Journal of Molecular Graphics and Modelling 54 (2014) 19–31

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Modeling structure–activity relationships of prodiginines with antimalarial activity using GA/MLR and OPS/PLS

Luana Janaína de Campos, Eduardo Borges de Melo*

Theoretical Medicinal and Environmental Chemistry Laboratory (LQMAT), Department of Pharmacy, Western Paraná State University (UNIOESTE), Cascavel, PR, Brazil

ARTICLE INFO

Article history:
Accepted 25 August 2014
Available online 6 September 2014

Keywords:
Prodiginines
Malaria
QSAR
PLS
OPS
Genetic algorithm

ABSTRACT

In the present study, we performed a multivariate quantitative structure-activity relationship (QSAR) analysis of 52 prodiginines with antimalarial activity. Variable selection was based on the genetic algorithm (GA) and ordered predictor selection (OPS) approaches, and the models were built using the multiple linear regression (MLR) and partial least squares (PLS) regression methods. The leave-N-out crossvalidation and ***y***-randomization tests showed that the models were robust and free from chance correlation. The mechanistic interpretation of the results was supported by earlier findings. In addition, the comparison of our models with those previously described indicated that the OPS/PLS-based model had a higher quality of external prediction. Thus, this study provides a comprehensive approach to the evaluation of the antimalarial activity of prodiginines, which may be used as a support tool in designing new therapeutic agents for malaria.

1. Introduction

Malaria is a mosquito-borne infectious disease caused by parasitic protozoa, which is considered the most prevalent parasitic disease in the world. Approximately one-third of the world population lives in malaria-endemic areas. About 250 million people in more than 109 countries are affected, with 90% of deaths occurring in Africa (80% in sub-Saharan Africa). The disease has a great impact on the public health and financial situation in these countries; the economic loss in affected African countries is estimated to be US$ 12 billion every year, thus complicating therapeutic intervention. In addition, malaria may look unattractive to private pharmaceutical industry, mainly because of low purchasing power of the affected population, which emphasizes the necessity to develop cost-effective approaches to diagnose and treat malaria [1–9].

Malaria is caused by the parasites of *Plasmodium* spp. (*P. vivax*, *P. falciparum*, *P. ovale*, and *P. malariae*) transmitted primarily by mosquitoes of the genus *Anopheles*. *P. vivax* and *P. falciparum* are responsible for 80% of human cases [10,11]. The disease is characterized by intermittent fever occurring every 2 or 3 days, headache, body aches, anemia, jaundice, and swelling of the liver and spleen [12]. Treatment is very complex and often based on two or three different drugs used in combined mode [10,13]. The most common chemotherapy involves chloroquine and sulfadoxine–pyrimethamine. In the recent years, the derivatives of artemisinin, a natural product extracted from a Chinese plant *Artemisia annua* have been introduced and are currently the treatment of choice in sub-Saharan Africa [14].

The development of resistance to antimalarial agents, including artemisinin derivatives, is one of the main factors underlying the need for the development of new antimalarial drugs. This need is reinforced by inadequate pharmacokinetic properties, adverse effects, toxicity and high cost of current antimalarial agents [10,13–15]. In this context, prodiginines (Fig. 1) gained attention as natural products with antimalarial activity. Prodiginines are a class of red-pigmented secondary metabolites produced by actinomycetes and other eubacteria [16,17]. These compounds have been described to have multiple activities, including antibacterial, anticancer, and immunosuppressive effects. The antimalarial activity was first described by Gerber [18] and Papireddy et al. [16], who showed that prodiginines exhibited in vitro activity against *Plasmodium* species at lower concentrations than other agents did. The chemical structure of prodiginines has been a focus of attention because the first prodiginine derivative with potential therapeutic activity, 2-(2-((3,5-dimethyl-1*H*-pyrrol-2-yl)methylene)-3-methoxy-2*H*-pyrrol-5-yl)-1*H*-indole (GX15-070; Obatoclax$^{®}$) [20] (Fig. 1), is currently being tested in clinical trials for the treatment of lung cancer, leukemia, and other types of

* Corresponding author at: Department of Pharmacy, UNIOESTE, 2069 Universitária St., 85819110 Cascavel, PR, Brazil. Tel.: +55 45 3220 3256.
E-mail address: eduardo.b.de.melo@gmail.com (E.B. de Melo).

http://dx.doi.org/10.1016/j.jmgm.2014.08.004
1093-3263/© 2014 Elsevier Inc. All rights reserved.

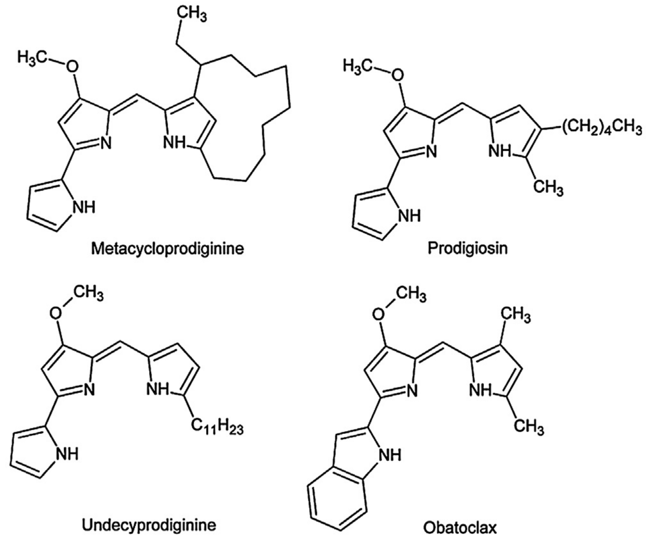


Fig. 1. Structures of naturally occurring prodiginines and commercial Obatoclax®.

malignancies [19], indicating that the compounds with this chemical structure may be safely used in clinical practice.

In this scenario, methods of quantitative structure–activity relationships (QSAR) should be useful tools for the development of new drugs. The approach is based on the assumption that the behavior of a set of structurally similar compounds in a biological system (in vitro or in vivo) can be quantitatively described by mathematical models, which can predict the activity of structural analogs not yet synthesized. The success of the QSAR methodology is a great assistance in reducing overhead costs, decreasing the time of obtaining positive results, reducing the use of laboratory animals as well as chemical and biological waste during drug development [21–25].

This study was aimed to obtain QSAR models with multiple linear regression (MLR) and partial least squares (PLS) based on a set of prodiginine derivatives described by Papireddy et al. as antimalarial agents [16]. The models were based on classical molecular descriptors and was constructed, with the aid of variable selection using genetic algorithms (GA) and ordered predictors selection (OPS), respectively [24,26,27].

2. Materials and methods

2.1. Softwares

Molecular modeling step was performed using the HyperChem software 7 [28] (structural design and optimization in molecular mechanics and semi empirical levels), Gaussian 09 [29] (optimizations in Hartree–Fock and density functional theory levels), Open Babel 2.3.1 [30] (conversion of file formats), Gauss View 5 [31] (visualization of structures and obtaining the electronic descriptors), Dragon 6 [32] (to obtain the other descriptors), QSARINS 1.1 [33] (variable selection by systematic search and genetic algorithm; regression using MLR; selection of the set of external validation; and validation of by MLR), QSAR Modeling [34] (variable selection by OPS; regression by PLS; internal validation of PLS models and robustness checks and random correlation for all the models), and Pirouette 4 [35] (for the refinement and external validation of PLS models). The r_m^2 metrics were obtained using the online server RmSquare Calculator (http://aptsoftware.co.in/rmsquare/). Additional test sets were obtained using the Dataset Division GUI 1.0. The validation was also performed with an “in-house” spreadsheet to calculate some statistical parameters of internal and external validation steps. The softwares QSARINS, QSAR Modeling and Dataset Division GUI 1.0 may be downloaded in http://www.qsar.it, http://lqta.iqm.unicamp.br, and http://dtclab.webs.com/software-tools, respectively.

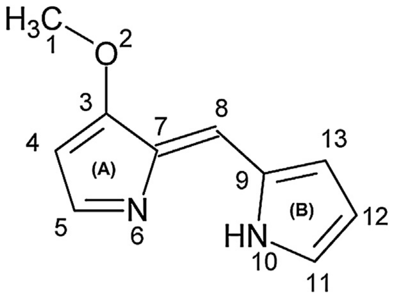


Fig. 2. The basic structure of novel prodiginine derivatives. Rings A and B are also identified.

