Letters in Applied NanoBioScience
Open-Access Journal (ISSN: 2284-6808)

Article
Volume 13, Issue 2, 2024, 90
https://doi.org/10.33263/LIANBS132.090

Efficient Classification and Regression Models for the QSAR of Chloroquine Analogues against Chloroquine-Sensitive and Chloroquine-Resistant *Plasmodium falciparum*

Thecla O. Ayoka [1], Uchenna J. Uchegbu [2], Cyprian C. Alabuike [2], Charles O. Nnadi [2,3,*]

[1] Department of Science Laboratory Technology (Biochemistry Unit), Faculty of Physical Sciences, University of Nigeria Nsukka, 410001 Enugu State Nigeria
[2] Department of Pharmaceutical Chemistry, Faculty of Pharmacy, Madonna University, PMB 05 Elele, Rivers State Nigeria
[3] Department of Pharmaceutical and Medicinal Chemistry, Faculty of Pharmaceutical Sciences, University of Nigeria Nsukka, 410001 Enugu State Nigeria
* Correspondence: charles.nnadi@unn.edu.ng;

Scopus Author ID 42962136300
Received: 26.03.2023; Accepted: 30.05.2023; Published: 30.06.2024

Abstract: The treatment of *Plasmodium falciparum* malaria using chloroquine (CQ) has been hampered by resistance, and there is a need to continually optimize CQ for a more effective analog by structural modification using the QSAR approach. The study developed classification and regression-based models for the optimization and prediction of activities of CQ derivatives against CQ-sensitive (*Pf*D-10) and CQ-resistant (*Pf*W-2) *P. falciparum* strains using a comparative molecular field analysis (CoMFA) and machine learning (ML) algorithms. This present work involves extracting the molecular features and SMILES of CQ analogs from RDKit software and combined to form X- and Y-matrices. The classification model was trained using Python and evaluated using the confusion matrix, accuracy, precision, recall, and F1-score. The CoMFA model was constructed with conformers of the training set (n=20) by the PLS method, cross-validated by leave-one-out, and externally predicted the IC_{50} of the test set (n=7) using Open3dqsar of the Open3dtools. The calibration R^2, cross-validation, Q^2, and prediction P^2 for the y-scrambled data set were 0.996 - 0.684 and 0.998 - 0.682 for *Pf*D-10 and *Pf*W-2 respectively. The decision tree classifier and Naïve Bayes algorithms outperformed the others in the classification model with accuracy scores of ≥ 0.88 against *Pf*D-10 and *Pf*W-2. A single ensemble model obtained by staking all the classification models significantly improved overall accuracy and reliability. The study developed models for the prediction of activity and provided an explanation for the SARs and strategies for further optimization of CQ analogs against *Pf*D-10 and *Pf*W-2 strains.

Keywords: chloroquine; classification; descriptors; *Plasmodium falciparum*; QSAR; regression

1. Introduction

Malaria is a disease caused by the parasite of the genus *Plasmodium* and transmitted through the saliva of female anopheles mosquitoes [1,2]. Sub-Saharan Africa is currently overwhelmed by *Plasmodium falciparum*. Chloroquine (CQ) and other heterocyclic quinoline derivatives are important chemotherapeutic classes and are still useful singly and in rare combinations with primaquine [3]. It is also a first-line treatment and prevention for *P. vivax* and *P. falciparum* malaria in some World Health Organization regions. In addition to a hypersensitivity reaction to CQ in certain individuals, there has been a rapid decline in the

efficacy of CQ, resulting in a further increase in malaria mortality [4]. Various structural modifications of CQ have been reported; however, some of these approaches are expensive and time-consuming and, therefore, could not be translated due to poor design [5]. Artemisinin-based combination therapy has remained the most potent first-line treatment for *P. falciparum*. The emergence and rapid spread of CQ- and artemisinin-resistant strains of *P. falciparum* are indications that a continuous search for a more efficacious remedy for malaria is imperative [6]. CQ's combined safety, favorable physicochemical properties, and cost-effectiveness make it a good candidate for structural modifications to overcome resistance and declining efficacy.

Previous attempts to modify CQ by semi-synthetic or metal incorporation approach yielded derivatives that showed an outstanding advantage over CQ [7-9]. Several quinolone-triazole derivatives showed mid-micromolar activity against *P. falciparum,* which improved significantly upon metal derivatization [10,11]. Several organometallic derivatives of CQ have also been tested against CQ –sensitive (*Pf*D-10) /resistant (*Pf*W-2) strains of *P. falciparum* [12,13]. The volume of molecules that have shown improved activity against *Plasmodium* species is enormous. Apart from ferroquine (organometallic of CQ) under clinical trial, other compounds are yet to be exploited maximally [14]. To develop an efficient statistical model for predicting the biological activities of new chemical entities, the quantitative structure-activity relationship (QSAR) method uses computation or mathematical modeling to uncover relationships between the physicochemical properties of chemical substances and their biological activities [15-17].

To redesign CQ derivatives with outstanding efficacy, a combined classification and regression approach to understanding the structure-activity relationship (SAR) of an existing data set is developed. The model is intended to design better CQ analogs using information from the molecular interaction field (MIF) descriptors (steric and electrostatic) of CQ analogs with bulky side chains [16]. The compounds in this study are grouped into (i) 7-chloro (or trifluoromethyl)-4-aminoquinoline linked to a quinolizidine at position 1, (ii) 7-chloro-4-aminoquinoline linked to a quinolizidine at position 9a and, (iii) 7-chloro (or trifluoromethyl)-4-thioquinoline linked to cysteine [18,19]. The congenericity of the data set assumed the quinoline nucleus as the biophore/pharmacophore for a common target-mediated antiplasmodial activity, while the IC_{50} obtained from a congruent assay in a previous study provided wide ranges of numerical potency of 2.70 against the *Pf*W-2 and 3.04 against *Pf*D-10 *P. falciparum* strains [20,21]. The substituents variations (position 1 or 9a of quinolizidine, and cysteine) connected via linkers (amino, thio, amino methylene, or thiomethylene), in addition to the considerably average number of data sets, provide chemical diversity necessary to capture vital SARs using MIFs-based descriptors in 3D-QSAR modeling of antimalarial activity of CQ derivatives.

This study, therefore, developed efficient models using regression (CoMFA and Multiple Linear Regression (MLR)) and classification (logistic regression, LR, Naive Bayes, NB, knearest neighbours classifier, kNNC, decision tree classifier, (DTC) and support vector classifier, SVC) algorithms to; (i) predict the pIC_{50} of any untested CQ analogs or similar derivatives against *Pf*D-10 and *Pf*W-2) strains; (ii) predict the bioactivity status (active or inactive) of such compounds and (iii) provide an explanation for SARs of CQ analogs against *Pf*D-10 and *Pf*W-2 strains.

2. Materials and Methods

2.1. Data collection.

A dataset of twenty-seven CQ analogs with bulky side chains previously tested against *Pf*W-2 and *Pf*D-10 was mined from a previous report [19]. The IC_{50} values (nM or µM) were converted to their corresponding molar (M) values, and the pIC_{50} was calculated from $-Log(IC_{50})$ as the dependent variables. The detailed chemical structures and the pIC_{50} values of the 27 CQ analogs used in this study are found in the supplementary Table S1. The dataset was randomly split into a training set (n = 20) and a test set (n = 7). For the classical 3D-QSAR CoMFA modeling, all the possible combinations of seven test sets from the 27 data sets were classified manually for the model development.

2.2. Preparation of dataset for MLR and classification modeling.

Based on the distribution of pIC_{50}, a critical decision boundary of $pIC_{50} = 7.00$ was set for the ML-based modeling of activity against *Pf*D-10. Compounds with $pIC_{50} \geq 7.00$ were set as active (label:1; 15), while those with $pIC_{50} < 7.00$ were set as inactive (Label:0; 12). For activity against *Pf*W-2, compounds with $pIC_{50} \geq 6.3$ were set as active (Label:1; 15), while CQ analogs with $pIC_{50} < 6.3$ were set as inactive (Label:0; 12). The RDKit library was used to generate molecular descriptors [22]. Canonical and isomeric SMILES strings for each compound were prepared and then converted to an RDKit molecule object. The detailed SMILES of the CQ analogs and the labels based on the critical decision boundary are found in the supplementary Table S2.

Four different molecular descriptors - cLogP (octanol-water partition coefficient), MW (molecular weight), nRB (number of rotatable bonds), and ArP (aromatic proportion = the number of aromatic atoms/numbers of heavy atoms) were calculated from SMILES of compounds. The cLogP, MW, and nRB were computed from RDKit, while ArP was computed manually from the ratio of the number of aromatic atoms to the total number of heavy atoms. The computed descriptors and *Pf*D-10; pIC_{50} column were combined on one hand, and the *Pf*W-2; pIC_{50} with the computed descriptors on the other hand to form the X-matrix, while the status 1 column of the dataset (Table 1) formed the Y-matrix, for the *Pf*D-10 and the status 2 column formed the Y-matrix for the *Pf*W-2 (Table 1). The X- and Y-matrix were split into a training set (n = 20) and a test set (n = 7), translating to 75% of the initial dataset as the training set, while 25% represented the test set.

Table 1. X-matrix for *Pf*D-10 and *Pf*W-2 activities.

