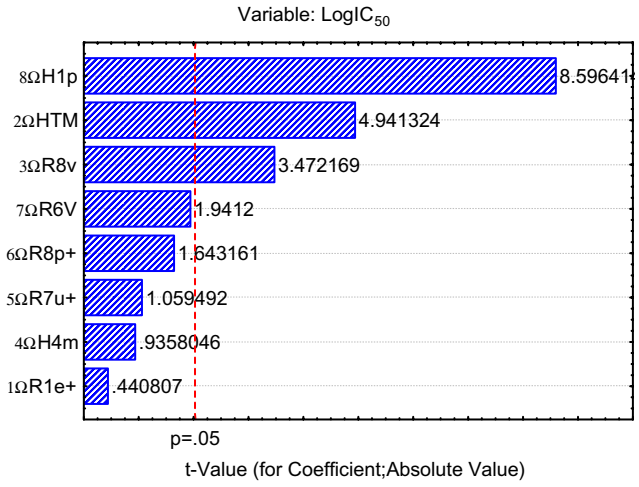



**Figure 3.** Paretto chart of *t*-values for variables' coefficients included in Eq. 2 after Randić's orthogonalization.

linearly correlated variables included in the equation, and has been reported in several preceding QSAR studies.[26–28]

After removing the irrelevant variables a new QSAR equation is obtained, as follows:

$$\begin{aligned}\log IC_{50} = &+0.314\ (\pm 0.067)\cdot {}^{2}\Omega\mathrm{HTm}\\ &-0.221\ (\pm 0.067)\cdot {}^{3}\Omega\mathrm{R8v}\\ &-0.546\ (\pm 0.067)\cdot {}^{8}\Omega\mathrm{H1p}\\ &-0.678\ (\pm 0.066)\end{aligned} \quad (3)$$

| $N$ | $R^2$ | $R^2_{Adj.}$ | $s$ | $F$ | $p$ | $\rho$ | $Q^2_{LOO}$ | $Q^2_{Boots.}$ | AIC | FIT |
|---|---|---|---|---|---|---|---|---|---|---|
| 42 | 0.726 | 0.704 | 0.426 | 33.558 | $<10^{-10}$ | 10.5 | 0.680 | 0.674 | 0.220 | 1.967 |

The new model, with only three variables, retained almost the same performance reached by the previous one containing eight variables. This model is able to explain 72.6% of the variance of antimalarial potency with a high robustness and predictivity, judging from the values of $Q^2_{LOO}$ (68.0%) and $Q^2_{Boots.}$ (67.4%) statistics.

Finally it is important to search for the presence of outliers that might be distorting our QSAR model. The presence of outliers was checked by withdrawal of 10% of the training data, as this is classically accepted in the literature[29] as the threshold for outliers search. Only one compound (2.38% of the training data) was detected as a statistical outlier (compound **26**) and thus removed from the training set. Compound **26** was the only bisbenzamidine showing a value of standardized residual (−3.0592) higher than two standard deviations from the mean value of $\log IC_{50}$, as well as possessing the highest value of delete residual (−1.3623) and Cook's distance (0.11). Compound **26** can also be visually detected as outlier by inspecting the scatter plots in Figure 3 resulting from the residual analysis of Eqs. 3 and 6 (before and after outlier removal). No structural explanation was found to justify the outlier nature of compound **26**, which is the most potent bisbenzamidine included in the training set. This compound is 2- to 3-fold more active than the second more potent bisbenzamidine (compound **18**). From the plethora of factors that could justify the obtained experimental data one could point out a putative difference into the mechanism of antiparasitic action. The ability of bisbenzamidines to exert their toxic effects in a multitude of potential targets is well known in the literature.[30–32] Being clearly an influential outlier, this compound should be removed from the training set to gather homogeneity of the biological response profiles among the compounds in the final training set.

Considering the outlier nature of compound **26** the reader may well call the applicability of the model to highly active bisbenzamidines like this one into question. To answer this question, the applicability domain of the model before and after removal of compound **26** has been determined.

A simple method to investigate the applicability domain of a prediction model is to carry out a leverage plot (plotting residuals vs. leverage of training compounds).[33,34] The leverage ($h$) of a compound in the original variable space which measures its influence on the model is defined as:

$$h_i = x_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x \quad (i = 1, \ldots, n) \quad (4)$$

where $x_i$ is the descriptor vector of the considered compound and $\boldsymbol{X}$ is the model matrix derived from the training set descriptor values. The warning leverage $h^*$ is defined as follows:

$$h^* = 3 \times p'/n \quad (5)$$

where $n$ is the number of training compounds and $p'$ is the number of model adjustable parameters.

Compounds with $h > h^*$($h^* = 0.2857$ before and 0.2926 after removal of compound **26**) are out of the model's applicability domain. As observed in Figure 4 all the compounds in training set lie within the model's applicability domain. Specifically, compound 26 ($h = 0.043$) although in the outlier zone is placed perfectly in the applicability domain before and after being removed. Consequently, we can assert that highly active bisbenzamidines like compound **26** can be predicted perfectly by using the model, even after their removal.

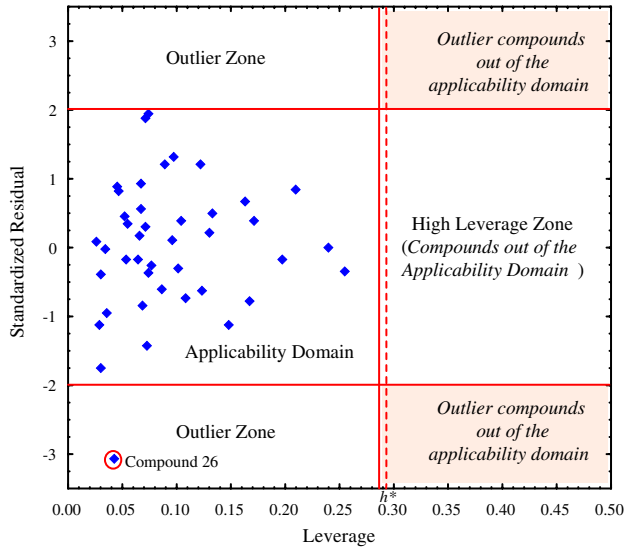



**Figure 4.** Model's applicability domain. Straight/dashed line is the $h^*$ value before/after removal of compound **26**.

In support of the above statement is important to note that 23 out of 42 compounds used for training (apart from compound **26**) were more potent than pentamidine against *P. falciparum*. So, even when this compound is removed from the training set highly active compounds (23 = 54.8% of training set) are steel represented.

When compound **26** was removed from the training set, the following model with the same three variables arises:

$$\begin{aligned}\log IC_{50} = {} & 0.307\ (\pm 0.059) \cdot {}^{2}\Omega HTm \\ & - 0.274\ (\pm 0.065) \cdot {}^{3}\Omega R8v \\ & - 1.506\ (\pm 0.164) \cdot {}^{8}\Omega H1p \\ & - 0.645\ (\pm 0.058) \end{aligned} \quad (6)$$

| $N$ | $R^2$ | $R^2_{Adj.}$ | $s$ | $F$ | $p$ | $\rho$ | $Q^2_{LOO}$ | $Q^2_{Boots.}$ | AIC | FIT |
|---|---|---|---|---|---|---|---|---|---|---|
| 41 | 0.776 | 0.759 | 0.372 | 42.893 | $<10^{-11}$ | 10.250 | 0.733 | 0.729 | 0.168 | 2.562 |

This last QSAR model is now able to explain the 77.62% of the variance of antimalarial potency of the training compounds. Contrasted to Eq. 3 this model also displays a higher predictive ability for new bisbenzamidines of 73.3/72.9% as established by leave-one-out cross-validation/bootstrapping analyses.

From Table 3, it is possible to access further details on the prediction of $IC_{50}$. Specifically, the error in predicting $IC_{50}$ by means of Eq. 6 is just slightly superior to the experimental error demonstrating the predictive ability of the proposed model. Note the very low difference between the mean of the experimental and prediction errors (0.047 μM), which is a very desirable property for a prediction model.

The next step is to find out if the basic assumptions for multiple linear regressions (MLR) are fulfilled.[35] As the name implies, MLR establishes a linear additive relationship between the molecular descriptors and the underlying bioactivity, which is the simplest functional form to adopt with no prior information. Visual inspection of the residuals' distribution for all bisbenzamidines (residuals vs. cases; see a2 in Fig. 5) supports this choice, as no systematic pattern is seen. This finding also corroborates the absence of multicollinearity in our model, an expected result since all the included variables were orthogonalized.[35]

The parametric assumption of homoscedasticity (i.e., homogeneity of variance of the variables) was also checked by simply plotting the square residuals for each predictor variable.[35] The plots in Figure 5(b2) reveal a broad scatter on the points, without any consistent pattern, *post-mortem* validating the pre-adopted assumption of homoscedasticity. As the term related to the error (represented by residuals) is not included in the MLR equation; the mean must be 0 which actually occurs (see Table 4).