2.2. Dataset

Papireddy et al. [16] synthesized 52 new prodiginines derivatives (Table 1) and tested their antimalarial activity against the chloroquine resistant D6 strain of *P. falciparum*. The antimalarial activity was measured as the concentration (nM) required to kill 50% of parasites (EC_{50}) using the methodology described by Smilkstein et al. [36] and Burgess et al. [37]. The observed EC_{50} values were converted to the corresponding $-\log EC_{50}$ (or pEC_{50}), resulting in vector ***y*** with a range of 4.34 logarithmic units (from 4.71 to 9.05). The dataset was divided into training (45 compounds) and test sets (compounds **26**, **28**, **36**, **41**, **49**, **60**, and **65**). The test set was randomly selected using the function of the QSARINS 1.1 [33], but it was verified a posteriori if the selected compounds represented adequately the pEC_{50} range as well as structural variations of the dataset. To ensure the quality of the external prediction, the original auxiliary models [38] was splitted in 14 different training and test sets, in a similar approach to that used recently that Kar et al. [39], using the Kennard–Stone and Euclidean Distance approaches [40]. The adopted identification code of each compound is the same used in the original reference [16].

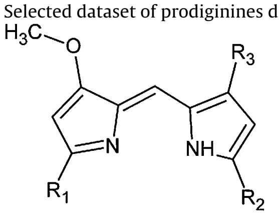
2.3. Molecular modeling

The dataset was built in the tautomeric form #1 (four possible), as described by Masand et al. [41], using as base the crystallographic structure of a synthetic prodiginine available in the support info. of García-Valverde et al. [42] (file jo301008c_si_003.cif). The basic structure of dataset is presented in Fig. 2. All molecules were initially optimized using molecular mechanics (MM+). In this step, the optimizations were alternated with cycles of molecular dynamics (1 ps, 300 K), until the energy obtained in MM+ did not vary more, indicating the obtention of a possible minimum energy structure. Next, simple optimizations were performed at Austin Model 1 (AM1), then in Hartree–Fock (HF/6-31G(d)), and finally density functional theory (DFT) level using the functional Becke, three-parameter, Lee–Yang–Parr (B3LYP), with the basis 6-311G(d,p). The optimization process was carried out in this sequence and steps to reduce the computation time required to obtain the optimized geometries at the level DFT/B3LYP, that was chosen because it has been reported to lead to satisfactory results when molecular geometries and energies are considered [43,44].

2.4. Molecular descriptors

Based on three-dimensional structures obtained by molecular modeling at DFT level, 29 electronic descriptors were obtained: Mulliken and Natural Bond Orders (NBO) partial charges for

Table 1
Selected dataset of prodiginines derivatives and their respective inhibition potencies against *P. falciparum* (D6 strain).



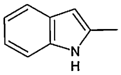
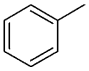
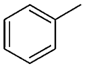
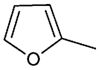
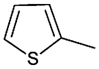
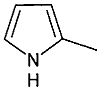
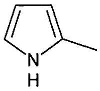
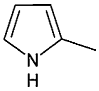
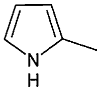
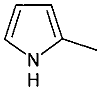
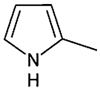
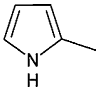
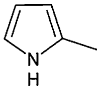
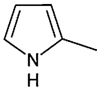
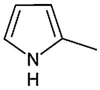
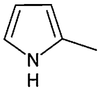
Compound[a]	R1	R2	R3	EC_{50} (nM) D6	pEC_{50}
18		CH_3	CH_3	4250	5.372
20		n-$C_{11}H_{23}$	H	10,470	4.980
21		CH_3	CH_3	19,410	4.712
22		n-$C_{11}H_{23}$	H	2920	5.535
24		n-$C_{11}H_{23}$	H	5940	5.226
26		n-C_3H_7	H	2300	5.638
27		n-C_4H_9	H	1780	5.750
28		n-C_6H_{13}	H	375	6.426
29		n-C_8H_{17}	H	80	7.097
30		n-$C_{16}H_{33}$	H	300	6.523
31		n-$C_{11}H_{22}NH_2$	H	1700	5.769
32		H	$(CH_2)_3COOCH_3$	4500	5.347
33		H	$CH_2CH(CH_3)_2$	460	6.337
34		H	n-C_4H_9	80	7.097
35		H	n-C_6H_{13}	28	7.553
36		H	n-C_8H_{17}	4.6	8.337

Table 1 (*Continued*)

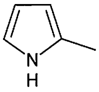
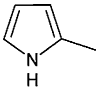
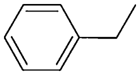
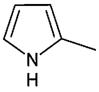
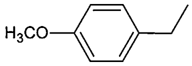
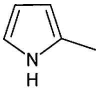
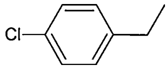
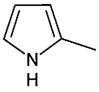
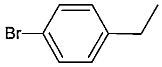
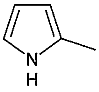
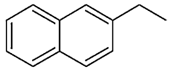
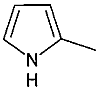
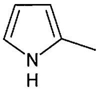
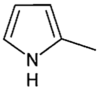
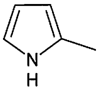
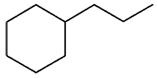
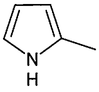
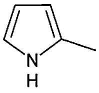
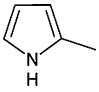
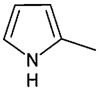
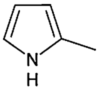
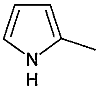
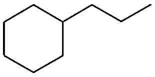
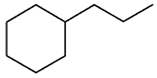
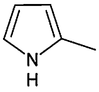
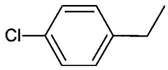
Compound[a]	R1	R2	R3	EC_{50} (nM) D6	pEC_{50}
37		H	n-$C_{10}H_{21}$	8.0	8.097
39		H		83	7.080
40		H		170	6.769
41		H		65	7.187
42		H		90	7.046
43		H		56	7.252
46		CH_3	CH_3	8900	5.051
47		n-C_6H_{13}	n-C_3H_7	4.5	8.347
48		n-C_8H_{17}	n-C_3H_7	2.9	8.538
49		n-C_3H_7		1.7	8.769
50		n-C_6H_{13}	n-C_6H_{13}	1.7	8.769
51		n-C_7H_{15}	n-C_6H_{13}	2.1	8.678
52		n-C_6H_{13}	n-C_8H_{17}	4.9	8.310
53		n-C_7H_{15}	n-C_8H_{17}	6.2	8.208
54		n-C_8H_{17}	n-C_8H_{17}	92	7.036
55				5.3	8.276
56		C_2H_5		6.3	8.201

Table 1 (*Continued*)

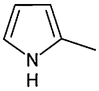
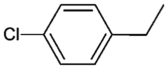
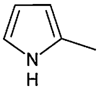
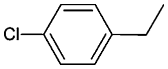
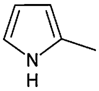
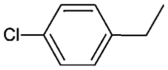
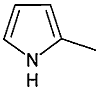
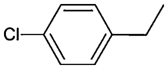
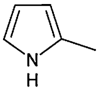
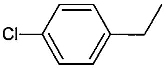
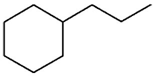
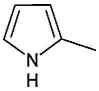
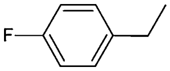
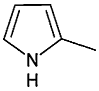
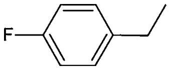
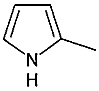
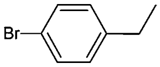
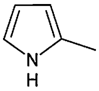
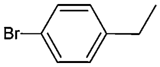
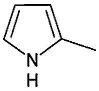
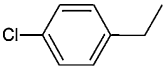
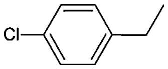
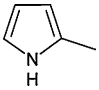
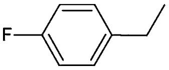
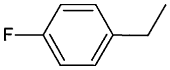
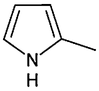
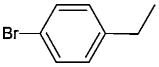
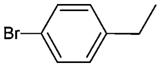
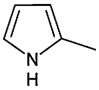
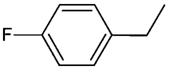
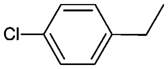
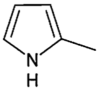
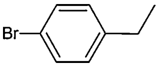
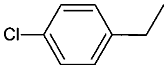
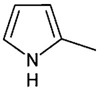
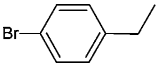
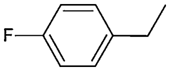
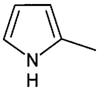
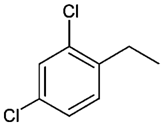
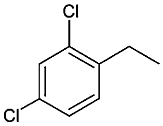
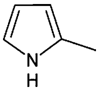
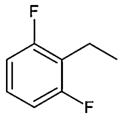
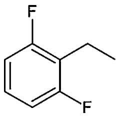
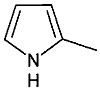
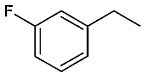
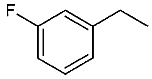
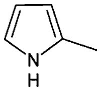
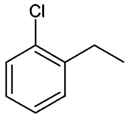
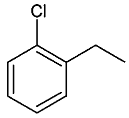
Compound[a]	R1	R2	R3	EC_{50} (nM) D6	pEC_{50}
57		n-C_3H_7		3.0	8.523
58		n-C_6H_{13}		2.0	8.700
59		n-C_7H_{15}		2.8	8.553
60		n-C_8H_{17}		16.0	7.796
61				3.9	8.409
62		n-C_6H_{13}		0.9	9.046
63		n-C_8H_{17}		1.3	8.886
64		n-C_6H_{13}		2.9	8.538
65		n-C_8H_{17}		4.0	8.398
66				6.1	8.215
67				5.6	8.252
68				14.0	7.854
69				6.1	8.215
70				8.3	8.081
71				5.7	8.244
72				12.6	7.900
73				14.7	7.833