S/N	cLogP	MW	nRB	ArP	*Pf*D-10	cLogP	MW	nRB	ArP	*Pf*W-2
1	5.2978	389.996	6.0	0.384615	7.270	5.2978	389.996	6.0	0.384615	6.586
2	5.6879	404.023	7.0	0.370370	7.567	5.6879	404.023	7.0	0.370370	6.808
3	5.6863	404.023	6.0	0.370370	7.609	5.6863	404.023	6.0	0.370370	7.083
4	6.0764	418.050	7.0	0.357143	7.636	6.0764	418.050	7.0	0.357143	6.920
5	6.0517	437.575	6.0	0.333333	7.251	6.0517	437.575	6.0	0.333333	7.078
6	5.8790	371.956	6.0	0.384615	7.648	5.8790	371.956	6.0	0.384615	6.323
7	6.6592	400.010	8.0	0.357143	7.566	6.6592	400.010	8.0	0.357143	6.711
8	7.4394	428.064	10.0	0.333333	8.046	7.4394	428.064	10.0	0.333333	7.587
9	4.4279	365.864	2.0	0.615385	6.432	4.4279	365.864	2.0	0.615385	6.197
10	5.0656	393.918	4.0	0.571429	6.026	5.0656	393.918	4.0	0.571429	5.751
11	5.4557	407.945	5.0	0.551724	6.934	5.4557	407.945	5.0	0.551724	6.525
12	6.2359	435.999	7.0	0.516129	7.483	6.2359	435.999	7.0	0.516129	7.113
13	5.4541	407.945	4.0	0.551724	6.045	5.4541	407.945	4.0	0.551724	5.979
14	4.9738	438.915	5.0	0.516129	6.938	4.9738	438.915	5.0	0.516129	6.058
15	5.3639	452.942	6.0	0.500000	7.203	5.3639	452.942	6.0	0.500000	6.262

S/N	cLogP	MW	nRB	ArP	*Pf*D-10	cLogP	MW	nRB	ArP	*Pf*W-2
16	6.1441	480.996	8.0	0.470588	7.783	6.1441	480.996	8.0	0.470588	7.376
17	5.0379	422.960	5.0	0.533333	7.017	5.0379	422.960	5.0	0.533333	6.320
18	5.8181	451.014	7.0	0.500000	7.924	5.8181	451.014	7.0	0.500000	7.614
19	6.9807	565.737	5.0	0.516129	7.323	6.9807	565.737	5.0	0.516129	7.042
20	7.7609	593.791	7.0	0.484848	7.775	7.7609	593.791	7.0	0.484848	7.507
21	3.4884	232.714	1.0	0.625000	6.001	3.4884	232.714	1.0	0.625000	6.091
22	5.2448	346.927	3.0	0.434783	5.143	5.2448	346.927	3.0	0.434783	5.344
23	5.6102	380.479	3.0	0.384615	5.000	5.6102	380.479	3.0	0.384615	5.107
24	5.9780	407.048	6.0	0.384615	5.618	5.9780	407.048	6.0	0.384615	5.740
25	6.3681	421.075	7.0	0.370370	5.758	6.3681	421.075	7.0	0.370370	5.849
26	6.1359	424.997	5.0	0.551724	5.007	6.1359	424.997	5.0	0.551724	4.911
27	6.5260	439.024	6.0	0.533333	5.406	6.5260	439.024	6.0	0.533333	5.285

S/N = compounds

2.3. Preparation of dataset for 3D-QSAR modeling.

2.3.1. Energy minimization of compounds.

The 2D structures of all the compounds were built from the fragments in the ChemSketch software module and loaded as a molecular database in the molecular operating environment window. The 2D structures were converted to 3D structures, and energy was minimized using the MMFF94x force field of the molecular operating environment. The 3D structures were subjected to a LowMode dynamic conformational search, adopting all the default settings. The lowest energy (dE=0 and within the global energy minimum of 3.0 kcal per mole) conformers were further subjected to energy minimization using Austin model 1 (AM1), a semi-empirical Hamiltonian algorithm of the MOPAC module [16]. The lowest AM1 energy conformers were then used for the molecular superposition.

2.3.2. Superposition of molecules.

The alignment of the molecules before modeling was done manually using the biophore hypothesis [21]. The structural alignment of the molecules was executed using compounds **8** (pIC_{50} = 8.045) and **18** (pIC_{50} = 7.61) as templates of CQ analogs with activities against *Pf*D-10 and *Pf*W-2, respectively. The alignment was performed using atom matching of the quinoline nucleus of **8** and **18** for the *Pf*D-10 and *Pf*W-2, respectively, and a query of similarity of the various substituents in the rest of the compounds. The aligned molecules and the pIC_{50} were saved in the Open3Dtools compatible file for the QSAR studies.

2.4. Building of classification and regression models.

2.4.1. Classification modeling.

The classification models, using LR, NB, *k*NNC, DTC, and SVC were built to predict a label (active or inactive) for the 27 CQ analogs with activities against the *Pf*D-10 and *Pf*W-2 strains as previously described [23].

2.4.2. Regression modeling.

For the regression model, the MLR model was built to predict the pIC_{50} values of the analogs with activities against both *Pf*D-10 and *Pf*W-2 strains, using the MLR equation shown in equation 1 [24].

$$y=c+b1X1+b2X2+b3X3\ldots+bmXm+e \quad - \quad - \quad 1$$

Where y = predicted variable; c = intercept; b1, b2, b3, bm = regression coefficients; X1, X2, X3, Xm = molecular descriptors; e = residual error

The parameter settings of the ML models are summarized in supplementary Table S3. The models were trained using a test size of 0.25 and a random state of 42.

2.4.3. Model evaluation.

Different evaluation metrics (confusion matrix, accuracy, precision, recall, F1-score) were deployed to evaluate the performance of the classification models [23]. Additionally, the squared coefficient of determination (R^2), mean squared error (MSE), mean absolute error (MAE), and root mean squared error (RMSE) were used to evaluate the performance of the MLR model [25]. The five classification models were combined into a single ensemble model using stacking. The models were plotted to measure their performance.

The parameters for the classification model, true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) were used for the evaluation of the models' performance [23]. Specificity, also called false positive, which is the proportion of the inactive compounds, and predicted as such was obtained. Sensitivity (recall), which is called true positives, is the proportion of true cases predicted as true by the model, and only positive cases were considered while leaving out negative cases. Precision, the proportion of true cases predicted as true, for both positive and negative cases were classified. The F1-score, which is the mean of precision and recall, and the accuracy representing the proportion of actual predicted cases out of the whole cases were ascertained.

The most often used metric for assessing regression-based models is MSE, which estimates how closely the fitted line matches the actual data points [26]. As a result, the training model uses MSE as a loss function that it seeks to minimize. The performance of the regression model improves with decreasing MSE values. In many studies, root mean squared error (RMSE), which is the square root of MSE, is utilized in place of MSE. In multiple regression models, the squared correlation between the actual and predicted values is equivalent to the coefficient of determination (R^2). The percentage of the outcome's variation that the predictor variables account for is known as R^2. A negative R^2 means the model performed poorly and did not match the data trend (i.e., chosen by chance). The R^2 value indicates how well a model fits the data and how closely the data resembles the regression line. An elevated R^2 indicates a good model fit [24]. The ML classifier took advantage of these correlations to provide effective results.

2.4.4. CoMFA-based modeling.

2.4.4.1. Pretreatment of aligned molecules.

The 3D-QSAR study was executed with an open-sourced Open3DQSAR software of the Open3dtools package [16]. The superposed molecule was loaded using script commands and enclosed in a grid box of step size 2 Å. The descriptors (steric and electrostatic MIFs) were computed in an MMFF94 field using charges and van der Waals parameters. High van der Waals and electrostatic grid energy values higher than 0.0001 kcal/mol were cut off. Other pretreatments such as energy filtration to +30 kcal/mol of MIF data of training set, block unscaled weighting, excluding variables which have a low SD, removal of N-level variables, and reducing the size of the PLS matrix were executed following the instruction text scripts [21].

2.4.4.2. Model calibration, internal validation and external prediction.

The model calibration was executed by the correlation of the MIF energies with the pIC_{50} using partial least squares (PLS) regression and extracting all the latent variables (PLS components (PCs)). The model was improved by combining close variables (smart region definition) and selecting the most relevant variables [16]. The model was cross-validated internally by the leave-one-out (LOO) method, and the performance was measured by the coefficient of determination, Q^2. The external predictive power of the model was evaluated by predicting the activities of randomized sets of test compounds, which were not used in the model calibration. The test set molecules that produced the best models for *Pf*D-10 (**4**, **7**, **9**, **19**, **21**, **25,** and **27**) and *Pf*W-2 (**2**, **9**, **11**, **16**, **21**, **25,** and **26**) strains were pretreated as the training set molecules. The predictive power of the PLS-built model was expressed as the coefficient of determination, R^2, for the correlation.

2.4.4.3. C1ontour mapping.

The contour maps of the developed models were visualized using the ReadMoeGrid module [21]. The regression coefficients exported from Open3dtools were plotted with emphasis on where the variations of Lennard-Jones and Coulombic potential properties in the structural spaces of the training set molecules lead to significant variation in the activity of CQ analogs against *Pf*D-10 and *Pf*W-2. The green and blue colors represented the positive influence of Lennard-Jones and Coulombic potential interactions, respectively, on the activity, while the negative influence of steric and electrostatic interactions on the activity was represented by yellow and red colors, respectively.