Moving on to the last but not least important parametric assumption of MLR, that is, multivariate normality of residuals, it was found that the residuals exhibit adequate values of skewness and kurtosis,[35] which is a sign of normal distribution fitting. Additionally, the hypothesis of multivariate normality of residuals is confirmed from their histograms, as well as the results of applying the Kolmogorov–Smirnov, including Lilliefors' correction, and Shapiro–Wilks hypothesis tests (Table 4).

### 2.2. Comparison of GETAWAY based QSAR model with other methodologies

A deep comparison between the QSAR model proposed here, which was developed using the family of GETAWAY descriptors, and other five different methodologies was carried out. The results obtained are presented in Table 5. Every model was developed using eight variables, except our last model (Eq. 6).

As seen, models from Constitutional, Geometrical, and WHIM descriptors have comparable $R^2$ values (0.618, 0.621, and 0.656, respectively) which are nevertheless lower than the one reported for the GETAWAY model. Galvez topological charge indexes and BCUT descriptors are included in another group with $R^2$ values even lower than those reported for the group analyzed previously (0.298 and 0.594, respectively). All of them are lower than 0.6.

It is important to highlight the low predictive ability of all of the tested models with respect to GETAWAY model. With the exception of the model developed by using WHIM descriptors (with a $Q^2$ value of 0.504), the $Q^2$ value for the other models was lower than 0.31. The only model with a discrete predictive capability besides that of the GETAWAY model is the WHIM model. However, statistical parameters of the WHIM model are significantly lower than those of the GETAWAY descriptors. Finally, one can assert that the proposed GETAWAY model is the simplest one that contains

**Table 3.** Performance of Eq. 6 in predicting the bisbenzamidines' antimalarial potency

| Compound[a] | Exp. $IC_{50}$ (μM) | Pred. $IC_{50}$ (μM) | $\Delta IC_{50}$ (μM) | SEM[b] | Obs. $\log IC_{50}$ | Pred. $\log IC_{50}$ | Res. |
|---|---|---|---|---|---|---|---|
| **1** | 0.085 | 0.24807 | −0.16307 | 0 | −1.07058 | −0.60542 | −0.46516 |
| **2** | 0.114 | 0.05638 | 0.05762 | ±0.04 | −0.94310 | −1.24884 | 0.30575 |
| **3** | 0.313 | 0.61614 | −0.30314 | 0 | −0.50446 | −0.21032 | −0.29414 |
| **4** | 0.671 | 0.39427 | 0.27673 | 0 | −0.17328 | −0.40420 | 0.23093 |
| **5** | 0.697 | 0.65850 | 0.03850 | ±0.06 | −0.15677 | −0.18144 | 0.02467 |
| **7** | 4.12 | 3.63560 | 0.48440 | ±0.11 | 0.61490 | 0.56058 | 0.05432 |
| **8** | 0.64 | 0.95243 | −0.31243 | ±0.04 | −0.19382 | −0.02117 | −0.17265 |
| **11** | 0.402 | 0.28732 | 0.11468 | ±0.01 | −0.39577 | −0.54163 | 0.14586 |
| **12** | 0.84 | 0.29740 | 0.54260 | ±0.03 | −0.07572 | −0.52666 | 0.45094 |
| **13** | 1.6 | 0.65024 | 0.94976 | ±0.04 | 0.20412 | −0.18693 | 0.39105 |
| **14** | 0.148 | 0.24087 | −0.09287 | ±0.05 | −0.82974 | −0.61821 | −0.21153 |
| **15** | 5.11 | 0.89963 | 4.21037 | ±0.2 | 0.70842 | −0.04594 | 0.75436 |
| **17** | 0.306 | 0.10974 | 0.19626 | ±0.04 | −0.51428 | −0.95962 | 0.44535 |
| **18** | 0.018 | 0.03628 | −0.01828 | ±0.009 | −1.74473 | −1.44030 | −0.30443 |
| **19** | 0.154 | 0.04386 | 0.11014 | ±0.02 | −0.81248 | −1.35790 | 0.54542 |
| **20** | 0.862 | 0.13960 | 0.72240 | ±0.01 | −0.06449 | −0.85511 | 0.79062 |
| **21** | 0.124 | 0.20096 | −0.07696 | ±0.06 | −0.90658 | −0.69689 | −0.20968 |
| **22** | 0.029 | 0.12749 | −0.09849 | ±0.01 | −1.53760 | −0.89454 | −0.64306 |
| **23** | 0.245 | 0.58907 | −0.34407 | ±0.05 | −0.61083 | −0.22984 | −0.38100 |
| **24** | 0.358 | 0.48463 | −0.12663 | ±0.04 | −0.44612 | −0.31459 | −0.13153 |
| **25** | 0.019 | 0.04685 | −0.02785 | ±0.004 | −1.72125 | −1.32930 | −0.39195 |
| **27** | 0.021 | 0.01567 | 0.00533 | ±0.004 | −1.67778 | −1.80495 | 0.12717 |
| **28** | 0.042 | 0.05674 | −0.01474 | ±0.005 | −1.37675 | −1.24610 | −0.13065 |
| **29** | 0.069 | 0.06089 | 0.00811 | ±0.019 | −1.16115 | −1.21543 | 0.05428 |
| **30** | 0.297 | 0.14308 | 0.15392 | ±0.015 | −0.52724 | −0.84441 | 0.31717 |
| **31** | 0.035 | 0.04542 | −0.01042 | ±0.005 | −1.45593 | −1.34279 | −0.11315 |
| **32** | 0.036 | 0.11916 | −0.08316 | 0 | −1.44370 | −0.92386 | −0.51983 |
| **33** | 0.048 | 0.03107 | 0.01693 | ±0.009 | −1.31876 | −1.50760 | 0.18884 |
| **34** | 0.036 | 0.09647 | −0.06047 | ±0.002 | −1.44370 | −1.01562 | −0.42808 |
| **35** | 0.138 | 0.13767 | 0.00033 | ±0.002 | −0.86012 | −0.86116 | 0.00104 |
| **36** | 0.055 | 0.04246 | 0.01254 | ±0.008 | −1.25964 | −1.37203 | 0.11239 |
| **37** | 0.088 | 0.09802 | −0.01002 | ±0.011 | −1.05552 | −1.00867 | −0.04684 |
| **38** | 0.075 | 0.44212 | −0.36712 | ±0.043 | −1.12494 | −0.35446 | −0.77048 |
| **39** | 3.89 | 1.60732 | 2.28268 | ±0.07 | 0.58995 | 0.20610 | 0.38385 |
| **40** | 0.101 | 0.05494 | 0.04606 | ±0.077 | −0.99568 | −1.26011 | 0.26443 |
| **41** | 8.8 | 8.80115 | −0.00115 | ±1.1 | 0.94448 | 0.94454 | −0.00006 |
| **42** | 8.32 | 5.57045 | 2.74955 | ±0.7 | 0.92012 | 0.74589 | 0.17423 |
| **43** | 0.089 | 0.07169 | 0.01731 | 0 | −1.05061 | −1.14455 | 0.09394 |
| **44** | 4.88 | 6.56586 | −1.68586 | ±0.46 | 0.68842 | 0.81729 | −0.12887 |
| **45** | 0.106 | 0.16179 | −0.05579 | ±0.016 | −0.97469 | −0.79106 | −0.18364 |
| Pentamidine | 0.278 | 0.59417 | −0.31617 | ±0.08 | −0.55596 | −0.22609 | −0.32987 |
| Mean | 1.07949 | 0.86418 | 0.21530 | ±0.08412 | −0.64179 | −0.64179 | 0.00000 |

[a] The compound numeration is the same as in the original data source.[25]
[b] Standard error.

the most significant descriptors in predicting the antimalarial activity of bisbenzamidines.

### 2.3. Desirability analysis: optimization of antimalarial potential of bisbenzamidines

A typical problem in drug development is to find a set of conditions, or levels of the predictive variables, that generates the most desirable product in terms of characteristics, or responses on the predictor variables. In the development of any drug whose characteristics are known to depend on its properties or molecular descriptors, generating the greatest potential product requires determining the effects of those, and then finding the balance that optimizes the overall desirability of the final product. Typically, this work involves two main steps: (1) predicting responses on the property under study (in this case, the antimalarial potency), by fitting the observed responses using an equation based on the levels of the predictive variables (in this case, the GETAWAY descriptors), and (2) finding the levels of the *X*-variables that simultaneously produce the most desirable predicted responses on the studied property.[36] Considering Eq. 6 as the best choice, a desirability analysis was performed based on the levels of the predictor variables used in this model. The optimal values for obtaining a highest antimalarial drug potential should be around −2.01, 2.03, and 2.01 for $^{2}\Omega$HTm, $^{3}\Omega$R8v, and $^{8}\Omega$H1p, respectively, fixing the other two variables at their present mean values (see Fig. 6). However, if the current values for the three variables (0.00967, 0.01712, and −0.0091 for $^{2}\Omega$HTm, $^{3}\Omega$R8v, and $^{8}\Omega$H1p, respectively) are used it is possible to obtain a desirability value for the antimalarial potency of 0.696. So, through the