Table 1 (*Continued*)

Compound[a]	R1	R2	R3	EC_{50} (nM) D6	pEC_{50}
74				5.1	8.292
75				3.6	8.444

[a] The identification code of each compound is the same used in the original reference.

prodiginine atoms in the basic structure (Fig. 2), molecular orbital energies (E_{HOMO-1}, E_{HOMO}, E_{LUMO}, and E_{LUMO+1}), dipole moment total (*D*) and the components in the axes *x*, *y*, and *z* (*Dx*, *Dy*, and *Dz*), (and total energy (ET). Moreover, using the equations described by Todeschini and Consonni [45] and the values for the energy of molecular orbital descriptors, the descriptors HOMO–LUMO energy gap (GAP), activation energy index (AEI), HOMO/LUMO energy fraction ($f_{H/L}$), and molecular electronegativity (χ) were derived, resulting in 33 electronic descriptors.

The low-energy structures obtained from molecular modeling stage were also used to obtain descriptors derived only from the information on atomic composition and connectivity (i.e., 0D, 1D, and 2D descriptors) and also the ones related to molecular geometry (3D descriptors). The obtained classes were: constitutional descriptors, ring descriptors, functional group counts and atom-centered fragments, topological descriptors, walk and path counts, connectivity indices, information indices, 2D matrix-based descriptors, 2D autocorrelations, edge adjacency indices, Burden eigen values, topological charge indices, eigen value-based indices, CATS2D finger prints, 2D atom pairs, P_VSA-like descriptors, ETA indices, atom-type E-state indices, geometrical descriptors, 3D matrix-based descriptors, 3D autocorrelations, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, Randic molecular profiles, and 3D atom pairs, molecular properties, and charge descriptors.

At the end, 4855 more descriptors were generated, totaling 4888. Considering the amount of generated descriptors, it was necessary to use approaches of variable reduction, which consists in the selection of a subset of variables able to preserve the essential information contained in the whole dataset, but eliminating redundancy [45]. Initially, the number was reduced by eliminating one member among pairs with highly correlated (greater than 0.9), constant and near-constant descriptors, and those with standard deviation less than 0.001. The resulting matrix underwent manual reduction in order to exclude descriptors with low variance not eliminated previously, resulting in 522 descriptors. The latter matrix was subjected to a final reduction step, which excluded the descriptors with an absolute Pearson correlation coefficient to biological activity ($|r|$) below 0.3, in order to eliminate descriptors with numerical variation, which poorly correlated with the vector $\boldsymbol{y}$. Thus, the matrix used in the selection of variables consisted of 238 descriptors for each derivative. There was no concern with the results obtained in relation to tautomeric form used, because previous studies with this dataset showed that for these molecules the QSAR results are independent of tautomerism [41].

2.5. Variable selection

Typically, the number of obtained descriptors in QSAR is much higher than the number of samples. Thus, it is necessary to use special methodology for the selection of those most contributing to the biological activity, to be included in the model [46]. The process of variable selection is based on finding combinations among the available variables capable of producing mathematical models that adequately describe the observed biological activity [24]. In this study, we employed the approaches of GA and OPS.

The principles of the selection of variables with GA are based on the theory of evolution, which is an example of an optimization process: highly complex species are evolved from the ancestor species with simpler structure. Initially, a set with *N* regression equations formed by two distinct variables (in this study) randomly selected among the available dataset is obtained. Each equation corresponds to an “individual” generation, while the variables correspond to “genes,” and the set of variables in the equation represent a “chromosome.” Each “individual” is characterized by its “chromosomes” and placed in descending order of fitness. In the “reproduction” process, new “individuals” are generated by “recombination” (crossover) and possible “mutations.” The process continues until a preset number of “generations” or copies of a “chromosome” are obtained [24,47].

The OPS method attaches importance to each descriptor according to a vector; then matrix descriptors are rearranged according to their relevancy so that the most important/relevant are represented by the first column of the matrix. The process is iterative: since this algorithm builds models using PLS, it is necessary to determine the maximum number of latent variables (LVs) that the user wishes to explore in order to obtain the models, and the number of LVs to be used for building the models. After this, the models are created by rearranging the data matrix of the descriptors, as outlined before. The models obtained in each OPS cycle are then organized according to statistical parameters generated in the process of model construction. The models are built with a reduced set of original variables and can be used either as the final model (which is not common) or as a reduced set of descriptors for further reiteration, until a combination of reduced descriptors and/or LVs with the statistical quality that meets the goals of the algorithm user are obtained [46].

During the process of variable selection, some statistical criteria should be used to classify the resulting models. In the equations obtained by GA, the coefficient of determination between the observed and predicted biological activities during the process of leave-one-out (LOO) cross-validation, Q^2_{LOO}, was used. When the models were generated using the OPS method, the best models had initially been classified by their *RMSECV* (root mean square error of cross-validation) to obtain a reduced set of descriptors that could lead to the models with low error. In the subsequent cycles, Q^2_{LOO} was used to obtain the models with better prediction.

2.6. Construction of models

After the selection of variables, the models were constructed using MLR to GA, and PLS to OPS. Since each method generated a number of models, the best of them were selected based on: (i) their respective statistical quality and (ii) the simplicity relative to the number of variables.

Table 2
Variable correlation matrix of Model 1.

Descriptor	SM11_AEA(dm)	R6u+	CATS2D_04_AL	MATS5v
SM11_AEA(dm)	1			
R6u+	−0.168	1		
CATS2D_04_AL	−0.220	0.257	1	
MATS5v	0.286	−0.278	−0.104	1

The advantage of MLR is its simple form and easily interpretable mathematical expression. Although utilized to great effect, MLR is vulnerable to descriptors that are correlated to one another, making it incapable of selecting the most significantly correlated sets [27]. In this step, after checking which models had the best Q^2_{LOO} values, the quality of the statistical parameters related to the external validation was tested.

On the other hand, the PLS regression method reduces the dimensionality of the data while retaining most of the variations in the dataset. In this process, the matrix of descriptors is correlated with the biological activity vector (or vector $\boldsymbol{y}$). Thus, the data are optimized for the estimation of $\boldsymbol{y}$ values, yielding orthogonal LVs, and the problem of collinearity is avoided [24,46,48].

In both the studies, it was necessary to perform data autoscaling. In this pre-processing approach, the value of each descriptor is subtracted by the correspondent average value (i.e., mean centering) and the result is divided by the standard deviation (i.e., scaling by variance). QSAR datasets consist of variables that differ in range, variation, and size. Consequently, autoscaling is usually applied prior to the regression analysis in QSAR [34].

2.7. Model validation

Validation constitutes a fundamental step in QSAR and is used to test the suitability of the model to perform reliable predictions of the modeled activity for a new derivative with an unknown response [45]. Validation is also recommended for the models describing a putative activity mechanism of the test compounds [24].

The statistical validation of QSAR models consists of internal and external validation [38,49–52]. In internal validation, explained variance by the model is measured using the R^2 (coefficient of determination), F-ratio test with a 95% confidence interval (α = 0.05), *RMSEC* (root mean square error of calibration) Q^2_{LOO}, *RMSECV*, and the scaled r^2_{m} (r^2_{m} (LOO)-scaled and average r^2_{m} (LOO)-scaled). Finally, in the internal validation step, it is also recommended to verify the robustness of the model and if the explained and predicted variances occur by chance. In this study, we used the leave-N-out cross-validation and the $\boldsymbol{y}$-randomization test, respectively, to evaluate these possibilities [38].

External validation is much more reliable to verify the predictive ability of the model than cross-validation, because it compares actual data of biological activities of compounds which has not been used in model construction with predicted activities [38]. To analyze the capacity for external prediction, R^2_{pred} (coefficient of determination of external validation), *RMSEP* (root mean square error of external validation), k and k' (Golbraikh–Tropsha slopes of the linear regression lines between the observed and the predicted activities in the external validation), $|R^2_0 - R'^2_0|$ (Golbraikh–Tropsha absolute values of the difference between the coefficients of multiple determination), and the scaled r^2_{m} metrics for external validation (delta r^2_{m}(pred)-scaled and average r^2_{m}(pred)-scaled) have been used [49–52]. These tests provide a thorough assessment of predictive QSAR models, thus ensuring maximum reliability, quality, and efficiency of the regression models for practical purposes [38].