2.4.4.4. Application of the model.

Based on the contour mapping, several putative CQ analogs (p-CQ) with potentially high synthetic accessibility were designed for further optimization of the model. Six (**18**a-f) and four (**8**a-d) derivatives of **18** and **8** were designed for PfW-2 and PfD-10 activity, respectively. The compounds were built and pretreated, aligned with **1**-**27**, and the MIF energies were re-calculated as par with the compounds used in building the model. The QSAR and PLS regression analysis were executed, with **8**a-d and **18**a-f as the test sets. To fit into the molecular field analysis of the 3D-QSAR module, the compounds were assigned the status of untested compounds (pIC_{50} =0), and the model regression was applied to predict the activity against PfW-2 and PfD-10.

3. Results and Discussion

3.1. Classification modeling.

The LR model had a model accuracy of 86% for both *Pf*D-10 and *Pf*W-2 strains. For the *Pf*D-10 strain, the precision for the ‘active’ showed 0.75, while that of the ‘inactive’ was 1.00. The model correctly predicted the inactive cases. For the *Pf*W-2 strain, the precision score for the ‘inactive’ was also 0.75, and 1.00 for the “inactive” case. Recall and F1 scores have the same values as the precision scores for the *Pf*W-2 strain. For the *Pf*D-10, the recall scores for both the ‘active’ and ‘inactive’ cases were 1.00 and 0.75, respectively. These showed the correct prediction of the actual positive cases. The F1 score was 0.86 for both ‘active’ and ‘inactive’ cases. The TP, TN, FP and FN counts were observed in each classification model

(Table 1). For the *Pf*D-10 strain, DTC and NB attained the highest TP of 4 out of 4, and the lowest FP of 0 out of 4. The LR and SVC were the next best-performed models with TP of 3 out of 4, and FN of 0 out of 3. The least performed model was *k*NNC attaining a TP of 0 out of 4 and an FP of 4 out of 4. For the *Pf*W-2 strain, DTC was the best-performed model with TP of 3 out of 3, TN of 4 out of 4, and lowest FN of 0 out of 4. LR and SVC were the next best-performed model achieving a TP of 3 out of 3, and an FN of 1 out of 4. The *k*NNC was the least performed model, having a TP of 0 out of 3, and an FN of 0 out of 4. The performance of the DTC models (*Pf*D-10 and *Pf*W-2 strains) was further described using a Confusion Matrix (Figure 1). It contains the actual and predicted values.

Table 1. Overall predictions of the models based on TP, TN, FP, and FN counts.

Models	***Pf*D-10**				***Pf*W-2**			
	TP	**TN**	**FP**	**FN**	**TP**	**TN**	**FP**	**FN**
LR	3	3	1	0	3	3	0	1
*k*NNC	0	2	4	1	0	4	3	0
DTC	4	2	0	1	3	4	0	0
SVC	3	3	1	0	3	3	0	1
Naïve Bayes	4	2	0	1	3	2	0	2

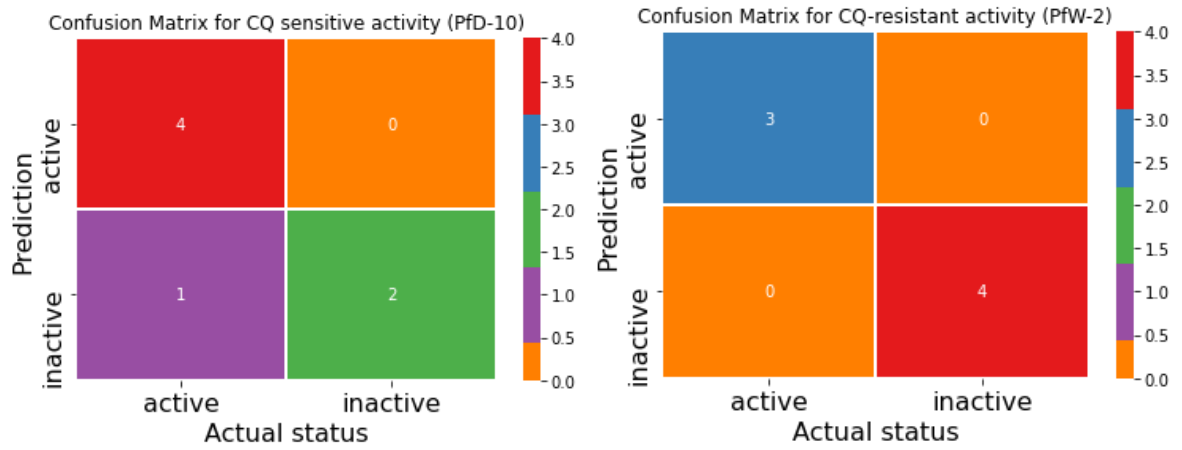


Figure 1. Confusion matrix for the modeled CQ analogs against *Pf*D-10 and *Pf*W-2 strains.

3.1.1. Comparison of the models.

The performance of the models was plotted and compared (supplementary Figure S1). The LR, DTC, and NB outperformed the other models. The LR model had an accuracy score of 0.88 ± 0.21 (PfD-10) and 0.85 ± 0.29 (*Pf*W-2). The accuracy scores for DTC and NB were 0.95 ± 0.15 (*Pf*D-10), 0.92 ± 0.19 (*Pf*W-2) and 0.90 ± 0.20 (*Pf*D-10), 0.95 ± 0.15 (*Pf*W-2) respectively. On the other hand, the accuracy score for *k*NNC was 0.45 ± 0.30 (*Pf*D-10) and 0.30 ± 0.32 (*Pf*W-2), while SVC had an accuracy score of 0.60 ± 0.20 (*Pf*D-10) and 0.53 ± 0.13 (*Pf*W-2). When the models were combined into a single ensemble model using staking, the staked model significantly improved the performance of the DTC model for both *Pf*D-10 and *Pf*W-2 strains.

3.2. Regression modeling.

Regression models were built for the 27 CQ analogs tested against *Pf*D-10 and *Pf*W-2 *P. falciparum* to predict the pIC_{50} values of compounds using MLR and 3D-QSAR CoMFA analyses.

3.2.1. MLR Model.

The MLR model was built to predict the pIC_{50} values of compounds against both *Pf*D-10 and *Pf*W-2 strains. Each feature is viewed as a dimension, and four x values were used in the MLR algorithm. Regression equations 2 and 3 were obtained for the modeling against *Pf*D-10 and *Pf*W-2, respectively.

pIC_{50}=7.02-1.01clogP+0.0088MW+0.47nRB-1.48ArP - - 2

pIC_{50}=5.39-0.56cLogP+0.0054MW+0.31nRB+0.70ArP - - 3

The results of the performance of the models are shown in Table 2. The MLR model for the *Pf*D-10 and *Pf*W-2 strains showed an R^2 of 0.70 and 0.47 for the test data, respectively. The MSE scores for the *Pf*D-10 and *Pf*W-2 strains were 0.17 and 0.50 for the test set, respectively. The correlation of the predicted and experimental pIC_{50} values against *Pf*D-10 and *Pf*W-2 is shown in Figure 2.

Table 2. Performance of the MLR model.

Evaluation metrics	*Pf*D-10		*Pf*W-2	
	Test set	Train set	Test set	Train set
Coefficient of determination (R^2)	0.70	0.51	0.47	0.40
Mean absolute error (MAE)	0.38	0.60	0.60	0.43
Mean Squared Error (MSE)	0.17	0.50	0.50	0.27
Root Mean Squared Error (RMSE)	0.41	0.71	0.70	0.52

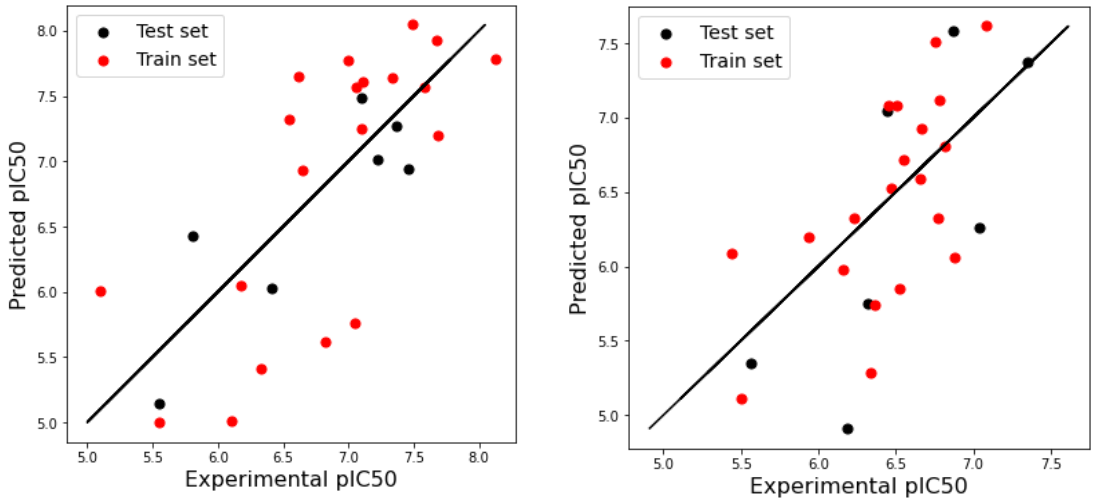


Figure 2. Scatter plot of the predicted and experimental pIC_{50} of modeled CQ analogs against *Pf*D-10 and *Pf*W-2 strains

3.2.2. 3D-QSAR CoMFA model.