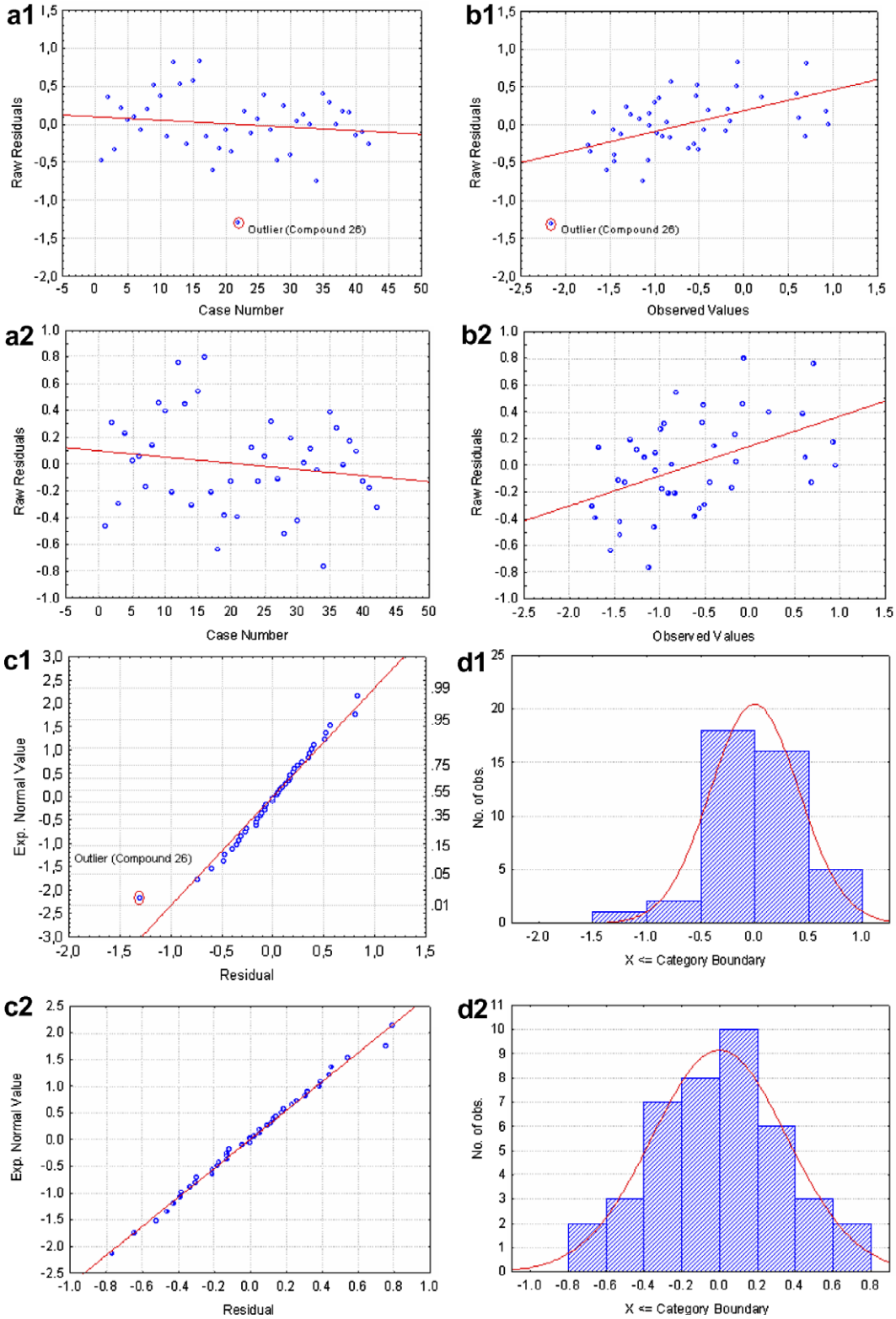



**Figure 5.** Residual analysis carried out to check parametrical assumptions for regression models (a1 and a2). Scatter plot of raw residual versus cases: checking the correct specification of the mathematical (linear) form of the model (b1 and b2). Scatter plot of residual versus observed values of $\log IC_{50}$: checking the homoscedasticity (c1 and c2). Histogram of residuals frequency distribution: probability plots of residuals (d1 and d2). Checking the normal distribution of residuals: a1, b1, c1, and d1 are related to Eq. 3 (that is to say, before withdrawal of outlier). Note in these graphs the outlier nature of compound **26**. a2, b2, c2, and d2 are related to Eq. 6.

developed model it is possible to obtain bisbenzamidines more potent than pentamidine (Recall that the desirability value of pentamidine was set at 0.5.).

The contour plots in Figure 7 show the overall response desirability produced by different level combinations of two independent variables (fixing the value of the

**Table 4.** Some important descriptive statistics in residual analysis and normality tests

| Valid *N* | Mean | Skewness | Kurtosis |
|---|---|---|---|
| *Residual analysis* | | | |
| 41 | $<10^{-15}$ | 0.1186 | −0.2034 |
| K–S *d* *p* | Lilliefors *p* | Sh–Wilk *W* | *p* |
| *Normality tests* | | | |
| 0.0631 >0.2 | >0.2 | 0.9929 | 0.9958 |

remaining third variable at its mean value). The plots were built by previously transforming scores on each of the three variables into desirability scores (they could range from 0.0—undesirable (in green) to 1.0—very desirable (in red). The red zone in the contour plots represents the zone of higher probability to obtain a drug with the best antimalarial profile, and *vice versa* for the green zone. As can be noticed in Figure 7, $^{3}$ΩR8v and $^{8}$ΩH1p have a positive and synergic influence over the antimalarial potential. On the contrary, $^{2}$ΩHTm affects the antimalarial properties of bisbenzamidines in a negative way.

## 2.4. Interpretation of the model

The molecular descriptors and desirability analyses led to get insight into the structural factors governing the affinity of bisbenzamidines for specific targets in *P. falciparum* structures. It is important to note that all descriptors included in MLR-QSAR model (Eq. 6) provide information on the molecular degree of interaction

**Table 5.** Comparison of the GETAWAY descriptors model with other methodologies[a]

| Family of mol. descriptors | No. var. | *N* | $R^2$ | *s* | *F* | $Q^2$ | $s(Q^2)$ |
|---|---|---|---|---|---|---|---|
| Galvez topological charge | 8 | 42 | 0.2978 | 0.741 | 1.604 | 0.099 | 0.916 |
| Geometrical | 8 | 42 | 0.6209 | 0.537 | 6.778 | 0.283 | 0.739 |
| Constitutional | 8 | 42 | 0.6177 | 0.541 | 6.646 | 0.304 | 0.729 |
| BCUT | 8 | 42 | 0.5944 | 0.556 | 6.044 | 0.314 | 0.724 |
| WHIM | 8 | 42 | 0.6561 | 0.504 | 7.886 | 0.504 | 0.615 |
| GETAWAY (Model 1) | **8** | **42** | **0.7829** | **0.407** | **14.879** | **0.689** | **0.487** |
| GETAWAY (Model 4) | **3** | **41** | **0.7762** | **0.372** | **42.893** | **0.733** | **0.407** |
| *Variables included within models* | | | | | | | |
| Galvez topological charge | GGI2, GGI3, GGI4, GGI5, GGI6, JGI3, JGI5, JGI6 | | | | | | |
| Geometrical | H3D, AGDD, MAXDP, G1, SPAM, MEcc, G(N...N), G(O...O) | | | | | | |
| Constitutional | MW, AMW, Ss, Ms, RBN, RBF, nH, nO | | | | | | |
| BCUT | BEHm2, BEHm3, BELm3, BELm5, BELm7, BELv6, BELe3, BEHp7 | | | | | | |
| WHIM | L2u, G3u, E2u, E3u, E2m, E2v, E1e, Du | | | | | | |
| GETAWAY (Model 1) | HTm, H1p, R6v, R8v, H4m, R1e$^{+}$, R7u$^{+}$, R8p$^{+}$ | | | | | | |
| GETAWAY (Model 4) | ΩHTm, ΩH1p, ΩR8v | | | | | | |

[a] The best model found is highlighted in bold.