Table 3
Contribution of descriptors in each latent variable in Model 2.

Descriptor	LV1	LV2
CATS2D_03_AL	−0.335	0.061
R6e+	−0.476	−0.468
F07[C—N]	0.570	0.378
SM11_AEA(dm)	0.471	−0.394
Q13NBO	0.337	−0.692

Finally, we evaluated the scaled r^2_{m} overall metrics, which analyzed the overall performance of the developed model based on predictions for both internal and external validation. As the r^2_{m} metrics are used for internal and external validation, the overall prediction ability of the model can be considered good if the average r^2_{m}(overall)-scaled is >0.5 and delta r^2_{m}(overall)-scaled is < 0.2 [51,52].

The mathematical definition of the statistical tools used in this study can be found in previous reports [38,45,50–52].

3. Results and discussion

3.1. QSAR results

Eqs. (1) and (2) correspond to the best models obtained in the GA/MLR and OPS/PLS steps, respectively. As Model 1 was built using MLR, the correlation between the descriptors was evaluated [27]. The highest value was obtained for the MATS5v and SM11_AEA (dm) descriptors (0.286) (Table 2), indicating that the correlations between the variables were acceptable. The OPS/PLS model consisted of five descriptors that generated two naturally orthogonal LVs (Table 3) [24], which accumulated 75.218% of information (LV1: 54.192%. LV2: 21.025%):

$$\begin{aligned}\text{pEC}_{50} = {} & -6.066 + 1.768 * (\text{SM11\_AEA(dm)}) - 88.467 * (\text{R6u+}) \\ & - 0.602 * (\text{CATS2D\_04\_AL}) + 3.171 * (\text{MATS5v})\end{aligned} \tag{1}$$

$$\begin{aligned}\text{pEC}_{50} = {} & 1.899 - 0.476 * (\text{CATS2D\_03\_AL}) - 70.067 * (\text{R6e+}) \\ & + 0.337 * (\text{F07[C—N]} + 0.690 * (\text{SM11\_AEA(dm)}) \\ & + 1.506 * (\text{Q13NBO})\end{aligned} \tag{2}$$

Table 4 presents information about the eight descriptors selected in the study. Only one descriptor (SM11_AEA(dm)) was selected for both models, although there were pairs of related descriptors: GEometry, Topology, Atom-Weights AssemblY (GETAWAY) R6u+ and R6e+, and the Chemically Advanced Template Search-2D CATS2D_03_AL and CATS2D_04_AL.

Statistical tests of internal validation (Table 5) evaluated the predictive ability of the models toward the activity of the compounds used in its construction, which should have a good degree of fit and significance [53]. R^2 is a measure of the data explained by the model; the acceptable minimum R^2 value for the regression models in QSAR is >0.6 (i.e., 60% of the explained variance) [38,49]. Furthermore, *RMSEC* assesses the variability not explained by the model, and thus should have the lowest possible value [38,49]. Model 2,

Table 4
Descriptors selected in the obtained models.[a]

Symbol	Descriptor	Class	Type	Model
SM11_AEA (dm)	Spectral moment of order 11 from augmented edge adjacency matrix weighted by dipole moment	Edge adjacency indices	2D	1 and 2
F07[C—N]	Frequency of C—N at topological distance 7	2D atom pairs	2D	2
CATS2D_04_AL	CATS2D acceptor-lipophilic at lag 04	Chemically Advanced Template Search-2D (CATS2D)	2D	1
CATS2D_03_AL	CATS2D acceptor-lipophilic at lag 03	CATS2D	2D	2
MATS5v	Moran autocorrelation of lag 5 weighted by van der Waals volume	2D autocorrelation	2D	1
R6u+	R maximal autocorrelation of lag 6/unweighted	GEometry, Topology, and Atom-Weights AssemblY (GETAWAY) descriptor	3D	1
R6e+	R maximal autocorrelation of lag 6/weighted by Sanderson electronegativity	GETAWAY descriptor	3D	2
Q13NBO	Partial charge of atom 13 obtained by Natural Bond Orders (NBO) theory	Electronic	3D	2

[a] Values of the descriptors for all compounds are available in supplementary material.

Table 5
Statistical comparison of Models 1 and 2.

Parameter	Model 1	Model 2
Number of compounds	45	42
Number of descriptors	4	4
Number of LV	–	2
Cumulated information	–	75.218%
Outlier	0	3
R^2	0.869	0.918
RMSEC	0.451	0.364
F (tabulated value)	66.180 (2.606)	218.472 (2.612)
Q^2_{LOO}	0.821	0.894
RMSECV	0.526	0.397
Average r^2_{m}(LOO)-scaled	0.750	0.850
Delta r^2_{m}(LOO)-scaled	0.104	0.085
R^2-Q^2_{LOO}	0.048	0.024
R^2_{pred}	0.830	0.833
RMSEP	0.440	0.435
Average r^2_{m}(pred)-scaled	0.759	0.813
Delta r^2_{m}(pred)-scaled	0.099	0.088
k	1.033	1.037
k′	0.966	0.963
$\lvert R^2_0 - R'^2_0 \rvert$	0.014	0.002
Average r^2_{m}(overall)-scaled	0.732	0.842
Delta r^2_{m}(overall)-scaled	0.002	0.055

which explain the largest amount of data ($R^2 = 0.905$, 90.5%) with lower *RMSEC*, is thus the one with the smaller amount of unexplainable information.

The *F*-test evaluates the ratio between the variability explained by the model and the variability which remains unexplained, and thus should be maximized in relation to values found in specific tables [34,53]. The result indicates that both models were significant, and the high value obtained for Model 2 reinforces the hypothesis that there is a relationship between variables and biological activity.

The internal predictability of the model was tested by leave-one-out (LOO) cross-validation. In the process, the Q^2_{LOO} (which should have a minimum acceptable value of 0.5 indicating 50% prediction of variability) was calculated [38,49]. In LOO cross-validation, the prediction statistics are also expressed by *RMSECV*, which, similar to *RMSEC*, should have the lowest possible value [38], and by r^2_{m}(LOO)-scaled parameters, which penalize the model more strictly than Q^2_{LOO} [52]. The results were consistent with the proposed limits (average r^2_{m}(overall)-scaled > 0.5 and delta r^2_{m}(overall)-scaled < 0.2). The predicted pEC_{50} values are available in the supplementary material.

The difference between R^2 and Q^2_{LOO} should be less than 0.3, because higher values may indicative overfitting of the model. Some authors recommend even lower limits, most commonly 0.1 [54,55]. Model 2 also had the lowest difference, but according to the LOO test, both models had good predictive capacity and, most likely, no overfitting.

Despite the quality of internal statistics, it is necessary to check the models for robustness and if the explained and predicted informations is not due to spurious correlation. The robustness, defined as the ability of the model to resist small and deliberate changes, was evaluated by leave-N-out (LNO) crossvalidation. This process, based on the same principles as that of LOO, in that it removes a number of samples, builds a new model, and predicts the biological activity of the removed samples. In LNO cross-validation, the robust models should exhibit small differences in the coefficient of determination (Q^2_{LNO}) between the observed and predicted biological activity for *N* samples (typically 25–30% of the samples used to build the model) [34]. In this study, *N* between 1 and 14 was used. Both models were considered robust (Fig. 3), because they retained their Q^2_{LNO} values with minor variations and within the limit allowed. The smallest difference between Q^2_{LOO} and average Q^2_{LNO} was observed for Model 1 (0.003 against 0.005 of Model 2), while Model 2 was shown to be more stable (average standard deviation of 0.009, against 0.015 of Model 1).

The ***y***-randomization test is performed in parallel models to detect and quantify chance correlations between the randomized vector values of biological activity and unchanged original descriptors. To quantify the chance correlation, we used the approach based on $|r|$ between the original and randomized vectors ***y*** [21]. At the end, the values of intercepts should be less than 0.3 for the regression based on R^2 and 0.05 for the regression based on Q^2_{LOO}. The models showed results within these limits (Fig. 4), demonstrating that the correlation between the selected descriptors and biological activity was real.