3.2.2.1. Alignment of molecules.

The lowest energy conformers of the AM1 minimized compound with the strongest activities against PfD-10 (**8**) and PfW-2 (**18**) strains were used as templates for the alignment of molecular structures. For each data set, the atoms of the quinoline moiety were used as the superposition point while querying its similarity with other analogs. The aligned molecules (supplementary Figure S2) showed a similar arrangement in both sets of compounds tested against PfD-10 and PfW-2, with slight conformational differences arising from the different superposition templates used,

The direct relationships between the MIF energy differences of the CQ analogs with the changes in the pIC_{50} of the molecules for both PfD-10 and PfW-2 strains were modeled

using the PLS regression method with up to 5PCs. The models were cross-validated and used to predict the pIC_{50} of compounds excluded from the model building. The detailed model statistical parameters for both models are summarized in Table 3.

Table 3. Model statistics for the activity of CQ analogs, 5PCs.

Parameter	*Pf*D-10	*Pf*W-2
R^2±SDEC, P value	0.996±0.061; <0.0001	0.998±0.033; <0.0001
Q^2±SDEP, P value	0.785±0.528; <0.0001	0.743±0.548; <0.0001
P^2±SDEP, P value	0.684±0.503; 0.0218	0.682±0.458; 0.0127
$Sy.x_{(MIE)}$	0.0627, 0.2996, 0.3634	0.0345, 0.2805, 0.2038
F-ratio	737.2	1493.5
$F\ (DFn, DFd)_{MIE}$	4905; 65.80; 10.80	3935; 71.40; 12.60
% (steric; electrostatic)	27.5; 72.5	38.6; 61.4
Regression equation	y = 0.5313x + 3.111	y = 0.4050x + 3.932

R^2 = non-cross validated; Q^2 = internal cross-validated; P^2 = external predicted coefficient of determination; F = Fisher value (critical F-values for the 95 % probability level are reported in parentheses); SDEC = standard deviation error in calculation; SDEP = standard deviation error in prediction; x = actual pIC_{50}; y = predicted pIC_{50}; MIE= method validation, internal validation and external; prediction.

3.2.2.2. PLS analysis for PfD-10 and PfW-2 activities.

The MIFs-pIC_{50} relationship of CQ analogs tested against *Pf*D-10 produced a better model compared with those tested against *Pf*W-2 strains with R^2 = 0.996, Q^2 = 0.785, and P^2 = 0.684. In this model, five latent variables were required to obtain the optimum model with the set of training and test data sets. The model statistics and the plots of predicted versus actual activity data are shown in Table 3 and Figure 3A, respectively. There was also a strong correlation between the MIF energies of CQ analogs with pIC_{50} against *Pf*W-2 strains. The best model comprised an optimum number of five latent variables (5PCs) and yielded a PLS coefficient of determination of R^2 = 0.998, Q^2 = 0.743, and P^2 = 0.682 for the non-cross validation, LOO cross-validation, and external prediction, respectively. The model statistics are shown in Table 3, and the plots of predicted pIC_{50} against actual pIC_{50} are shown in Figure 3B.

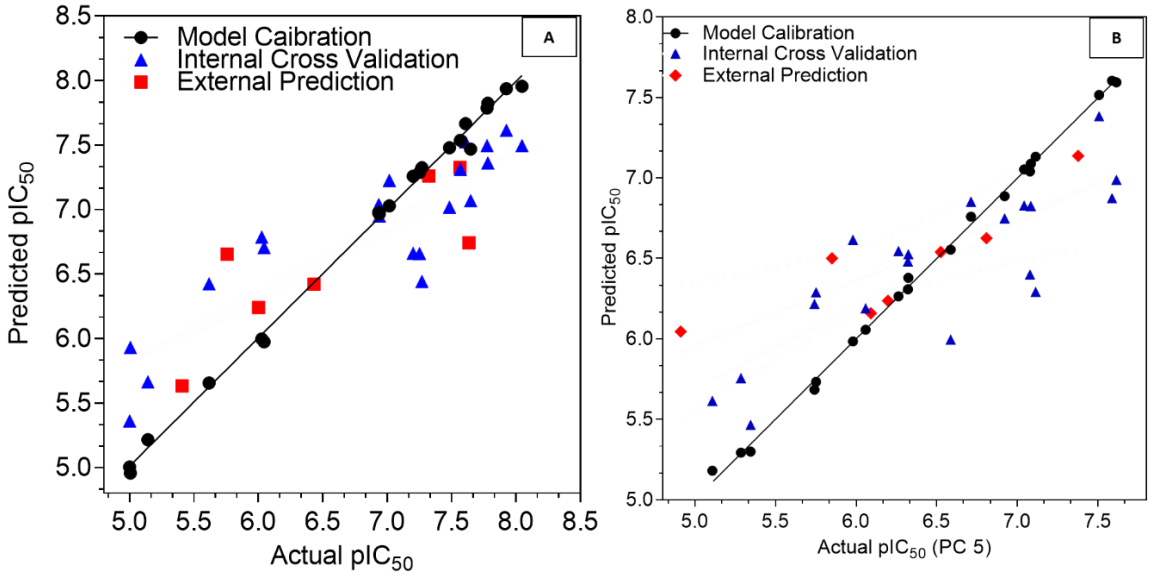


Figure 3. Scatter plots of predicted vs actual pIC_{50} of modeled CQ analogs against PfD-10 **(A)** and PfW-2 **(B)** strains. Data are from the PLS model with 5PCs and represent non-cross-validated (black), internal predictions LOO cross-validation (blue), and test set predictions (red). The black trend line in D is for the non-cross-validated scatter plot.

3.2.2.3. Contour maps of CQ analogs active against PfD-10 and PfW-2.

The influence of the electrostatic effect on the activity of the CQ analogs against *Pf*D-10 was found to be stronger at the positions of the amino group of the 4-aminoquinoline and the N atom of the other moieties (Supplementary Figure S3). The electrostatic interaction between the positively charged probe of the receptor site and the 4-amino group was found to enhance the activity of the CQ analogs, while the second nitrogen decreased the activity. The longer chain length was observed to increase the distance between the N atom and the red contours, making their influence less important than the 4-amino group. The steric effect of CQ analogs was strongest in the vicinity of 4-amino or 4-thio substituents. There was a heavy presence of yellow contour around the moderately shorter chain length at position 4 of the quinoline ring. There were also green contours around the N than the S heteroatom of the substituent, where C-4 nitrogenous ligands showed higher activity.

In the contour maps of CQ analogs with *Pf*W-2 activity, significant electrostatic effects are observed within the vicinity of the quinoline ring, 4-amino group, and the amine, nitro, or bromine substituents on the 4-amino substituents (Blue region, Figure S4) and the nitrogen atoms of the quinoline ring and 4-amino group (red regions, Supplementary Figure S4. The N-atoms of the 4-amino group and the quinoline ring interact with the blue and red contours. However, there is stronger interaction with the blue contours, as seen in the shorter distances between the atoms and the blue regions. The interaction of the blue contour with the N and O atoms of the 4-amino substituent was also prominent. The steric effects on the activity of CQ analogs against *Pf*W-2 are predominant on the carbon chain (green area, Figure S4A) and the aromatic ring of the 4-amino substituent (yellow density, Figure S4B). The carbon chain connecting the 4-amino group to the aromatic ring of the substituent interacts strongly with the green contour at all sides. At the same time, the yellow area protrudes from the aromatic ring and slightly extends to the sp3 hybridized chains of the pyridine ring (**9**-**20**, **26**, **27**).

3.2.2.4. Application of the CoMFA model to the design of CQ analogues.

Ten CQ analogs, derived from **8** (CQ-sensitive *P. falciparum* active) and **18** (CQ-resistant *P. falciparum* active), were subjected to all algorithms executed in building the model to validate the robustness further. The derivatives of **8** and **18** used for the assessment and the aligned molecules of **8**, **18**, **8**a-d, and **18**a-f are shown in Figure 4. The alignment of the designed CQ analogs with **8** or **18** showed similar patterns with **1**-**27**.