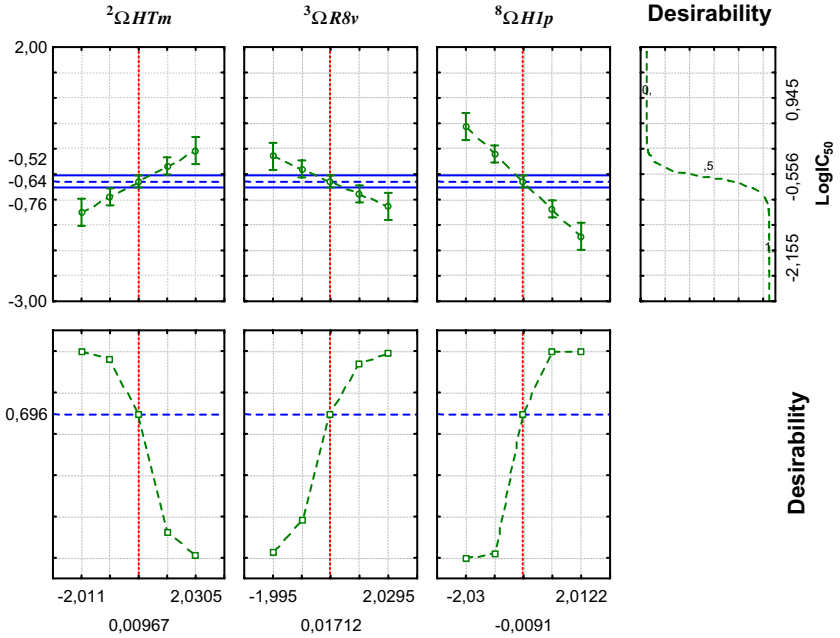



**Figure 6.** Profiles for $\log IC_{50}$ predicted values and desirability derived from Eq. 6.

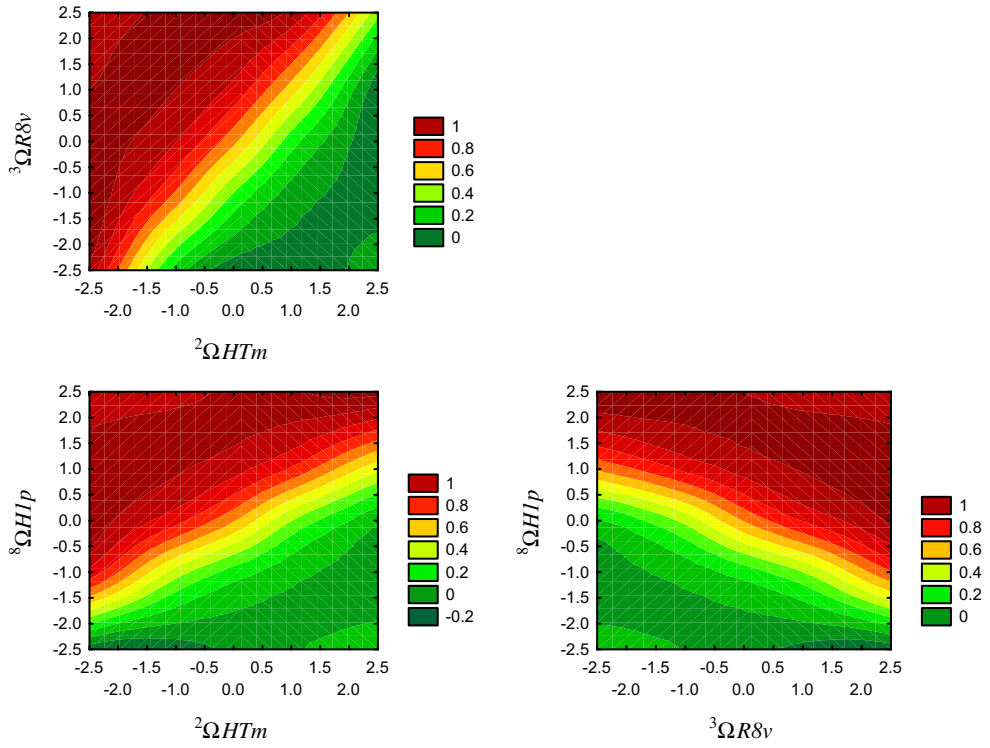



**Figure 7.** Desirability surface/contours plots of $\log IC_{50}$ and desirability derived from Eq. 6.

determined by different atomic properties. The first inference that could be deduced is that the molecular degree of interaction between the pair of atoms in the molecule is determinant for the antimalarial activity of this type of drugs. Probably, this type of interaction(s) orients the preferred molecular conformation for optimal binding to its macromolecular target in the parasite. The data are consistent with the experimental results since it was proposed that conformationally restricted bisbenzamidines revealed a significant increase in antimalarial efficacy.[25] These inferences concerning mechanistic aspects although not entirely novel agree

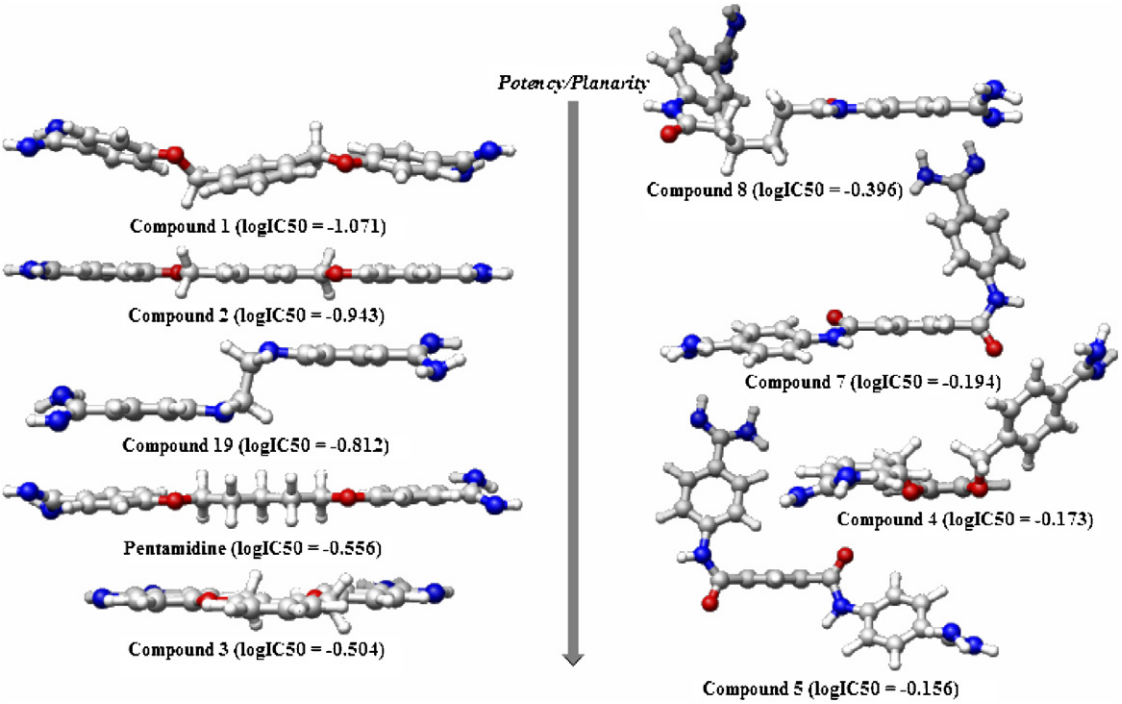



**Figure 8.** Minimal energy conformation of several pentamidine analogues with different linkers. Note the tendency to the decrease of the potency with the loss of the planarity.

with previous experimental findings, reaffirming the validity of the model.

However, when we analyzed the minimal energy conformations of compounds **1**–**8** including compound **19** and pentamidine (pentamidine analogues with dissimilar linkers) some relation between a planar conformation and the antimalarial potency was observed (decrease of antimalarial potency is related with a loss of the planar conformation; see Fig. 8). This finding suggests that the restricted linker is not the only requirement for an improved antimalarial potency, pointing out to the planar conformation (determined by the molecular degree of interaction between the pair of atoms in the molecule) as an additional requirement.

To prove the above mentioned, some modifications were made to the 1,4-piperazinedyl linker of compound **19** (the pentamidine analogue with the most restricted linker) and the antimalarial potency of the new analogues generated was predicted by using Eq. 6 (see Fig. 9). All the modifications lead to a loss of the antimalarial potency according to predictions made with Eq. 6. A complete loss of the planar conformation of all the analogues generated suggests that a planar conformation is required for the antimalarial activity. At the same time, the conformation of the molecular system is mainly determined by the linker as can be noted in Figure 9.

In particular, HTm provides information on the degree of interaction between all the molecule atoms, determined by the atomic masses of every individual atom in the molecule. Judging from Eq. 6 and the results of desirability analysis, elevated atomic masses play a negative role in inducing molecular interactions. This could lead to conformational changes affecting its ability to interact with specific macromolecular targets in *P. falciparum*. On the other hand, H1p provides information on the degree of interaction between atom pairs, determined by their respective atomic polarizabilities. Specifically, it informs about the chance of a given atom *i* to