Once internally validated, the model was used to predict pEC_{50} of the tested compounds. Whereas the prediction of biological activities for novel compounds is the most important goal in QSAR studies, many authors have argued that the models should be externally validated to be considered suitable for this purpose [22,45,56–58]. Our results show that the model had an adequate external predictability (Table 5). The values of *k*, *k*′, and $|R^2_0 - R'^2_0|$ were within the acceptable ranges (0.85–1.15 for *k* and *k*′, and $|R^2_0 - R'^2_0| < 0.3$) [50]. The *RMSEP* value also indicated a low probability of errors for the prediction of new derivatives [38]. The calculation of average r^2_{m}(pred)-scaled and delta r^2_{m}(pred)-scaled values confirmed the predictive power of the model. These data suggest that all the tested models can be used for the prediction of antimalarial activity in potential for new prodiginine analogs. The

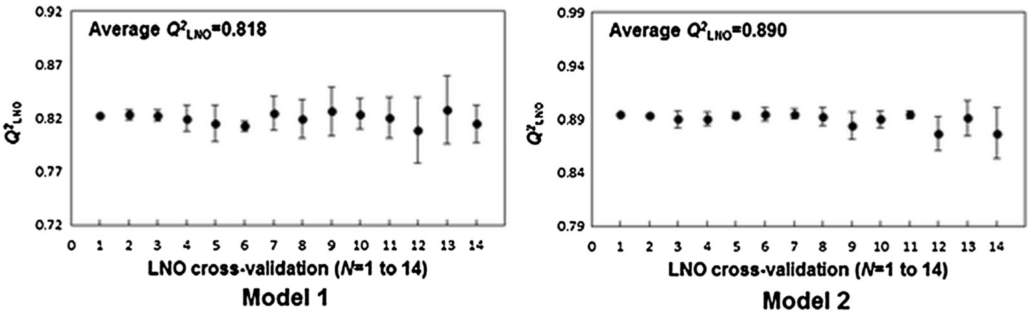


Fig. 3. LNO crossvalidation test. In the plots, each point refers to the average value of six tests and bars indicate standard deviation.

predicted pEC_{50} values of the test set are available in the supplementary material.

To confirm the quality of external prediction model, it was verified that the statistical parameters related to external validation of 14 different test sets generated using the Kennard–Stone and Euclidean Distance (seven sets per method) approaches as well as the autoscaled coefficients of the different models generated by the corresponding training sets (i.e., the importance of each descriptor in the model) would present an acceptable range of variation. The results of this study can be found in the supplementary material, Tables S4–S7. For Model 1, the coefficient values showed little variation and the results of external statistical parameters proved equivalents to those obtained for the original test set, which is demonstrated by the average values of these parameters. Just a test set (compounds **24**, **33**, **41**, **58**, **61**, **67**, and **70**) presented a parameter ($|R^2_0 - R'^2_0|$) slightly above the limit. For Model 2, all parameters in all test sets were approved. These results show that the quality of external prediction is reliable, particularly for Model 2.

Finally, the r^2_m(overall)-scaled metrics, which are based on the prediction of a comparably large number of compounds, can be more reliable for prediction purposes. These parameters can also be used for the selection of the best predictive models among a set of comparable models [54]. The obtained results (Table 5) confirmed the predictive quality of the models.

The presence of outliers (i.e., compounds with an atypical value that does not correspond to the distribution of the other values in the dataset) was verified using the studentized residuals (σ) versus the leverage (h) plot [59–61]. The maximal accepted deviation was $2.5 \times \sigma$. It was also verified if the compounds had a h higher than an established cut-off (in this work, $h^* = 2^*(p/n)$, more restrictive than the commonly used.

3.2. Mechanistic interpretation

In Model 1, the order of the descriptor importance based on autoscaled coefficients was: SM11_AEA(dm) (0.572) > R6u+ > (−0.448) CATS2D_04_AL (−0.249) > MATS5v (0.101), and in Model 2 the order is F07[C—N] (0.407) > R6e+ (−0.367) > SM11_AEA(dm) (0.208) > CATS2D_03_AL (−0.188) > Q13NBO (0.074). The most important information that could be extracted was the selection of the 2D frequency fingerprint descriptor F07[C—N] in Model 2. This descriptor stands out because it has already been selected by Mahajan et al. [62] in another GA/MLR study with the same dataset and reinforced by the Unrestricted Structure Activity Relationships General (GUSAR) analytical approach [63]. It could be observed that the four most active compounds (**49**, **50**, **62**, and **63**) had nitrogen in the basic structure in a topological distance of seven chemical bonds in relation to carbon atoms in ring B, both for R1 and R2. Fig. 5 shows a comparison between the compounds **49** and three of the less active (**21**, **46**, **32**), demonstrating the importance of the substituents on R1 and R2 in defining these paths. These results make a correlation between the selection of this descriptor and antimalarial activity quite plausible.

Mahajan et al. [62] have proposed that lipophilicity is important for the antimalarial activity of prodiginines. This feature can be observed in CATS2D_03_AL and CATS2D_04_AL. The CATS2D descriptors were very similar to 2D atom pairs descriptors, but they assigned atoms to defined pharmacophore points. Both descriptors defined the same pharmacophore points: hydrogen-bond acceptor (A) and lipophilic (L) at topological distances of 3 and 4 bonds between the points, and are probably also related to existing nitrogen atoms in the rings (the hydrogen bond-acceptor point) and hydrophobic side chains. The importance of this feature was also previously described in a pharmacophoric modeling study carried out by Singh et al. [63] (Fig. 6), and the distance between the points are equivalent to the encoded topological distances in the descriptors. But the interpretation of these descriptors in QSAR models can be considered complex, since both had negative coefficients, indicating that these groups at these topological distances tended to reduce the antimalarial potency of the molecule. Some compounds that had acceptor groups in the lateral chain (as **32** and **40**) are moderately or least actives, and for this reason have higher values for

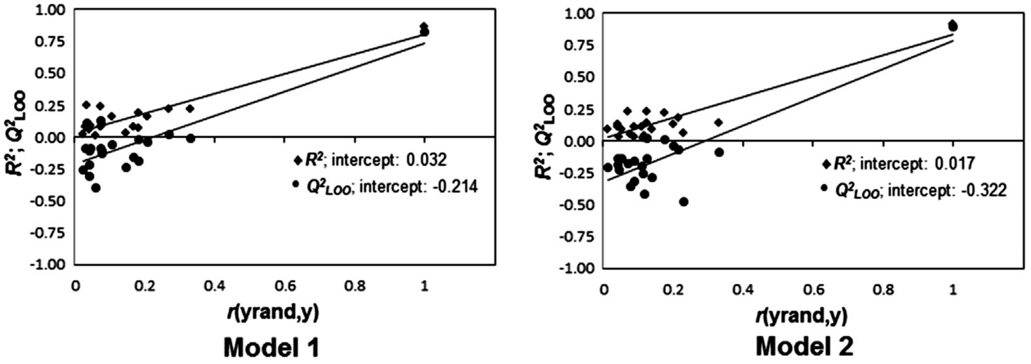


Fig. 4. Plots of $\boldsymbol{y}$-randomization test. In the plots, r ($\boldsymbol{y}_{\text{rand}}$, $\boldsymbol{y}$) refers to the absolute values of the Pearson correlation coefficient (r) between the randomized ($\boldsymbol{y}_{\text{rand}}$) and original ($\boldsymbol{y}$) vectors $\boldsymbol{y}$.

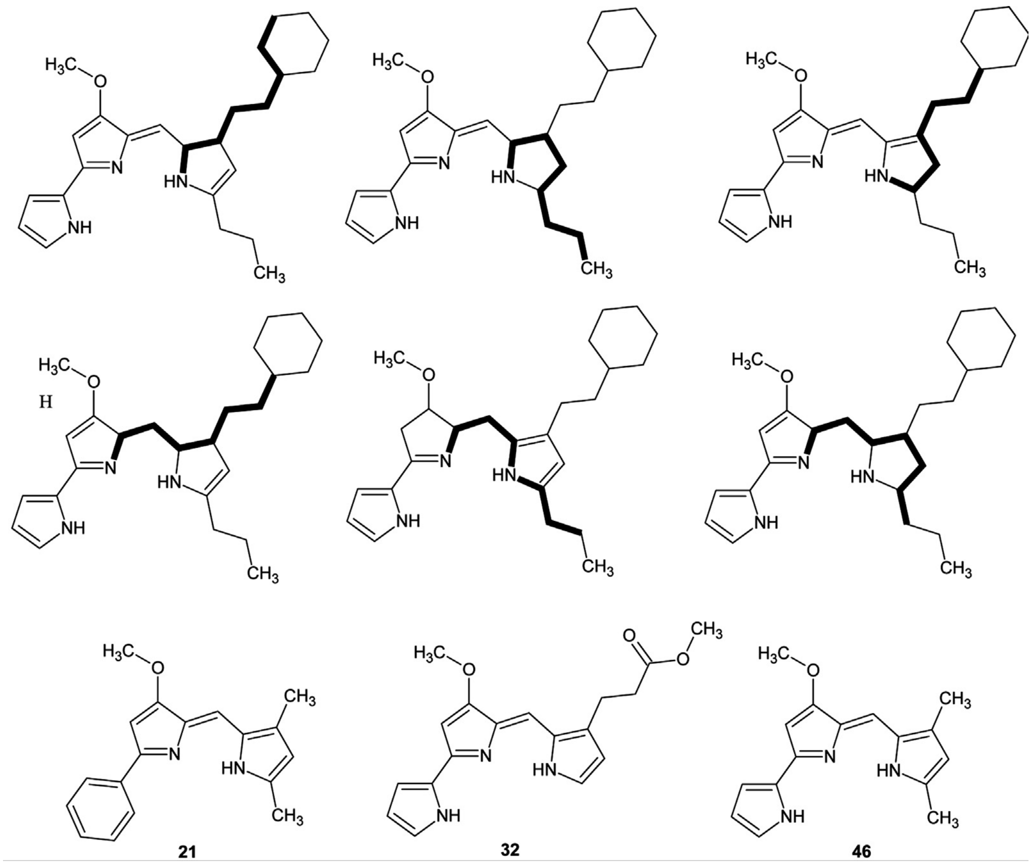


Fig. 5. Comparison of the structure of compound **46**, one of the most potent compounds of the dataset, with three of the least active compounds (**21**, **32**, **46**), highlighting the topological distances of **46** defined by R1 and R2.