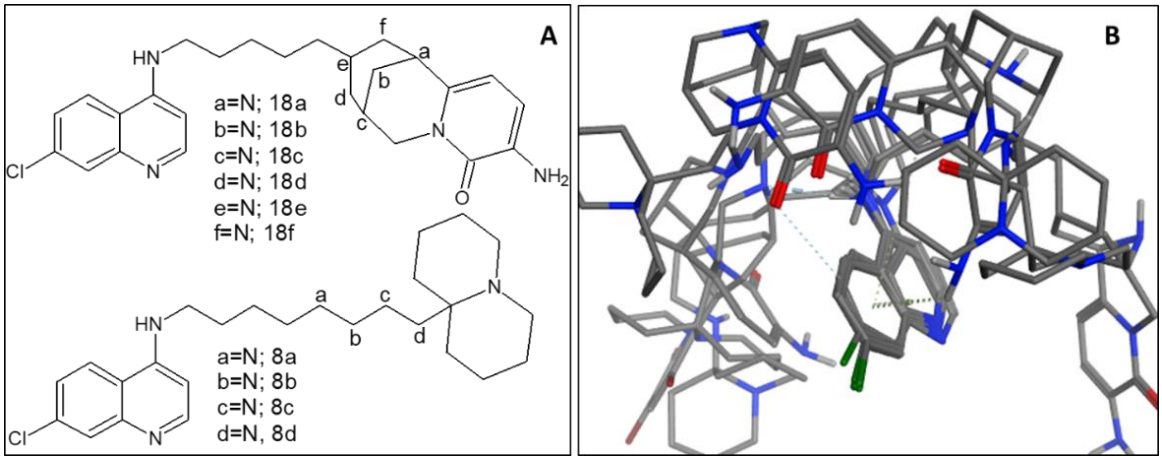


Figure 4. Designed structures of p-CQ of **8** and **18 (A)** and 3D aligned molecules **(B).**

The predicted pIC_{50} for the 5PCs (Table 4) showed that the designed **8** series showed higher activity against the PfD-10 strain, while the PfW-2 strain was more susceptible to compound **18** series. Of the putative compounds, analogs **8**a and **8**b showed improved activity against PfD-10, while analogs **18**b, **18**c, and **18**e were found to inhibit the growth PfW-2 compared with all tested compounds.

Table 4. Predicted pIC_{50} of proposed CQ derivatives.

8-CQ	pIC_{50}, 5PCs		18-CQ	pIC_{50}, 5PCs	
	***Pf*D-10**	***Pf*W-2**		***Pf*D-10**	***Pf*W-2**
8a	8.653	5.876	**18**a	5.948	6,045
8b	8.754	5.684	**18**b	5.876	7.673
8c	6.702	6.342	**18**c	6.365	8.004
8d	7.012	6.734	**18**d	6.093	6.903
			18e	6.713	9.692
			18f	6.325	6.039

3.3. Discussion.

The study developed classification and regression models to predict the pIC_{50} of CQ analogs against PfD-10 and PfW-2 strains of *P. falciparum*, representing a significant step in optimizing these CQ analogs to discover more potent CQ chemotherapy. This complements other studies that investigated the prediction of other chemical analogs against *P. falciparum* [27, 28]. The models developed in this study may be instructive for application to other antimalarial drug optimization [29, 30]. Studies have developed several efficient approaches to designing more effective CQ analogs for treating malaria. For instance, assembling two or more different pharmacophoric hybrids via a linker is vital to CQ optimization. This is to provide one molecule with different targets for interaction and, hence, different mechanisms of action. A previous study has established the potential of aminoquinoline linked to quinolizidine or cysteine via methylene or thio moiety as antiplasmodial agents. The study provided some SARs of the CQ analogs against PfD-10 and PfW-2 strains without empirical or in silico biochemical explanations. In addition, there were no insights into the interaction mechanism with putative targets in *P. falciparum* or optimization strategy for more effective CQ analogs. The present study, in addition to rationalizing the SARs using computational tools for further optimization, was designed to construct more efficient classification and regression-based models for the prediction of the activities of CQ analogs against both PfD-10 and PfW-2 strains of *P. falciparum* using ML and CoMFA approaches.

Regression-based modeling is a statistical or chemometric technique that forms the foundation of QSAR. The study adopted linear regression analysis (MLR) and multivariate data analysis (PLS) methods to correlate the CQ properties of a large number of its features with the pIC_{50} against PfD-10 and PfW-2 strains of *P. falciparum*. In the PLS regression, the large-sized MIFs relative to the latent variables were correlated with the pIC_{50} using the quinoline nucleus as the pharmacophore. The alignment of molecules is the rate-determining step in CoMFA modeling. The aligned CQ analogs, which assumed the quinoline as the biophore and represented the possible mechanism of action, provided interesting data on the interactions with the putative targets. The biophore (Figure 4) showed a three-pronged 3D interaction of the aligned molecules, representing the three moieties linked to the quinoline template of CQ analogs in this study [19]. These moieties included a link at position 1 or 9a of quinolizidine and cysteine coupled to quinoline via amino, thio, amino methylene, or thiomethylene linkers, suggesting that the linked moieties could serve as pharmacophores in

addition to the quinoline nucleus used as the superposition template in this study. This hybrid approach to the design of antimalarial compounds has provided one molecule-multiple mechanism of action model in recent times, such as the inhibition of hemozoin formation by alkylation of heme, activation of the efflux pump of the digestive vacuoles, inhibition of PfCRT or inhibition of falcipain 2 enzymes [31, 32].

The importance of the hybrid assembly design of quinoline and quinolizidine or cysteine, in addition to their significant pIC_{50}s against PfD-10 and PfW-2 strains, led to this attempt to optimize the CQ analogs for improved activities. The CoMFA model for activity against PfD-10 showed strong correlation coefficients of R^2 = 0.996, Q^2 = 0.785, and P^2 = 0.684; F-ratio of 737.2 and 72.5% electrostatic effect on the pIC_{50} with all the correlation coefficients higher than 0.5, $\delta(Q^2, P^2)$ of 0.101 upon randomization of the data set [21]. Similarly, a promising model statistics of R^2 = 0.998, Q^2 = 0.743, P^2 = 0.682; F-ratio of 1493.5, 61.4% electrostatic effect and $\delta(Q^2, P^2)$ = 0.041 were obtained for the activity against PfW-2. In both models, the $\delta(Q^2, P^2)$ was significantly low, SDEP << P^2, F-ratio >> 500, and all the correlation coefficients were higher than 0.50. These statistical parameters were indicators of the model's robustness, reliability, high efficiency and reproducibility, and low possibility of chance correlations [16]. A major factor that produced the indicators was the significantly high randomization adopted in this study's train and test sets selection. To ensure robustness, reproducibility and eliminate chance correlation, a total of 888,030 ($^{27}C_7$) models involving the randomized train and test data sets were generated to select the best models considering the PCs' R^2, Q^2 and P^2 [21]. To further understand the correlation, MLR model was built. However, the coefficients of correlation R^2 were lower, 0.70 and 0.47 (test set) and 0.51 and 0.40 (train set) for PfD-10 and PfW-2 strains, respectively. However, there was consistency in declined R^2 for PfW-2 compared with PfD-10 for both CoMFA and MLR regression models. One important finding in this study is the correlation of the pIC_{50} with the MIFs (electrostatic and steric effects) on the one hand and some Lipinski parameters (MW, cLogP,) of the CQ analogs against both PfD-10 and PfW-2 strains of *P. falciparum*; a rare combination of parameters that can partly predict or describe the oral bioavailability of the CQ analogs [33].

The correlation coefficients of the activities of CQ analogs against PfD-10 and PfW-2 strains translated into contour maps revealed three major factors contributing to the SARs of the compounds. The nature and number of the carbon chains of the linker, the substitutions on the bulky basic side chain, and the type of bulky basic side chain (quinolizidine or cysteine) are surrounded by different contour colors. A strong electrostatic interaction between the cysteine side chain and the binding target leads to the higher pIC_{50} of the compounds with the cysteine side chain (compounds **9**-**20**). However, few exceptions existed, with compounds **9**, **10**, and **13** having lower pIC_{50} than the rest. This observation was due to the relatively lower number of carbon chains, which is 2 [19]. The long-chain carbon (C > 2) was enclosed in a green pocket of the contour, suggesting a strong positive steric effect on compounds with a high number of carbon linkers. The nature of the linker also influenced the activity in some ways. Four main types of linkers were found to influence the activity of the compounds: -N-, N-C-, N-C-S-C-, and -S-C- bridges. The ten CQ derivatives designed to test the applicability of the CoMFA models utilized the three factors (linker chain length >>5 bridged by N, quinolizidinylalkyl, and cytisinylalkyl derivatives) that contributed maximally to the potency of the CQ analogs [19]. On bridging the quinolizidinylalkyl linker (n = 8 for **8**a-**8**e) with an N atom at a different position, additional electrostatic interaction with the target probe was obtained for **8**a and **8**b, resulting in higher pIC_{50} of 8.65 and 8.75 respectively against PfD-10

with no corresponding effect on PfW-2. Such bridging in the cytisinylalkyl derivatives of **18** (n = 5) with substituents on the pyridine ring significantly enhanced the pIC_{50} of **18**b, **18**c, and **18**e against PfW-2 which could be attributed to the electrostatic influence of the electrophilic group and the peptidomimetic pyridine ring [34, 35].

Machine Learning (ML) models were built to tackle a binary classification problem to predict a label (active or inactive) in both PfD-10) and PfW-2 *P. falciparum* datasets. Different evaluation metrics were used to assess the performance of the models [36]. Model evaluation was necessary to select the best-performing model for each strain. Precision scores show that the prediction of positive cases was truly positive.