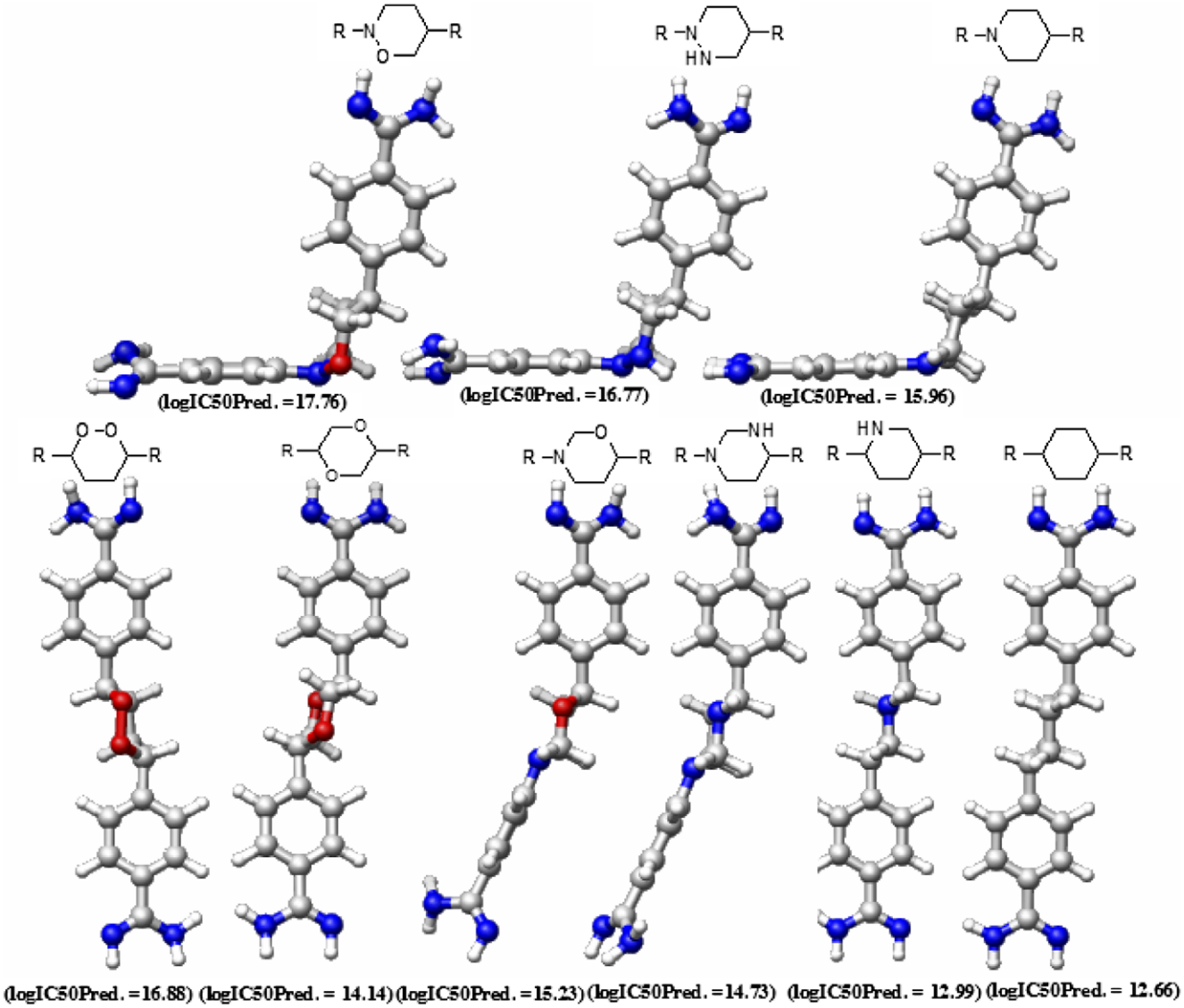



**Figure 9.** Modifications made to the 1,4-piperazinedyl linker of compound **19**.

interact with the atoms $j$ at a topological distance $d_{ij} = 1$ due to their respective atomic polarizabilities. This can be related to the influence of the atomic polarizability over the probability of interaction between the nearer pair of atoms (probably adjacent atoms). In this case, elevated atomic polarizabilities of adjacent atoms play a positive role to favor conformational changes for the antimalarial activity.

The R8v indices, defined in analogy to H-indices, encode similar information. These indices additionally consider the 3D geometric distances between each pair of atoms. In particular, R8v provides information on the chance of a given atom $i$ to interact with those atoms $j$ at a topological distance $d_{ij} = 8$, due to their respective atomic van der Waals volumes. This performance can be related to the influence of the atomic van der Waals volumes over the probability of interaction between distant pair of atoms. So, high atomic van der Waals volumes may contribute positively to the preferential conformational changes for acquiring an optimized activity.

Moreover, the fact that a descriptor derived from the matrix **H** at a topological distance $d_{ij} = 1$ (matrix **H** contains some useful information on the molecular geometry allowing to discriminate among the atoms according to their position in the 3D molecular space with respect to the molecule center) was significantly included on the MLR model corroborates the importance of the central linker of bisbenzamidines for the antimalarial activity. These findings are in agreement with the literature in which a key role for the central linker of bisbenzamidines is proposed to be related with their antiparasitic efficacy.[24,30,37]

Concluding, one can assume that high atomic masses/atomic polarizabilities/atomic van der Waals volumes could play a negative/positive/positive role, respectively, in the molecular interactions responsible for the desired drug conformation required for the optimal binding to its macromolecular target.

## 3. Conclusions

In this work, a QSAR model was developed which successfully fits the antimalarial bisbenzamidines used for training. The model is statistically significant and robust possessing a high predictive effectiveness. It totally fulfills the main MLR assumptions and possesses total superiority, when compared to different methodologies, evidencing the superiority of GETAWAY descriptors in predicting the antimalarial potency of bisbenzamidine drugs.

The desirability analysis performed, based on the final QSAR model, was shown to be a useful way of selecting the predictive variable levels necessary to obtain potent bisbenzamidines. In a near future this type of rational design will allow to overcome some of the problems related with the current antimalarial drugs, such as pentamidine.

From the final model it was also possible to infer that elevated high atomic masses/polarizabilities/van der Waals volumes could play a negative/positive/positive role in the molecular interactions responsible for the desired drug conformation, which is required for the optimal binding to the macromolecular target. Thus, the model provides a feasible and practical tool for looking for new and potent antimalarial bisbenzamidines.

## 4. Materials and methods

### 4.1. Data set

The compounds used to develop our predictive model are included in a library of 52 bisbenzamidines reported by Huang T.L. et al.[25] However, only 42 compounds were reported in exact figures suitable for the regression analysis. Pentamidine was included as positive control. The in vitro anti-plasmodial activity of the lead analogues has been determined against chloroquine-sensitive strain of *P. falciparum* (D6, Sierra Leone) using a lactate dehydrogenase (LDH) assay.[38] Chloroquine and pentamidine were used as the positive controls, while DMSO was the negative control. $IC_{50}$ values (50% inhibitory concentration that inhibits parasite growth by 50% in relation to drug-free control) for each compound were computed from the dose–response curves.[25] The structure of the training set compounds along with the reported experimental $IC_{50}$ values are depicted in Table 6.

### 4.2. Computational strategies

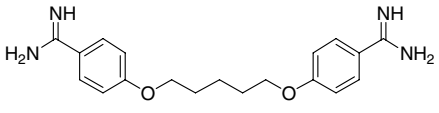
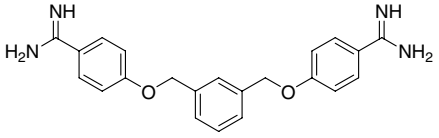
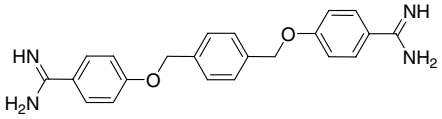
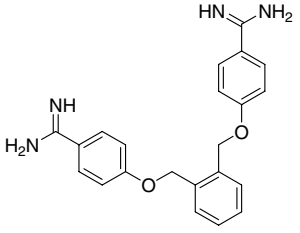
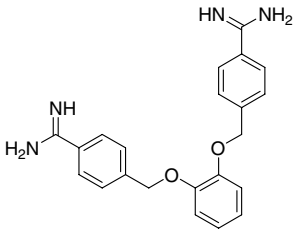
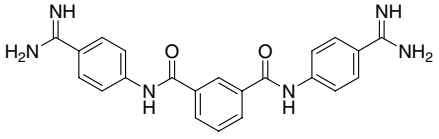
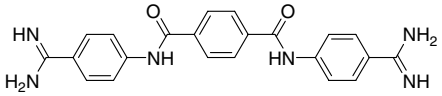
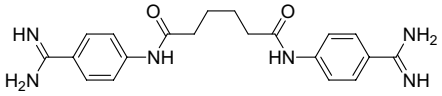
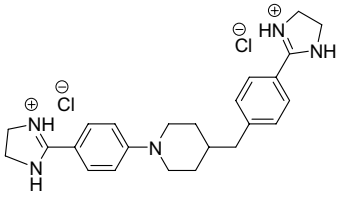
The geometries of the minimum energy conformation of all compounds were obtained by optimization with the semi-empirical PM3 method implemented in MOPAC 6.0.[39] GETAWAY descriptors for the given compounds were calculated on the minimal energy conformations, using software DRAGON 2.1.[40]

A total of 197 GETAWAY molecular descriptors were calculated to describe the structural diversity of the compounds, following the procedure explained in detail by Consonni et al.[17,21,41] Table 7 depicts the names and meanings of the molecular descriptors used in this work. Additionally, Galvez topological charge indices (21 descriptors), Geometrical (58), Constitutional (47), BCUT (64), and WHIM (99) descriptors were computed in order to test the performance of these families in predicting the property under study.[42] The full list of this type of molecular descriptors, and their meaning, is provided in the DRAGON 2.1 package.[40]

To reduce redundant and inadequate information constant or near constant values, molecular descriptors with a pairwise correlation greater than 0.95 were excluded. Thus, from an initial set of 197 GETAWAY molecular descriptors only 115 remained for further variable selection.