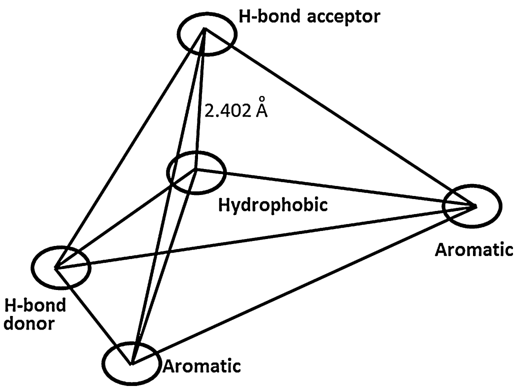


Fig. 6. Scheme based on pharmacophore model published by Singh et al. [63] to prodigininas active against *P. falciparum* strain D6. Only the distance between the H-bond acceptor and hydrophobic feature is presented.

these descriptors (Table S3). It is possible that both are representing this unfavorable structural feature for activity. Another possibility may be related with the proposition suggested by Masand et al. [64], wherein the activity would be favored by small alkyl chain in the R2 position. An equivalent suggestion can be seen in the results obtained by Singh et al. [63,65], who proposed that the negative coefficient of Wiener Topological Index (W) means that smaller carbon chains are favorable to activity [65]. Some of the less active compounds (**20**, **22** and **24**) really have long side chains in R2.

In this study, the electronic descriptor Q13NBO was selected despite its minor importance indicated by the autoscaled coefficient (0.074), which is interesting, considering that none of the previous studies [62–66] used electronic descriptors. Quantum-mechanical descriptors are usually easy to interpret when compared to geometric or topological descriptors [67–70], which can lead to potentially important information about the mechanism of action of a dataset. The selected descriptor is a partial atomic charge of carbon 13 in ring B of the basic prodiginine structure (Fig. 2), has been obtained using the Natural Bond Orders theory (NBO) [71]. One of the uses of electronic descriptors is describe electronic aspects of whole molecules and particular regions such as atoms and molecular fragments. Electrostatic forces in specific molecular regions may be important in the interaction between drugs and their receptors. With C13 atom as a substitution point, it is possible that the influence of Q13NBO on the biological activity of prodiginine analogs is related to the effect of the different substituents positioned at this point of the basic structure. According to the second model, charge increase at this position correlated with the increase in antimalarial activity. It is interesting to note that compounds with electronegative elements in substituents attached to this position demonstrated an increase of atom 13 charge (i.e., reduction of the electron density). The most active compounds of the dataset (**62** and **63**) have a halogen atom in the substituent at C13, suggesting that electron-withdrawing groups in the substituent improve the activity. This possibility is supported by the results of Masand et al. [64], but is at variance with the findings of Mahajan et al. [62], who reported that electron-withdrawing groups had insignificant impact on the activity of synthetic prodiginines. But is possible that this descriptor is correlated to some underlying factor. This structural feature has not been sufficiently investigated, but the results presented here reinforce the hypothesis of the influence of these characteristics on the biological activity of prodiginines.

Table 6
Statistics of previously published models.

Reference	Models	Method	R^2	r	*RMSEC*	*F*	Q^2_{LOO}	*RMSECV*	R^2_{pred}
[62]	A	GA/MLR	0.920	–	0.560	48.130	0.800	0.590	0.710
	B	GUSAR	0.800	–	0.350	16.000	0.760	–	0.860
[64]	C	GA/MLR	0.924	–	0.372	87.960	0.901	0.572	0.798
	D	GUSAR	0.940	–	0.321	54.894	0.912	0.299	0.948
[66]	E	COMSIA/PLS	0.911	–	0.390	122.117	0.738	–	–
[63]	F	CP/MLR	–	0.910	0.592	64.099	0.790	–	0.616
	G	CP/MLR	–	0.921	0.561	54.739	0.809	–	0.668
[65]	H	MRA	0.778	0.882	0.684	35.110	0.616	–	0.608
	I	3D-QSAR PHASE	0.993	–	0.138	530.000	–	–	0.655
	J	3D-QSAR PHASE	0.994	–	0.122	678.000	–	–	0.451

The low value of the autoscaled coefficient would indicate that this descriptor could be removed from Model 2, but the external validation step for the resulting model (not shown) downgraded the results ($R^2_{\text{pred}} = 0.720$, average $r^2_{\text{m}}(\text{pred}) = 0.712$ and delta $r^2_{\text{m}}(\text{pred}) = 0.145$). This change in results may be related to the contribution of this descriptor (Table 3) in the second LV of this model, which accumulates 21.025% of the model information, and may influence its ability for external prediction.

Moreover, the selection of the GETAWAY R6e+ descriptor weighted by electronegativity can emphasize the relevance of electronic effects to the antimalarial activity of prodiginines. In Model 1, statistically inferior to Model 2, the corresponding unweighted descriptor was selected, which may indicate that in addition to the geometric information encoded by this descriptor, electronic properties are actually relevant. The importance of the electronic effect can also be enhanced by the descriptor SM11_AEA(dm). This edge adjacency index, a type of a topological descriptor, is weighted by the dipole moment of specific bond types and uses predetermined values for such bonds as carbon—halogen [72]. The selection of this descriptor is supported by the fact that the majority of the most active compounds has halogen in the side chain and shows the highest values for this descriptor. The importance of this effect is further emphasized by its selection in both models. Singh et al. [63] reported that, despite not having used electronic descriptors, they also obtained QSAR models where the importance of the electronic features can be observed through geometrical descriptors where the electronegativity is a weighting factor, and also by the selection of topological charge indices [73].

The influence of steric factors on prodiginine activity can be defined by the descriptor MATS5v that corresponds to 2D autocorrelation between atom pairs within the topological distance of 5. This descriptor is calculated by summing products of terms, including the atomic weight (in this case, van der Waals volume), for the terminal atoms in the topological distance. The atomic property is directly proportional to the descriptor [45]; therefore, if the atoms have high van der Waals volume, the value of the descriptor will be also high. The selection of this descriptor is supported by the study of Masand et al. [64], where the descriptor Broto–Moreau autocorrelation of lag 7/weighted by atomic van der Waals volume (ATS5v) and a related 2D autocorrelation descriptor have been selected. The positive coefficient of both descriptors demonstrated in the previous as well as our study indicates that the compounds with higher values tend to be more potent. In this study, MATS5v was the least important descriptor in Model 1, suggesting that, similar to charge (Q13NBO), the molecular size is moderately favorable for the antimalarial activity.

The disadvantage of Model 2 is that three compounds (**20**, **29**, and **54**) were classified as outliers. In LOO crossvalidation, all the three compounds showed a difference in residue prediction greater than one unit, which led to σ of −2.436, 2.530, and −2.975, and h of 0.129, 0.115 and 0.020, respectively. The h^* for Model 2 is 0.095, and thus the compounds **20** and **29** were removed. The compound **54** was removed due to your higher σ. In the Model 1,**54** also presented a higher σ (−2.890), but is ths case this sample is more distant of the h^* line (0.178), and thus it is less influential on the quality of the model. Therefore, this sample was kept in the Model 1, thereby ensuring that it presents a better structural representation of the dataset under study. No structural justification was detected for these behaviors. However, according to Gramatica [74], the presence of compounds with high studentized residuals and leverages below the limit value (h^*) may indicate a failure in the experimental assessment of biological activity.