On the other hand, recall computes the prediction of positive cases from the total positive cases. High precision and recall scores show that models predicted true positive results. Accuracy scores show the ratio of the true predicted cases and the total observations; as a result, high accuracy scores measure the positive predictions of the model [37]. In the modeling for both PfD-10) and PfW-2 *P. falciparum* activities, LR, DTC, and NB algorithms produced more efficient models that compare well with the ensemble model formed from the five algorithms. The best-performed models yielded an accuracy higher than 0.75. An accuracy score of 0.86 means that 86% of the predicted cases are correct, which can be confirmed by a high F1 score. These results agreed with similar works on predicting antimalarial activities of CQ and artemisinin-based combination treatment against *P. falciparum* [38]. The study achieved an accuracy of 81% and 79% with Extreme Gradient Boost and Artificial Neural Network, respectively. In another similar work, ML algorithms were used to validate and develop predictive antimalarial models, achieving an accuracy of 85%, and using the Support Vector Machine algorithm, an accuracy of 85.94% was achieved [39, 40]. When the General Regression Neural Network was used to predict antimalarial activity against *P. falciparum*, it achieved an accuracy of 88.9% [23].

4. Conclusions

The study developed efficient classification and regression-based models to predict the activity of CQ analogs against *Pf*D-10 and *Pf*W-2 strains of *P. falciparum*. The regression (CoMFA and MLR) and classification (ensemble) models are robust, have high accuracy, and can be applied to predict the activity of untested CQ analogs. The study also showed that a cysteine moiety linked to the quinoline nucleus through a –N-C- linker, with a carbon chain of at least 3 gives more potent CQ analogs against *Pf*D-10 and *Pf*W-2 strains of *P. falciparum*. The study further provided strategies for designing more active CQ analogs through structural modification of the quinoline substituents and the nature of the linkage between the quinoline nucleus and moiety.

Funding

This research received no external funding.

Acknowledgments

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Oguike, E.O.; Ugwuishiwu, C.H.; Asogwa, C.N.; Nnadi, C.O.; Obonga, W.O.; Attama, A.A. Systematic review on the application of machine learning to quantitative structure–activity relationship modeling against *Plasmodium falciparum*. *Molec Divers* **2022**, *26*, 3447-3462. https://doi.org/10.1007/s11030-022-10380-1
2. Ikerionwu, C.; Ugwuishiwu, C.; Okpala, I.; James, I.; Okoronkwo, M.; Nnadi, C.; Orji, U.; Ebem, D.; Ike, A. Application of machine and deep learning algorithms in optical microscopic detection of *Plasmodium* parasites: A malaria diagnostic tool for the future. *Photodiagn Photodyn Ther* **2022**, *40*, 103198. https://doi.org/10.1016/j.pdpdt.2022.103198
3. Njiro, B.J.; Mutagonda, R.F.; Chamani, A.T.; Mwakyandile, T.; Sabas, D.; Bwire, G.M. Molecular surveillance of chloroquine-resistant *Plasmodium falciparum* in sub-Saharan African countries after withdrawal of chloroquine for treatment of uncomplicated malaria: A systematic review. *J Inf Public Health* **2022**, *15*, 550-557. https://doi.org/10.1016/j.jiph.2022.03.015
4. Cortopassi, W.A.; Gunderson, E.; Annunciato, Y.; Silva, A.E.; dos Santos Ferreira, A.; Teles, C.B.G.; Pimentel, A.S.; Ramamoorthi, R.; Gazarini, M.; Meneghetti, M.R.; Guido, R.V. Fighting *Plasmodium* chloroquine resistance with acetylenic chloroquine analogues. *Int J Parasitol: Drugs Drug Resistance* **2022**, *20*, 121-128. https://doi.org/10.1016/j.ijpddr.2022.10.003
5. Wattanakul, T.; Chotsiri, P.; Scandale, I.; Hoglund, R.M.; Tarning, J. A pharmacometric approach to evaluate drugs for potential repurposing as COVID-19 therapeutics. *Expert Rev Clin Pharmacol* **2022**, *15*, 945-958. https://doi.org/10.1080/17512433.2022.2113388
6. Cheema, H.S.; Singh, M.P. Drug resistance in plasmodium, future malaria management strategies and importance of medicinal plants. *J Ayurvedic Herbal Med* **2022**, *8*, 107-112. https://doi.org/10.31254/jahm.2022.8209
7. Fisher, G.M.; Tanpure, R.P.; Douche, A.; Andrews, K.T.; Poulsen, S.A. Synthesis and evaluation of antimalarial properties of novel 4-aminoquinoline hybrid compounds. *Chem Biol Drug Design* **2014**, *84*, 462-472. https://doi.org/10.1111/cbdd.12335
8. Aguiar, L.; Biosca, A.; Lantero, E.; Gut, J.; Vale, N.; Rosenthal, P.J.; Nogueira, F.; Andreu, D.; Fernàndez-Busquets, X.; Gomes, P. Coupling the antimalarial cell-penetrating peptide TP10 to classical antimalarial drugs primaquine and chloroquine produces strongly hemolytic conjugates. *Molecules* **2019**, *24*, 4559. https://doi.org/10.3390/molecules24244559
9. Vandekerckhove, S.; D'hooghe, M. Quinoline-based antimalarial hybrid compounds. *Bioorg Med Chem Let*t **2015**, *23*, 5098-5119. https://doi.org/10.1016/j.bmc.2014.12.018
10. Li, Y.; de Kock, C.; Smith, P.J.; Chibale, K.; Smith, G.S. Synthesis and evaluation of a carbosilane congener of ferroquine and its corresponding half-sandwich ruthenium and rhodium complexes for antiplasmodial and β-hematin inhibition activity. *Organometallics* **2014**, *33*, 4345-4348. https://doi.org/10.1021/om500622p
11. Arise, R.O.; Elizabeth, S.N.; Farohunbi, S.T.; Nafiu, M.O.; Tella, A.C. Mechanochemical synthesis, in vivo antimalarial and safety evaluation of amodiaquine-zinc complex. *Acta Facultatis Medicae Naissensis* **2017**, *34*, 221-233. https://doi.org/10.1515/afmnai-2017-0024
12. Pradines. B.; Fusai. T.; Daries. W.; Laloge. V.; Rogier. C.; Millet. P.; Panconi. E.; Kombila. M.; Parzy. D. Ferrocene-chloroquine analogues as antimalarial agents: in vitro activity of ferrochloroquine against 103 Gabonese isolates of *Plasmodium falciparum*. *J Antimicrob Chemother* **2001**, *48*, 179-184. https://doi.org/10.1093/jac/48.2.179
13. Domarle, O.; Blampain, G.; Agnaniet, H.; Nzadiyabi, T.; Lebibi, J.; Brocard, J.; Maciejewski, L.; Biot, C.; Georges, A.J.; Millet, P. In vitro antimalarial activity of a new organometallic analog, ferrocene-chloroquine. *Antimicrob Agents Chemother* **1998**, *42*, 540-544. https://doi.org/10.1128/AAC.42.3.540
14. McCarthy, J.S.; Rückle, T.; Djeriou, E.; Cantalloube, C.; Ter-Minassian, D.; Baker, M.; O'Rourke, P.; Griffin, P.; Marquart, L.; Hooft van Huijsduijnen, R.; Möhrle, J.J. A phase II pilot trial to evaluate the safety and efficacy of ferroquine against early *Plasmodium falciparum* in an induced blood-stage malaria infection study. *Malaria J* **2016**, *15*, 1-9. https://doi.org/10.1186/s12936-016-1511-3
15. Arabi, A.A. Artificial intelligence in drug design: algorithms, applications, challenges and ethics. *Future Drug Discov 59* **2021**, *3*, 2631–3316. https://doi.org/10.4155/fdd-2020-0028