As in QSAR studies the choice of appropriate descriptors is a complicated task, because there are no absolute

**Table 6.** Chemical structures for the 42 bisbenzamidines used in regression analysis and their respective 50% inhibitory concentration, inhibiting D6 Sierra Leone strain of *P. falciparum* growth by 50% in relation to drug-free control ($IC_{50}$ values)

| Compound[a] | Structure | $IC_{50}$ (μM) |
|---|---|---|
| Pentamidine |  | 0.278 (±0.08) |
| **1** |  | 0.085 (±0.00) |
| **2** |  | 0.114 (±0.04) |
| **3** |  | 0.313 (±0.00) |
| **4** |  | 0.671 (±0.00) |
| **5** |  | 0.697 (±0.06) |
| **7** |  | 4.12 (±0.11) |
| **8** |  | 0.64 (±0.04) |
| **11** |  | 0.402 (±0.01) |

(*continued on next page*)

**Table 6** (*continued*)

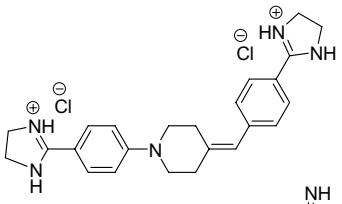
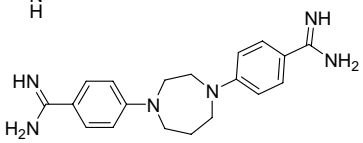
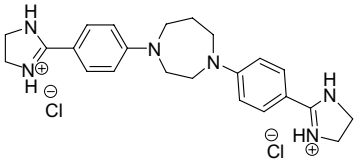
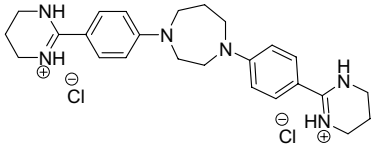
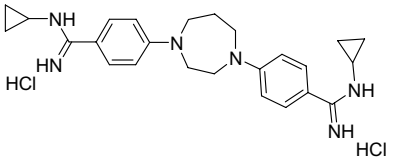
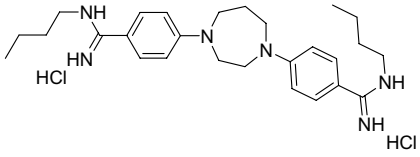
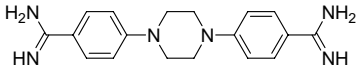
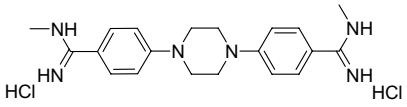
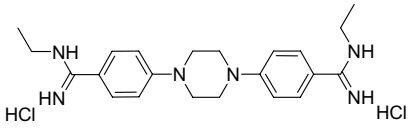
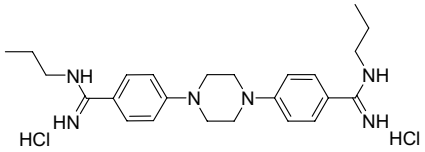
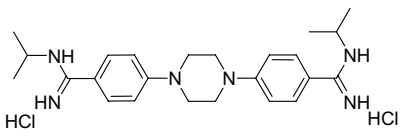
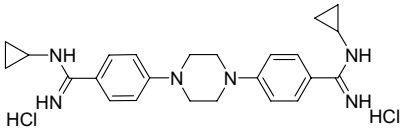
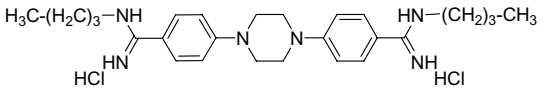
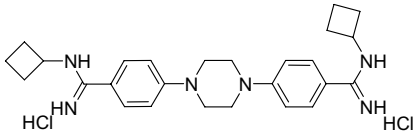
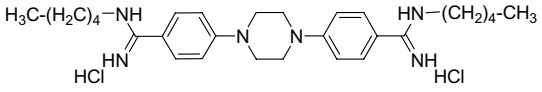
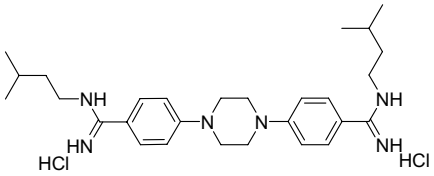
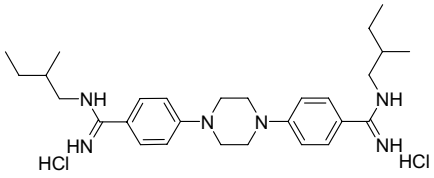
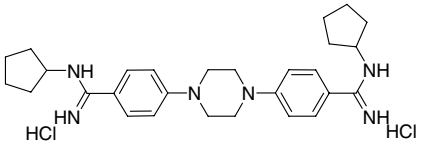
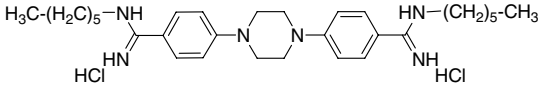
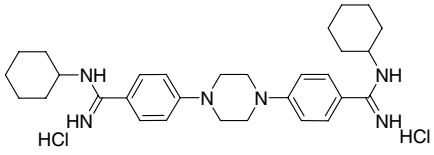
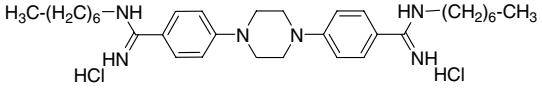
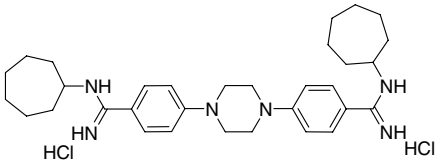
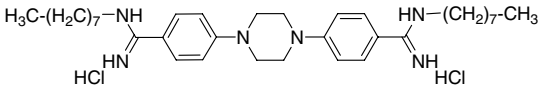
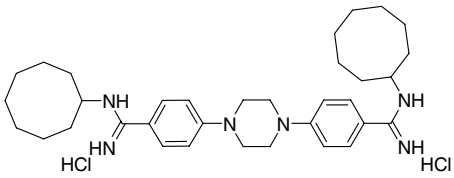
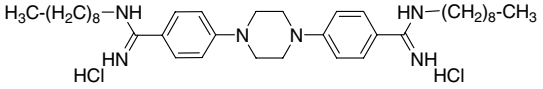
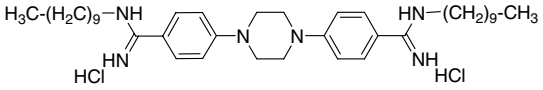
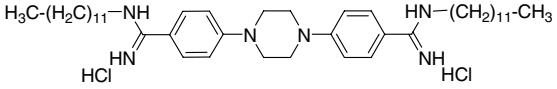
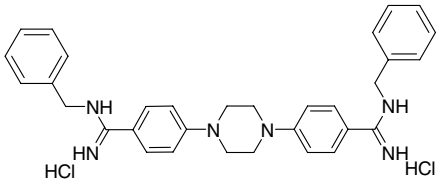
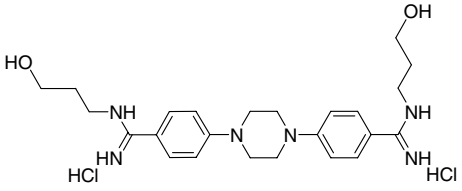
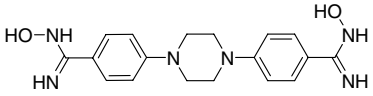
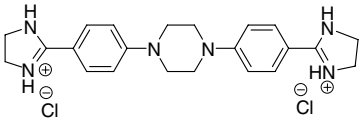
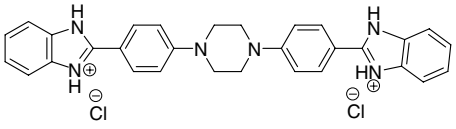
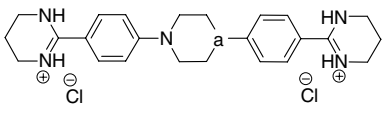
| Compound[a] | Structure | $IC_{50}$ (μM) |
|---|---|---|
| **12** |  | 0.84 (±0.03) |
| **13** |  | 1.6 (±0.04) |
| **14** |  | 0.148 (±0.05) |
| **15** |  | 5.11 (±0.2) |
| **17** |  | 0.306 (±0.04) |
| **18** |  | 0.018 (±0.009) |
| **19** |  | 0.154 (±0.02) |
| **20** |  | 0.862 (±0.01) |
| **21** |  | 0.124 (±0.06) |
| **22** |  | 0.029 (±0.01) |
| **23** |  | 0.245 (±0.05) |

**Table 6** (*continued*)

| Compound[a] | Structure | $IC_{50}$ (μM) |
|---|---|---|
| **24** |  | 0.358 (±0.04) |
| **25** |  | 0.019 (±0.004) |
| **26** |  | 0.007 (±0.001) |
| **27** |  | 0.021 (±0.004) |
| **28** |  | 0.042 (±0.005) |
| **29** |  | 0.069 (±0.019) |
| **30** |  | 0.297 (±0.015) |
| **31** |  | 0.035 (±0.005) |
| **32** |  | 0.036 (±0.00) |
| **33** |  | 0.048 (±0.009) |
| **34** |  | 0.036 (±0.002) |