3.3. Comparison with previous models

The comparison of our models with those previously published (Table 6) showed that Models 1 and 2 had good statistical quality. These models are stronger in cross-validation than models A, B, E, F, G and H, but weaker than C and D. This test was not carried out for models I and J. R^2_{pred} for Model 2 was greater than that of other models (including Model 1), except for models B (only 0.027 units) and D (this test was not carried out for model E). However, it is important to remember that the models are not standardized, and therefore, the algorithms and validation procedures are very different. Thus, the comparison between the statistical parameters is insufficient to guarantee the predictive superiority of a particular model. The greatest advantage of Models 1 and 2 in relation to previous models is that they were subjected to and approved in several statistical tests currently recommended for the validation of QSAR, which increases their reliability in predicting the antimalarial activity of new prodiginine derivatives. The high statistical significance of Model 2 (Table 5) (about 68 times higher than its tabulated critical value) indicates that it is the most suitable model for the prediction of antimalarial activity of new prodiginine analogs.

4. Conclusion

In this study, we validated two multivariate QSAR models using a set of 52 structural derivatives of antimalarial prodiginines. All the models were validated both internally and externally, indicating that they were significant, robust, had no random correlation, and possessed good predictive ability. Model 2, obtained using the OPS/PLS approach, is the one that performed the best, despite having three outliers. Furthermore, most of the descriptors were interpretable and were corroborated by the results of previous studies, which used the same dataset. A special significance of the descriptor F07[C—N] indicates that this topological feature is relevant to the antimalarial activity of prodiginines. The selection of the electronic descriptor Q13NBO may indicate that, despite lower relevance, the electronic distribution is also relevant to the antimalarial activity. Finally, the comparison with other published models demonstrates that our Model 2 provides the quality of external prediction superior to that of most other models, with the exception of a GUSAR model. The comparison, however, has been hampered by inter-study differences in statistical tools, types of descriptors used, and mathematical approaches. In this study, we used a variety of stringent validation tools, which ensure high reliability of the presented models as support tools for the design of new lead compounds with antimalarial activity.

Acknowledgements

Financial support: Fundação Araucária (www.fappr.pr.gov.br, grant 2010/7354); Scientific initiation scholarship: UNIOESTE/Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Suggestions: Dr. João Paulo Ataíde Martins (IESB: Instituto de Educação Superior de Brasília), and Dr. Márcia Miguel Castro Ferreira (UNICAMP: Universidade Estadual de Campinas).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jmgm.2014.08.004.

References

[1] M.A. Biamonte, J. Wanner, K.G. Le Roch, Recent advances in malaria drug discovery, Bioorg. Med. Chem. Lett. 23 (2013) 2829–2843.

[2] A.P. Costa, C.S. Bressan, R.S. Pedro, R. Valls-de-Souza, S. Silva, P.R. Souza, L. Guaraldo, M.F. Ferreira-da-Cruz, C.T. Daniel-Ribeiro, P. Brasil, Diagnóstico tardio de malária em área endêmica de dengue na extra-Amazônia Brasileira: experiência recente de uma unidade sentinela no estado do Rio de Janeiro, Rev. Soc. Bras. Med. Trop. 43 (2010) 571–574.

[3] G. Pasvol, Management of severe malaria: interventions and controversies, Infect. Dis. Clin. North Am. 19 (2005) 211–240.

[4] V.D. Bastos, Laboratórios farmacêuticos oficiais e doenças negligenciadas: perspectivas de política pública, Revista do BNDS 13 (2006) 269–329.

[5] J.N.M. Boechat, Era uma vez doenças negligenciadas, Rev. Virtual Quím. 4 (2012) 195–196.

[6] J.A. Lindoso, A.A. Lindoso, Neglected tropical diseases in Brazil, Rev. Inst. Med. Trop. São Paulo 51 (2009) 247–253.

[7] Ministério da Saúde, Doenças negligenciadas: estratégias do Ministério da Saúde, Rev. Saúde Públ. 44 (2010) 200–202.

[8] F. Pontes, Doenças negligenciadas ainda matam 1 milhão por ano no mundo, Rev. Inovação em Pauta 6 (2009) 69–73.

[9] G.L. Werneck, M.H. Hasselmann, T.G. Gouvea, An overview of studies on nutrition and neglected diseases in Brazil, Cien. Saude Colet. 16 (2011) 39–62.

[10] L.S. Garcia, Malaria, Clin. Lab. Med. 30 (2010) 93–129.

[11] K.N. Suh, K.C. Kain, J.S. Keystone, Malaria, Can. Med. Assoc. J. 170 (2004) 1693–1702.

[12] T.C.C. França, M.G. Santos, J.D. Figueroa-Villar, Malária: aspectos históricos e quimioterapia, Quím. Nova 31 (2008) 1271–1278.

[13] K. Na-Bangchang, J. Karbwang, Current status of malaria chemotherapy and the role of pharmacology in antimalarial drug research and development, Fundam. Clin. Pharmacol. 23 (2009) 387–409.

[14] B. Mordmuller, New medicines for malaria, Wien. Klin. Wochenschr. 122 (2010) 19–22.

[15] K. Kaur, M. Jain, T. Kaur, R. Jain, Antimalarials from nature, Bioorg. Med. Chem. 17 (2009) 3229–3256.

[16] K. Papireddy, M. Smilkstein, J.X. Kelly, Shweta, S.M. Salem, M. Alhamadsheh, S.W. Haynes, G.L. Challis, K.A. Reynolds, Antimalarial activity of natural and synthetic prodiginines, J. Med. Chem. 54 (2011) 5296–5306.

[17] N.R. Williamson, P.C. Fineran, F.J. Leeper, G.P.C. Salmond, The biosynthesis and regulation of bacterial prodiginines, Nat. Rev. Microbiol. 4 (2006) 887–899.

[18] N.N. Gerber, A new prodiginine (prodigiosin-like) pigment from streptomyces. Antimalarial activity of several prodiginines, J. Antibiot. 28 (1975) 194–199.

[19] ClinicalTrials.go, Obatoclax. http://clinicaltrials.gov/ct2/results?intr=Obatoclax (accessed 17.01.14).

[20] N.R. Williamson, P.C. Fineran, T. Gristwood, S.R. Chawrai, F.J. Leeper, G.P. Salmond, Anticancer and immunosuppressive properties of bacterial prodiginines, Future Microbiol. 2 (2007) 605–618.

[21] L. Eriksson, J. Jaworska, A.P. Worth, M.T. Cronin, R.M. McDowell, P. Gramatica, Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs, Environ. Health Perspect. 111 (2003) 1361–1375.

[22] Organization for Economic Co-Operation and Development, Guidance Document on the Validation of (Quantitative) Structure–Activity Relationship [(Q)SAR] Models. March 2007. http://search.oecd.org/officialdocuments/displaydocumentpdf/?doclanguage=en&cote=env/jm/mono%282007%292 (accessed 17.01.14).

[23] R. Dayam, N. Neamati, Small-molecule HIV-1 integrase inhibitors: the 2001–2002 update, Curr. Pharm. Des. 9 (2003) 1789–1802.

[24] M.M.C. Ferreira, C.A. Montanari, A.C. Gaudio, Seleção de variáveis em QSAR, Quím. Nova 25 (2002) 439–448.

[25] L.C. Tavares, QSAR: a abordagem de Hansch, Quím. Nova 27 (2004) 631–639.

[26] E.B. de Melo, Modeling physical and toxicity endpoints of alkyl (1-phenylsulfonyl) cycloalkane-carboxylates using the Ordered Predictors Selection (OPS) for variable selection and descriptors derived with SMILES, Chemom. Intell. Lab. Syst. 118 (2012) 79–87.

[27] P. Liu, W. Long, Current mathematical methods used in QSAR/QSPR studies, Int. J. Mol. Sci. 10 (2009) 1978–1998.

[28] HyperChem, version 7. Hypercube Inc., USA, 2002. http://www.hyper.com/

[29] Gaussian, version 09, Gaussian Inc., Wallingford, USA, 2009. http://www.gaussian.com/g_prod/g09.htm

[30] The Open Babel Package, version 2.3.1, 2011. http://openbabel.org

[31] GaussView, version 5. Semichem Inc., Shawnee Mission, USA, 2009. http://www.gaussian.com/g_prod/gv5.htm

[32] Dragon 6, TALETE srl, Milano, Italy, 1997. http://www.talete.mi.it/products/dragon_description.htm

[33] P. Gramatica, N. Chirico, E. Papa, S. Cassani, S. Kovarich, QSARINS: a new software for the development, analysis, and validation of QSAR MLR models, J. Comp. Chem. 34 (2014) 2121–2132.

[34] J.P.A. Martins, M.M.C. Ferreira, QSAR modeling: a new open source computational package to generate and validate QSAR models, Quím. Nova 36 (2013) 554–560.