16. Ikwuka, C.E.; Asogwa, C.M.; Ikwuka, O.J.; Ogbonna, J.E.; Onah, C.E.; Ohama, C.C.; Nnadi, C.O. Insights into the In-vivo Antiplasmodial Activity of Trisdimethylamino Pyrimidine Derivative in Plasmodium berghei Infected Mouse Model. *J Pharm Res Int* **2022**, *34*, 33-40. https://doi.org/10.9734/JPRI/2022/v34i49B36430
17. Danishuddin, M.G.; Malik, M.; Subbarao, N. Development and rigorous validation of antimalarial predictive models using machine learning approaches. *SAR and QSAR Environ Res* **2019**, *30*, 543–560. https://doi.org/10.1080/1062936X.2019.1635526
18. Melis, D.R.; Barnett, C.B.; Wiesner, L.; Nordlander, E.; Smith, G.S. Quinoline-triazole half-sandwich iridium (III) complexes: Synthesis, antiplasmodial activity and preliminary transfer hydrogenation studies. *Dalton Transactions* **2020**, *49*, 11543-11555. https://doi.org/10.1039/D0DT01935F
19. Tasso, B.; Novelli, F.; Tonelli, M.; Barteselli, A.; Basilico, N.; Parapini, S.; Taramelli, D.; Sparatore, A.; Sparatore F. Synthesis and antiplasmodial activity of novel chloroquine analogues with bulky basic side chains. *ChemMedChem* **2015**, *10*, 1570-1583. https://doi.org/10.1002/cmdc.201500195
20. Nnadi, C.O.; Ayoka, T.O.; Okorie, H.N. A ligand-based approach to lead optimization of N, N′-substituted diamines for leishmanicidal activity. *Biointerface Res Appl Chem* **2022**, *12*, 7429-7437. https://doi.org/10.33263/BRIAC126.74297437
21. Nnadi, C.O.; Althaus, J.B.; Nwodo, N.J.; Schmidt, T.J. A 3D-QSAR study on the antitrypanosomal and cytotoxic activities of steroid alkaloids by comparative molecular field analysis. *Molecules* **2018**, *23*, 1113. https://doi.org/10.3390/molecules23051113
22. Sydow, D.; Morger, A.; Driller, M.; Volkamer, A. TeachOpenCADD : A teaching platform for computer-aided drug design using open-source packages and data. *J Cheminform* **2019**, *11*, 29–36. https://doi.org/10.1186/s13321-019-0351-x
23. Liu, Q.; Deng, I.; Liu, M. Classification models for predicting the antimalarial activity against *Plasmodium falciparum*. *SAR and QSAR Environ Res* **2020**, *31*, 313–324. https://doi.org/10.1080/1062936X.2020.1740890
24. Stanelle, S.T.; Crouse, S.F.; Heimdal, T.R.; Riechman, S.E.; Remy, A.L.; Lambert, B.S. Predicting muscular strength using demographics, skeletal dimensions, and body composition measures. *Sports Med Health Sci* **2021**, *3*, 34–39. https://doi.org/10.1016/j.smhs.2021.02.001
25. Vijaya, S. Personalized drug-disease prediction using multiple linear regression with ReLU. *J Phys: Conference Series Paper* **2021**, *2115*, 012035. https://doi.org/10.1088/1742-6596/2115/1/012035
26 Du, Z.; Yang, H.; Lv, W.J.; Zhang, X.Y.; Zhai, H.L. prediction of the inhibitory concentrations of chloroquine derivatives using deep neyral network models. *J Biomolec Struct Dyn* **2021**, *39*, 672-680. https://doi.org/10.1080/07391102.2020.1714486
27 dara, S.; Dhamercherla, S.; Jadav, S.S.; Babu, C.M.; Ahsan, M.J. Machine learning in drug discovery: A review. *Artif Intell Rev* **2022**,*55*, 1947-1999. https://doi.org/10.1007/s10462-021-10058-4
28. Zhang, H.; Guo, J.; Li, H.; Guan, Y. Machine learning for artemisinin resistance in malaria treatment across in vivo-in vitro platforms. *ISCIENCE* **2022**, *25*, 103910. https://doi.org/10.1016/j.isci.2022.103910
29 El Rhabori, S.; El Aissouq, A.; Chitta, S.; Khali, F. QSAR, molecular docking and ADMET studies of quinoline, isoquinoline and quinazoline derivatives against *Plasmodium falciparum* malaria. *Struct Chem* **2022**, *34*, 585-603. https://doi.org/10.1007/s11224-022-01988-y
30. Manirakiza, G.; Kassaza, K.; Taremwa, I.M.; Bazira, J.; Byarugaba, F. Molecular identification and antimalarial drug resistance profile of *Plasmodium falciparum* from patients attending Kisoro Hospital, southwestern Uganda. *Malaria J* **2022**, *21*, 1–10. https://doi.org/10.1186/s12936-021-04023-3
31 Nema, S.; Verma, K.; Mani, A.; Maurya, N.S.; Tiwari, A.; Bharti, P.K. Identification of potential antimalarial drug candidates targetting falcipain 2 protein of malaria parasite- A computational strategy. *Biotech* **2022**, *11*, 54. https://doi.org/10.3390/biotech11040054
32 Rajguru, T.; Bora, D.; Modi, M.K. Identification of promising inhibitors for *Plasmodium* haemoglobinase falcipain-2, using virtual screening, molecular docking and MD stimulation. *J Molec Strct* **2022**, *1248*, 131427, https://www.x-mol.net/paper/article/1434958529760653312.
33 Otieno, F.; Walekhwa, M. In silico screening of chloroquine analogues for compounds with more affinity for the Plasmodium falciparum chloroquine transporter as potential antimalarial drugs. *F1000Research*, **2022**, *11*, 188. https://doi.org/10.12688/f1000research.108733.1
34 Ling, Y.; Hao, Z.Y.; Liang, D.; Zhang, C.L.; Liu, Y.F.; Wang, Y.. The expanding role of pyridine and dihydropyridine scaffolds in drug design. *Drug Design Dev Ther*, **2021**, *15*, 4289-4338. https://doi.org/10.2147/DDDT.S329547

35 Ettari, R.; Previti, S.; Di Chio, C.; Zappalà, M. Falcipain-2 and Falcipain-3 inhibitors as promising antimalarial agents. *Curr Med Chem* **2021**, *28*, 3010-3031. https://doi.org/10.2174/0929867327666200730215316

36. Afuwape, A.A.; Xu, Y.; Anajemba, J.H.; Srivasta, G. Performance evaluation of secured network traffic classification using machine learning approach. *Comput Standards Interfaces* **2021**, *78*, 103545, https://doi.org/10.1016/j.csi.2021.103545.

37 Natu, M.; Bachute, M.; Gite, S.; Kotecha, K.; Vidyarthi, A. Review on epileptic seizure prediction: machine learning and deep learning approaches. *Computational and Mathematical Methods in Medicine*, **2022**. ID 7751263, 17 . https://doi.org/10.1155/2022/7751263

38. Mswahili, M.E.; Martin, G.L.; Woo, J.; Choi, G.J.; Jeong, Y. Antimalarial drug predictions using molecular descriptors and machine learning against *Plasmodium falciparum*. *Biomoleculrs* **2021**, *11*, 1750–1770. https://doi.org/10.3390/biom11121750

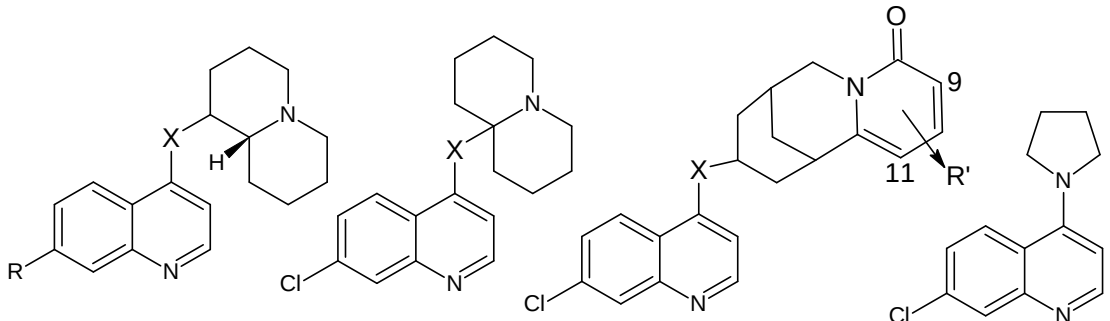
39. Sharma, P.P.; Kumar, S.; Kaushik, K.; Singh, A.; Singh, I.K.; Grishina, M.; Pandey, K.C.; Singh, P.; Potemkin, V.; Poonam, Singh, G.; Rathi, B. In silico validation of novel inhibitors of malarial aspartyl protease, plasmepsin V and antimalarial efficacy prediction. *J Biomolec Struct Dyn*. **2022**, *40*, 8352-8364. https://doi.org/10.1080/07391102.2021.1911855

40. Egieyeh, S.; Syce, J.; Malan, S.; Christoffels, A. Predictive classifier models built from natural products with antimalarial bioactivity using a machine learning approach. *PLoS One* **2018**, *13*, e0204644. https://doi.org/10.1371/journal.pone.0204644

https://doi.org/10.33263/LIANBS132.090

Supplementary materials

Table S1. The chemical data set of CQ analogs used in the study.



1-5, 22-25 **6-8** **9-20, 26, 27** **21**

Compounds	R	X	R´	pIC$_{50}$ *Pf*W-2	pIC$_{50}$ *Pf*D-10
1	Cl	$NHCH_2CH_2SCH_2$	-	6.586030	7.270026
2	Cl	$NH(CH_2)_3SCH_2$	-	6.808549	7.567031
3	Cl	$NHCH(CH_3)CH_2SCH_2$	-	7.083020	7.609065
4	Cl	$NHCH(CH_3)CH_2CH_2SCH_2$	-	6.921181	7.636388
5	CF_3	$NHCH(CH_3)CH_2SCH_2$	-	7.078314	7.251037
6	-	$NH(CH_2)_4$	-	6.323032	7.647817
7	-	$NH(CH_2)_6$	-	6.711527	7.566230
8	-	$NH(CH_2)_8$	-	7.586700	8.045757
9	-	NH	H	6.196679	6.431681
10	-	$NHCH_2CH_2$	H	5.750581	6.026180
11	-	$NH(CH_2)_3$	H	6.524765	6.933674
12	-	$NH(CH_2)_5$	H	7.112946	7.482804
13	-	$NHCH(CH_3)CH_2$	H	5.979390	6.044649
14	-	$NHCH_2CH_2$	9-NO_2	6.058091	6.938171
15	-	$NH(CH_2)_3$	9-NO_2	6.262330	7.202732
16	-	$NH(CH_2)_5$	9-NO_2	7.376751	7.782516
17	-	$NH(CH_2)_3$	9-NH_2	6.320209	7.016825
18	-	$NH(CH_2)_5$	9-NH_2	7.614394	7.924453
19	-	$NH(CH_2)_3$	9,11-Br_2	7.042393	7.323306
20	-	$NH(CH_2)_5$	9,11-Br_2	7.507240	7.774691
21	-	-	-	6.091033	6.001479
22	Cl	SCH_2	-	5.343902	5.143271
23	CF_3	SCH_2	-	5.106793	5.000000
24	Cl	$SCH_2CH_2SCH_2$	-	5.740406	5.617623
25	Cl	$S(CH_2)_3SCH_2$	-	5.848937	5.757707
26	-	$S(CH_2)_3$	H	4.910766	5.006961
27	-	$S(CH_2)_4$	H	5.284833	5.406492

Table S2. Sample of the unprocessed data.