(*continued on next page*)

**Table 6** (*continued*)

| Compound[a] | Structure | $IC_{50}$ (μM) |
|---|---|---|
| **35** |  | 0.138 (±0.002) |
| **36** |  | 0.055 (±0.008) |
| **37** |  | 0.088 (±0.011) |
| **38** |  | 0.075 (±0.043) |
| **39** |  | 3.89 (±0.07) |
| **40** |  | 0.101 (±0.077) |
| **41** |  | 8.8 (±1.1) |
| **42** |  | 8.32 (±0.7) |
| **43** |  | 0.089 (±0.00) |
| **44** |  | 4.88 (±0.46) |
| **45** |  | 0.106 (±0.016) |

[a] The numeration used for each compound here is the same employed in the original data source.[25]

**Table 7.** Molecular descriptors used on the regression models, generated by the DRAGON software

| Descriptor | Meaning |
|---|---|
| *Molecular walk counts* | |
| MWC05 | Molecular walk count of order 05 |
| MWC07 | Molecular walk count of order 07 |
| MWC09 | Molecular walk count of order 09 |
| MWC10 | Molecular walk count of order 10 |
| TWC | Total walk count |
| SRW04 | Self-returning walk count of order 04 |
| SRW08 | Self-returning walk count of order 08 |
| SRW10 | Self-returning walk count of order 10 |
| *Galvez topological charge indices* | |
| GGI2 | Topological charge index of order 2 |
| GGI3 | Topological charge index of order 3 |
| GGI4 | Topological charge index of order 4 |
| GGI5 | Topological charge index of order 5 |
| GGI6 | Topological charge index of order 6 |
| JGI3 | Mean topological charge index of order3 |
| JGI5 | Mean topological charge index of order5 |
| JGI6 | Mean topological charge index of order6 |
| *Geometrical descriptors* | |
| H3D | 3D-Harary index |
| MAXDP | Maximal electrotopological positive variation |
| AGDD | Average geometric distance degree |
| G1 | Gravitational index G1 |
| SPAM | Average span R |
| MEcc | Molecular eccentricity |
| G(N...N) | Sum of geometrical distances between N...N |
| G(O...O) | Sum of geometrical distances between O...O |
| *Constitutional descriptors* | |
| MW | Molecular weight |
| AMW | Average molecular weight |
| Ss | Sum of Kier-Hall electrotopological states |
| Ms | Mean electrotopological state |
| RBN | Number of rotatable bonds |
| RBF | Rotatable bond fraction |
| nH | Number of Hydrogen atoms |
| nO | Number of Oxygen atoms |
| *BCUT descriptors* | |
| BEHm2 | Highest eigenvalue $n \cdot 2$ of Burden matrix/weighted by atomic masses |
| BEHm3 | Highest eigenvalue $n \cdot 3$ of Burden matrix/weighted by atomic masses |
| BELm3 | Lowest eigenvalue $n \cdot 3$ of Burden matrix/weighted by atomic masses |
| BELm5 | Lowest eigenvalue $n \cdot 5$ of Burden matrix/weighted by atomic masses |
| BELm7 | Lowest eigenvalue $n \cdot 7$ of Burden matrix/weighted by atomic masses |
| BELv6 | Lowest eigenvalue $n \cdot 6$ of Burden matrix/weighted by atomic van der Waals volumes |
| BELe3 | Lowest eigenvalue $n \cdot 3$ of Burden matrix/weighted by atomic Sanderson electronegativities |
| BEHp7 | Highest eigenvalue $n \cdot 7$ of Burden matrix/weighted by atomic polarizabilities |
| *Weighted holistic invariant molecular (WHIM) descriptors* | |
| L2u | 2nd component size directional WHIM index/unweighted |
| G3u | 3st component symmetry directional WHIM index/unweighted |
| E2u | 2nd component accessibility directional WHIM index/unweighted |
| E3u | 3rd component accessibility directional WHIM index/unweighted |
| E2m | 2nd component accessibility directional WHIM index/weighted by atomic masses |
| E2v | 2nd component accessibility directional WHIM index/weighted by atomic van der Waals volumes |
| E1e | 1st component accessibility directional WHIM index/weighted by atomic Sanderson electronegativities |
| Du | D total accessibility index/unweighted |
| *Geometry, topology, and atom-weights assembly GETAWAY) descriptors* | |
| H4m | H autocorrelation of lag 4/weighted by atomic masses |
| HTm | H total index/weighted by atomic masses |
| H1p | H autocorrelation of lag 1/weighted by atomic polarizabilities |
| $R7u^+$ | R maximal autocorrelation of lag 7/unweighted |
| R6v | R autocorrelation of lag 6/weighted by atomic van der Waals volumes |
| R8v | R autocorrelation of lag 8/weighted by atomic van der Waals volumes |
| $R1e^+$ | R maximal autocorrelation of lag 1/weighted by atomic Sanderson electronegativities |
| $R8p^+$ | R maximal autocorrelation of lag 8/weighted by atomic polarizabilities |
| H0e | H autocorrelation of lag 0/weighted by atomic Sanderson electronegativities |

rules governing this selection, an optimization technique—the Genetic Algorithm (GA)—was applied for variable selection.[43–46] In GA, the essence of selection in nature is that under certain environmental conditions, species of high fitness can prevail in the next generation, and the best species may be reproduced by crossover, together with random mutations of chromosomes in the surviving species. As to the modeling technique, we opted for a regression-based approach; in this case, the regression coefficients and statistical parameters were obtained by Multiple Linear Regression (MLR) analysis.

GA and MLR analysis were performed using Mobydigs software.[47] Accordingly, a population of 100 regression models was ordered taking into account the leave-one-out cross-validated correlation coefficients $Q^2$. The GA was stopped when increasing the model size, the $Q^2$ value did not significantly increase.

Last but not least, it is well known that if any multiple linear-statistical based approach is applied to data sets exhibiting collinearities among the *X*-variables, the calculated coefficients get unstable and their interpretability breaks down. A practical approach to avoid multicollinearity is the *orthogonal descriptors* technique proposed by Randić,[48–51] which also provides useful structural input in building QSAR relationships. The *X*-variables were thus orthogonalized following the order selected by the GA search scheme, that is, according to the level of significance to the contribution for the antimalarial property estimated by the increase on model's $Q^2$ after successive additions of each *X*-variable to the model.

### 4.3. Validation of the models

Goodness of fit was assessed by examining the determination coefficient, $R^2$, the adjusted determination coefficient, $R^2_{\text{Adj.}}$, the standard deviation ($s$), Fisher's statistic, $F$, as well as the ratio between the number of compounds and the number of adjustable parameters in the model, known as the $\rho$ statistics. Besides these classic regression parameters, other important indices were adopted, that is, the Kubinyi function (FIT) and Akaike's information criterion (AIC), which give enough criteria for comparing models with different parameters, number of variables, and number of points. The robustness and predictive ability of the model was evaluated by both $Q^2$ based on leave-one-out cross-validation (LOO-CV) and bootstrapping analyses.[34] LOO-CV procedure consists of removing one data point from the training set and constructing the model only on the basis of the remaining training data and then testing on the removed point. In this way, all of the training data points were tested and $Q^2_{\text{LOO}}$ then calculated. In the bootstrapping procedure, $K$ $n$-dimensional groups are generated by a randomly repeated selection of $n$-objects from the original data set. The model obtained on the first selected objects is used to predict the values for the excluded sample, and then $Q^2_{\text{Boots.}}$ is calculated for each model. The bootstrapping was repeated 5000 times for each validated model.

### 4.4. Desirability analysis

Response/desirability profiling allows one to trace the response surface produced by fitting the observed responses using an equation based on levels of the predictor variables.[36] That is to say, one can inspect the predicted values for the dependent variable at different combinations of levels of the *X*-variables, specify desirability functions for the dependent variable, and search for the levels of the *X*-variables that produce the most desirable responses on the dependent variable. In the present work, the desirability analysis was conducted with STATISTICA 6.0,[52] by setting the current level of each predictor variable to the respective mean. Curvature $s$ and $t$ parameters were fixed at 9.00 (Notice that the values chosen for $s$ and $t$ determine the curvature in the desirability function between the low and medium inflection points, and between the medium and high inflection points, respectively.). Large values for $s$ and $t$ must be selected and predicted to be close to the most desirable value of *Y*-observed ($Y_{\text{MAX. DES.}}$). $Y_{\text{MAX. DES.}}$ can be selected anywhere between the minimum acceptable value ($Y_{\text{MIN}}$) of the response and the maximum acceptable value ($Y_{\text{MAX}}$).[36] For this particular case, $Y_{\text{MAX}}$ matches with $Y_{\text{MAX. DES.}}$, thus high $s$ and $t$ values are required. The values of $Y_{\text{MAX}}$ and $Y_{\text{MIN}}$ coincide with the $\log \text{IC}_{50}$ values of the more and less potent compound used for training the model and their respective desirability values were set as 1.00 (highest) for $Y_{\text{MAX}}$ and 0.00 (lowest) for $Y_{\text{MIN}}$. The $\log \text{IC}_{50}$ value of pentamidine was used as the medium acceptable value ($Y_{\text{MED}}$) with a desirability value of 0.500 (medium desirability value). Pentamidine was used as medium threshold value in a desirability analysis allowing the design of a more potent compound for malaria treatment. Least squares method was used for fitting the desirability function and surface/contours maps.