[35] Pirouette, version 4. Infometrix Inc., Woodinville, USA, 2011. http://www.infometrix.com/software/pirouette.html

[36] M. Smilkstein, N. Sriwilaijaroen, J.X. Kelly, P. Wilairat, M. Riscoe, Simple and inexpensive fluorescence-based technique for high-throughput antimalarial drug screening, Antimicrob. Agents Chemother. 48 (2004) 1803–1806.

[37] S.J. Burgess, A. Selzer, J.X. Kelly, M.J. Smilkstein, M.K. Riscoe, D.H. Peyton, A chloroquine-like molecule designed to reverse resistance in *Plasmodium falciparum*, J. Med. Chem. 49 (2006) 5623–5625.

[38] R. Kiralj, M.M.C. Ferreira, Basic validation procedures for regression models in QSAR and QSPR studies: theory and application, J. Braz. Chem. Soc. 20 (2009) 770–787.

[39] S. Kar, A. Gajewicz, T. Puzyn, K. Roy, J. Leszczynski, Periodic table-based descriptors to encode cytotoxicity profile of metal oxide nanoparticles: a mechanistic QSTR approach, Ecotoxicol. Environ. Safe. 107 (2014) 162–169.

[40] T.M. Martin, P. Harten, D.M. Young, E.N. Muratov, A. Golbraikh, H. Zhu, A. Tropsha, Does rational selection of training and test sets improve the outcome of QSAR modeling? J. Chem. Inf. Model. 52 (2012) 2570–2578.

[41] V.H. Masand, D.T. Mahajan, T.B. Hadda, R.D. Jawarkar, A.M. Alafeefy, V. Rastija, M.A. Ali, Does tautomerism influence the outcome of QSAR modeling? Med. Chem. Res. 23 (2014) 1742–1757.

[42] M. Garcia-Valverde, I. Alfonso, D. Quinonero, R. Quesada, Conformational analysis of a model synthetic prodiginine, J. Org. Chem. 77 (2012) 6538–6544.

[43] F.A. Molfetta, A.T. Bruni, F.P. Rosselli, A.B.F. da Silva, A partial least squares and principal component regression study of quinone compounds with trypanocidal activity, Struct. Chem. 18 (2007) 49–57.

[44] J. Lameira, I.G. Medeiros, M. Reis, A.S. Santos, C.N. Alves, Structure–activity relationship study of flavone compounds with anti-HIV-1 integrase activity: a density functional theory study, Bioorg. Med. Chem. 14 (2006) 7105–7112.

[45] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 2009.

[46] R.F. Teófilo, J.P. Martins, M.M.C. Ferreira, Sorting variables by using informative vectors as a strategy for feature selection in multivariate regression, J. Chemom. 23 (2009) 32–48.

[47] K.M.G. Oliveira, Estudos QSAR de compostos com atividade leishmanicida, Universidade Estadual de Campinas, Campinas, 2009, http://biq.iqm.unicamp.br/arquivos/teses/000469855.pdf (accessed 17.01.2014).

[48] L. Eriksson, L.P. Andersson, E. Johansson, M. Tysklind, Megavariate QSAR analysis of environmental data. Part I—the basic framework founded on main component analysis (PCA), partial least squares (PLS), and statistical molecular design (SMD), Mol. Divers. 10 (2006) 169–186.

[49] A. Golbraikh, A. Tropsha, Beware of q2! J, Mol. Graph. Model. 20 (2002) 269–276.

[50] A. Golbraikh, M. Shen, Z.Y. Xiao, Y.D. Xiao, K.H. Lee, A. Tropsha, Rational selection of training and test sets for the development of validated QSAR models, J. Comput. Aided Mol. Des. 17 (2003) 241–253.

[51] K. Roy, I. Mitra, S. Kar, P.K. Ojha, R.N. Das, H. Kabir, Comparative studies on some metrics for external validation of QSPR models, J. Chem. Inf. Model. 52 (2012) 396–408.

[52] K. Roy, I. Mitra, On the use of the metric rm^2 as an effective tool for validation of QSAR models in computational drug design and predictive toxicology, Mini-Rev. Med. Chem. 12 (2012) 491–504.

[53] A.C. Gaudio, E. Zandonade, Proposição, validação e análise dos modelos que correlacionam estrutura química e atividade biológica, Quím. Nova 24 (2001) 658–671.

[54] P.P. Roy, S. Paul, I. Mitra, K. Roy, On two novel parameters for validation of predictive QSAR models, Molecules 14 (2009) 1660–1701.

[55] N. Chirico, P. Gramatica, 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. 52 (2012) 2044–2058.

[56] E. Besalú, L. Vera, Internal test set (ITS) method: a new cross-validation technique to assess the predictive capability of QSAR models. Application to a benchmark set of steroids, J. Chil. Chem. Soc. 53 (2008) 1576–1580.

[57] A. Tropsha, Best practices for QSAR model development, validation, and exploitation, Mol. Inform. 29 (2010) 476–488.

[58] P. Gramatica, P. Pilutti, E. Papa, Approaches for externally validated QSAR modeling of nitrated polycyclic aromatic hydrocarbon mutagenicity, SAR QSAR Environ. Res. 18 (2007) 169–178.

[59] S.K. Dogra, Outlier, http://www.qsarworld.com/qsar-statistics-outlier.php (accessed 25.01.12), in: QSARWorld – A Strand Life Sciences Web Resource, 2012.

[60] R.P. Verma, C. Hansch, An approach toward the problem of outliers in QSAR, Bioorg. Med. Chem. 13 (2005) 4597–4621.

[61] E. Papa, J.C. Dearden, P. Gramatica, Linear QSAR regression models for the prediction of bioconcentration factors by physicochemical properties and structural theoretical molecular descriptors, Chemosphere 67 (2007) 351–358.

[62] D. Mahajan, V. Masand, K. Patil, T. Hadda, V. Rastija, Integrating GUSAR and QSAR analyses for antimalarial activity of synthetic prodiginines against multi drug resistant strain, Med. Chem. Res. 22 (2013) 2284–2292.

[63] B. Singh, R.A. Vishwakarma, S.B. Bharate, QSAR and pharmacophore modeling of natural and synthetic antimalarial prodiginines, Curr. Comput. Aided Drug. Des. 9 (2013) 350–359.

[64] V.H. Masand, D.T. Mahajan, K.N. Patil, T.B. Hadda, M.H. Youssoufi, R.D. Jawarkar, I.G. Shibi, Optimization of antimalarial activity of synthetic prodiginines: QSAR, GUSAR, and CoMFA analyses, Chem. Biol. Drug Des. 81 (2013) 527–536.

[65] P. Singh, N. Shekhawat, Chemometric descriptors in the rationale of antimalarial activity of natural and synthetic prodiginines, J. Curr. Chem. Pharm. Sci. 2 (2012) 244–260.

[66] D.T. Mahajan, V.H. Masand, K.N. Patil, T. Ben Hadda, R.D. Jawarkar, S.D. Thakur, V. Rastija, CoMSIA and POM analyses of anti-malarial activity of synthetic prodiginines, Bioorg. Med. Chem. Lett. 22 (2012) 4827–4835.

[67] M. Karelson, V.S. Lobanov, A.R. Katritzky, Quantum-chemical descriptors in QSAR/QSPR studies, Chem. Rev. 96 (1996) 1027–1043.

[68] J.O. Morley, T.P. Matthews, Structure–activity relationships in nitrothiophenes, Bioorg. Med. Chem. 14 (2006) 8099–8108.

[69] E.B. de Melo, M.M.C. Ferreira, Multivariate QSAR study of 4,5-dihydroxypyrimidine carboxamides as HIV-1 integrase inhibitors, Eur. J. Med. Chem. 44 (2009) 3577–3583.

[70] K.L. Lang, I.T. Silva, V.R. Machado, L.A. Zimmermann, M.S.B. Caro, C.M.O. Simões, E.P. Schenkel, F.J. Duránd, L.S.C. Bernardes, E.B. de Melo, Multivariate SAR and QSAR of cucurbitacin derivatives as cytotoxic compounds in a human lung adenocarcinoma cell line, J. Mol. Graph. Model. 48 (2014) 70–79.

[71] E.D. Glendening, C.R. Landis, F. Weinhold, Natural bond orbital methods, WIREs Comput. Mol. Sci. 2 (2012) 1–42.

[72] A. Mauri, V. Consonni, M. Pavan, R. Todeschini, Dragon software: an easy approach to molecular descriptor calculations, Commun. Math. Comput. Chem. 56 (2006) 237–248.

[73] J. Gálvez, R. García-Domenech, A.C. de Gregorio, J.V. de Julián-Ortiz, L. Popa, Pharmacological distribution diagrams: a tool for de novo drug design, J. Mol. Graph. 14 (1996) 272–276.

[74] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR Comb. Sci. 26 (2007) 694–701.