## Acknowledgments

The authors acknowledge the Portuguese *Fundação para a Ciência e a Tecnologia* (FCT) for financial support. Cruz-Monteagudo, M. thanks partial financial support from Physico-Chemical Molecular Research Unit, Department of Organic Chemistry, Faculty of Pharmacy and Department of Chemistry, Faculty of Sciences, University of Porto, Porto, Portugal.

## References and notes

1. Lopez, A. D.; Mathers, C. D.; Ezzati, M.; Jamison, D. T.; Murray, C. J. *Lancet* **2006**, *367*, 1747.
2. Guinovart, C.; Navia, M. M.; Tanner, M.; Alonso, P. L. *Curr. Mol. Med.* **2006**, *6*, 137.
3. Chen, L. L.; Yao, J.; Yang, J. B.; Yang, J. *Acta Pharmacol. Sin.* **2005**, *26*, 1322.
4. Snow, R. W.; Guerra, C. A.; Noor, A. M.; Myint, H. Y.; Hay, S. I. *Nature* **2005**, *434*, 214.
5. Shepard, D. S.; Ettling, M. B.; Brinkmann, U.; Sauerborn, R. *Trop. Med. Parasitol.* **1991**, *42*, 199.
6. Attanayake, N.; Fox-Rushby, J.; Mills, A. *Trop. Med. Int. Health* **2000**, *5*, 595.

7. Wongsrichanalai, C.; Pickard, A. L.; Wernsdorfer, W. H.; Meshnick, S. R. *Lancet Infect. Dis.* **2002**, *2*, 209.
8. Peters, W. *Chemotherapy and Drug Resistance in Malaria*; Academic Press: London, 1987, 543.
9. Wernsdorfer, W. H.; Payne, D. *Pharmacol. Ther.* **1991**, *50*, 95.
10. Wongsrichanalai, C.; Sirichaisinthop, J.; Karwacki, J. J.; Congpuong, K.; Miller, R. S.; Pang, L.; Thimasarn, K. *Southeast Asian J. Trop. Med. Public Health* **2001**, *32*, 41.
11. Gonzalez-Diaz, H.; Cruz-Monteagudo, M.; Vina, D.; Santana, L.; Uriarte, E.; De Clercq, E. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1651.
12. Gonzales-Diaz, H.; Gia, O.; Uriarte, E.; Hernadez, I.; Ramos, R.; Chaviano, M.; Seijo, S.; Castillo, J. A.; Morales, L.; Santana, L.; Akpaloo, D.; Molina, E.; Cruz, M.; Torres, L. A.; Cabrera, M. A. *J. Mol. Model. (Online)* **2003**, *9*, 395.
13. Santana, L.; Uriarte, E.; Gonzalez-Diaz, H.; Zagotto, G.; Soto-Otero, R.; Mendez-Alvarez, E. *J. Med. Chem.* **2006**, *49*, 1149.
14. Xie, A.; Sivaprakasam, P.; Doerksen, R. J. *Bioorg. Med. Chem.* **2006**, *14*, 7311.
15. Marrero-Ponce, Y.; Iyarreta-Veitia, M.; Montero-Torres, A.; Romero-Zaldivar, C.; Brandt, C. A.; Avila, P. E.; Kirchgatter, K.; Machado, Y. *J. Chem. Inf. Model.* **2005**, *45*, 1082.
16. Santos-Filho, O. A.; Mishra, R. K.; Hopfinger, A. J. *J. Comput. Aided Mol. Des.* **2001**, *15*, 787.
17. Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682.
18. Saiz-Urra, L.; Gonzalez, M. P.; Fall, Y.; Gomez, G. *Eur. J. Med. Chem.* **2006**.
19. Gonzalez, M. P.; Teran, C.; Teijeira, M.; Gonzalez-Moa, M. J. *Eur. J. Med. Chem.* **2005**, *40*, 1080.
20. Gonzalez, M. P.; Teran, C.; Teijeira, M.; Besada, P. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2641.
21. Consonni, V.; Todeschini, R.; Pavan, M.; Gramatica, P. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 693.
22. Cushion, M. T.; Walzer, P. D.; Ashbaugh, A.; Rebholz, S.; Brubaker, R.; Vanden Eynde, J. J.; Mayence, A.; Huang, T. L. *Antimicrob. Agents Chemother.* **2006**, *50*, 2337.
23. Mayence, A.; Vanden Eynde, J. J.; LeCour, L., Jr.; Walker, L. A.; Tekwani, B. L.; Huang, T. L. *Eur. J. Med. Chem.* **2004**, *39*, 547.
24. Mayence, A.; Vanden Eynde, J. J.; Krogstad, F. M.; Krogstad, D. J.; Cushion, M. T.; Huang, T. L. *J. Med. Chem.* **2004**, *47*, 2700.
25. Huang, T. L.; Vanden Eynde, J. J.; Mayence, A.; Donkor, I. O.; Khan, S. I.; Tekwani, B. L. *J. Pharm. Pharmacol.* **2006**, *58*, 1033.
26. Saiz-Urra, L.; Gonzalez, M. P.; Teijeira, M. *Bioorg. Med. Chem.* **2006**, *14*, 7347.
27. Gonzalez, M. P.; Suarez, P. L.; Fall, Y.; Gomez, G. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 5165.
28. Morales, A. H.; Cabrera Perez, M. A.; Gonzalez, M. P. *J. Mol. Model. (Online)* **2006**, *12*, 769.
29. Devillers, J.; Lipnick, R. L. In *Practical Applications of Quantitative Structure–Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*; Karcher, K., Devillers, J., Eds.; Kluwer: Dordrecht, 1990; p 129.
30. Brendle, J. J.; Outlaw, A.; Kumar, A.; Boykin, D. W.; Patrick, D. A.; Tidwell, R. R.; Werbovetz, K. A. *Antimicrob. Agents Chemother.* **2002**, *46*, 797.
31. Roberts, S. C.; Scott, J.; Gasteier, J. E.; Jiang, Y.; Brooks, B.; Jardim, A.; Carter, N. S.; Heby, O.; Ullman, B. *J. Biol. Chem.* **2002**, *277*, 5902.
32. Bray, P. G.; Barrett, M. P.; Ward, S. A.; de Koning, H. P. *Trends Parasitol.* **2003**, *19*, 232.
33. Atkinson, A. C. *Plots, Transformations and Regression*; Clarendon Press: Oxford, 1985.
34. Eriksson, L.; Jaworska, J.; Worth, A. P.; Cronin, M. T.; McDowell, R. M.; Gramatica, P. *Environ. Health Perspect.* **2003**, *111*, 1361.
35. Stewart, J.; Gill, L. *Econometrics*; Prentice Hall: London, 1998.
36. Derringer, G.; Suich, R. *J. Qual. Technol.* **1980**, *12*.
37. Tao, B.; Huang, T. L.; Zhang, Q.; Jackson, L.; Queener, S. F.; Donkor, I. O. *Eur. J. Med. Chem.* **1999**, *34*, 531.
38. Makler, M. T.; Ries, J. M.; Williams, J. A.; Bancroft, J. E.; Piper, R. C.; Gibbins, B. L.; Hinrichs, D. J. *Am. J. Trop. Med. Hyg.* **1993**, *48*, 739.
39. Frank, J.; Mopac 6.0; Seiler Research Laboratory, US Air Force Academy, Colorado Springs, CO, 1993.
40. Todeschini, R.; Consonni, V.; Pavan, M.; Dragon 2.1; Milano Chemometrics: Milano, 2002.
41. *Rational Approaches to Drug Design*; Todeschini, R., Ed.; Prous Science: Barcelona, Spain, 2001.
42. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
43. Leardi, R.; Boggia, R.; Terrile, M. *J. Chemom.* **1992**, *6*, 267.
44. Yasri, A.; Hartsough, D. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1218.
45. Hou, T. J.; Wang, J. M.; Liao, N.; Xu, X. J. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 775.
46. Hasegawa, K.; Kimura, T.; Funatsu, K. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 112.
47. Todeschini, R.; Consonni, V.; Pavan, M. Talete srl Milan, Italy, Software for Multilinear Regression Analysis and Variable Subset Selection by Genetic Algorithm, 2002
48. Randic, M. *New J. Chem.* **1991**, *15*, 517.
49. Randic, M. *J. Mol. Struct. (THEOCHEM)* **1991**, *233*, 45.
50. Randic, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311.
51. Klein, D. J.; Randic, M.; Babic, D.; Lucic, B.; Nikolic, S.; Trinajstic, N. *Int. J. Quantum Chem.* **1997**, *63*, 215.
52. Statistica 6.0; Statsoft Inc., 2001.