S/N	CQ analogs	*Pf*D-10		*Pf*W-2	
		pIC$_{50}$	Status	pIC$_{50}$	Status
1	Clc1ccc2c(c1)nccc2NCCSC[C@@H]1CCCN2CCCC[C@H]12	7.270	1	6.586	1
2	Clc1ccc2c(c1)nccc2NCCCSC[C@@H]1CCCN2CCCC[C@H]12	7.567	1	6.808	1
3	Clc1ccc2c(c1)nccc2NC(C)CSC[C@H]1CCCN2CCCC[C@@H]12	7.609	1	7.083	1
4	Clc1ccc2c(c1)nccc2NC(C)CCSC[C@@H]1CCCN2CCCC[C@H]12	7.636	1	6.920	1
5	FC(F)(F)c1ccc2c(c1)nccc2NC(C)CSC[C@H]1CCCN2CCCC[C@@H]12	7.251	1	7.078	1
6	Clc1ccc2c(c1)nccc2NCCCCC12CCCCN2CCCC1	7.648	1	6.323	1
7	Clc1ccc2c(c1)nccc2NCCCCCCC12CCCCN2CCCC1	7.566	1	6.711	1
8	Clc1ccc2c(c1)nccc2NCCCCCCCCCC12CCCCN2CCCC1	8.046	1	7.587	1
9	Clc1ccc2c(c1)nccc2NC1CC2CC(C1)CN1C(=O)C=CC=C21	6.432	0	6.197	0
10	Clc1ccc2c(c1)nccc2NCCC1CC2CC(C1)CN1C(=O)C=CC=C21	6.026	0	5.751	0
11	Clc1ccc2c(c1)nccc2NCCCC1CC2CC(C1)CN1C(=O)C=CC=C21	6.934	0	6.525	1
12	Clc1ccc2c(c1)nccc2NCCCCCC1CC2CC(C1)CN1C(=O)C=CC=C21	7.483	1	7.113	1
13	Clc1ccc2c(c1)nccc2NC(C)CC1CC2CC(C1)CN1C(=O)C=CC=C21	6.045	0	5.979	0

14	[O-][N+](=O)C1=CC=C2C3CC(CC(C3)CCNc3ccnc4cc(Cl)ccc43)CN2C1=O	6.938	0	6.058	0
15	[O-][N+](=O)C1=CC=C2C3CC(CC(C3)CCCNc3ccnc4cc(Cl)ccc43)CN2C1=O	7.203	1	6.262	0
16	[O-][N+](=O)C1=CC=C2C3CC(CC(C3)CCCCCNc3ccnc4cc(Cl)ccc43)CN2C1=O	7.783	1	7.376	1
17	Clc1ccc2c(c1)nccc2NCCCC1CC2CC(C1)CN1C(=O)C(N)=CC=C21	7.017	1	6.320	1
18	Clc1ccc2c(c1)nccc2NCCCCCC1CC2CC(C1)CN1C(=O)C(N)=CC=C21'	7.924	1	7.614	1
19	Clc1ccc2c(c1)nccc2NCCCC1CC2CC(C1)CN1C(=O)C(Br)=CC(Br)=C21	7.323	1	7.042	1
20	Clc1ccc2c(c1)nccc2NCCCCCC1CC2CC(C1)CN1C(=O)C(Br)=CC(Br)=C21	7.775	1	7.507	1
21	Clc1ccc2c(ccnc2c1)N1CCCC1	6.001	0	6.091	0
22	Clc1ccc2c(c1)nccc2SC[C@H]1CCCN2CCCC[C@@H]12	5.143	0	5.344	0
23	FC(F)(F)c1ccc2c(c1)nccc2SC[C@H]1CCCN2CCCC[C@@H]12	5.000	0	5.107	0
24	Clc1ccc2c(c1)nccc2SCCSC[C@H]1CCCN2CCCC[C@@H]12	5.618	0	5.740	0
25	Clc1ccc2c(c1)nccc2SCCCSC[C@@H]1CCCN2CCCC[C@H]12	5.758	0	5.849	0
26	Clc1ccc2c(c1)nccc2SCCCC1CC2CC(C1)CN1C(=O)C=CC=C21	5.007	0	4.911	0
27	Clc1ccc2c(c1)nccc2SCCCCC1CC2CC(C1)CN1C(=O)C=CC=C21	5.406	0	5.285	0

S/N = Compound; active = 1; inactive = 0.

Table S3 Parameter settings of ML models.

Model	Setting
Logistic Regression (LR)	Penalty = l2 C = 1e5 Class weight = None Multi_class = auto Solver = liblinear
Naïve Bayes	var_smoothingfloat = 1e-9
k-Nearest Classifier (*k*NNC)	n-neighbours = 5 Metric = Minkowski p = 2
Decision Tree Classifier (DTC)	Random-state = 456
Support Vector Classifier (SVC)	Kernel = linear C = 1 Gamma = auto
Multiple Linear Regression (MLR)	copy_X = True, fit_intercept = True, n_jobs = None, normalize = False

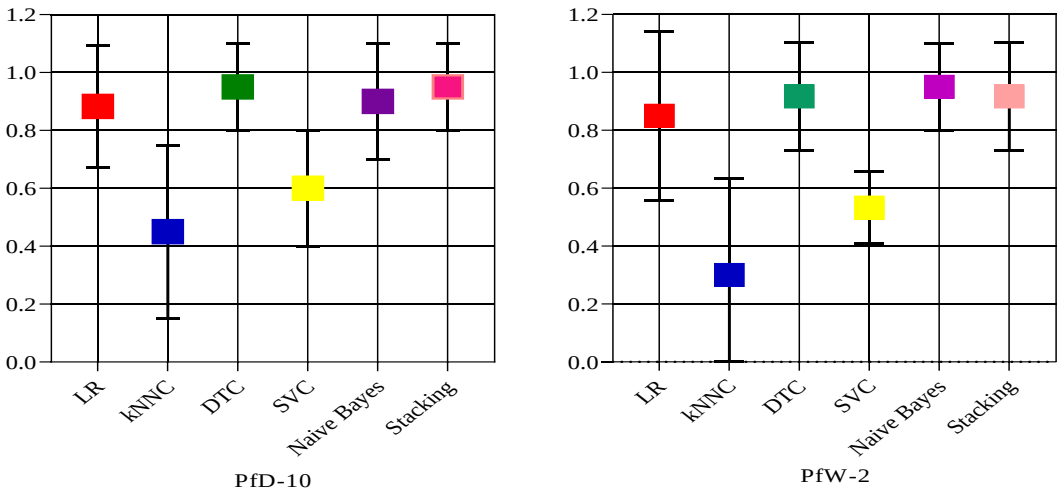


Figure S1. Cross-validation results for the best models. Values are Mean accuracy scores ± SD based on 3 iterations.

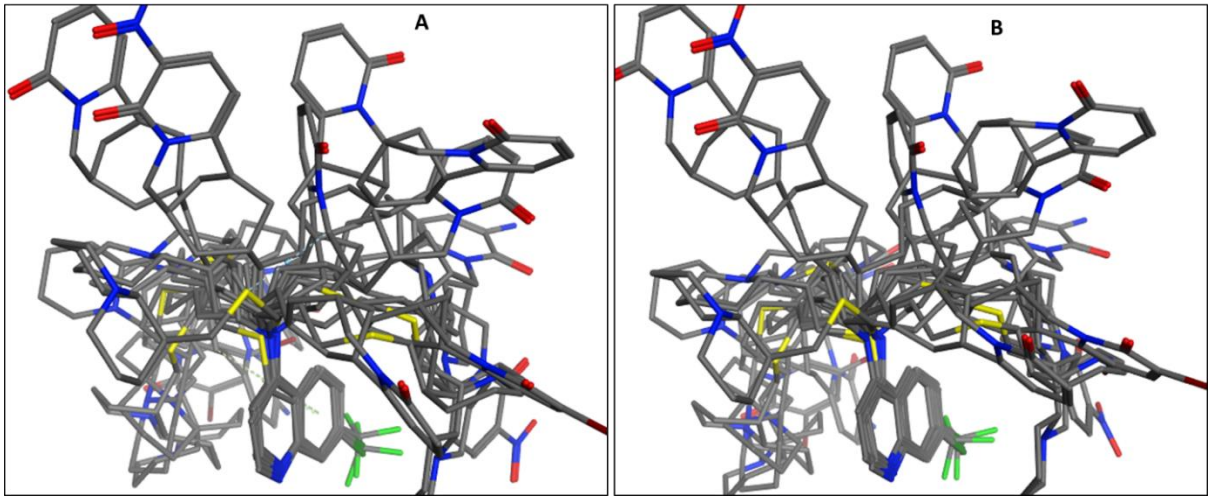


Figure S2. Aligned CQ analogs with activity against *Pf*D-10 (**A**) and *Pf*W-2 (**B**).

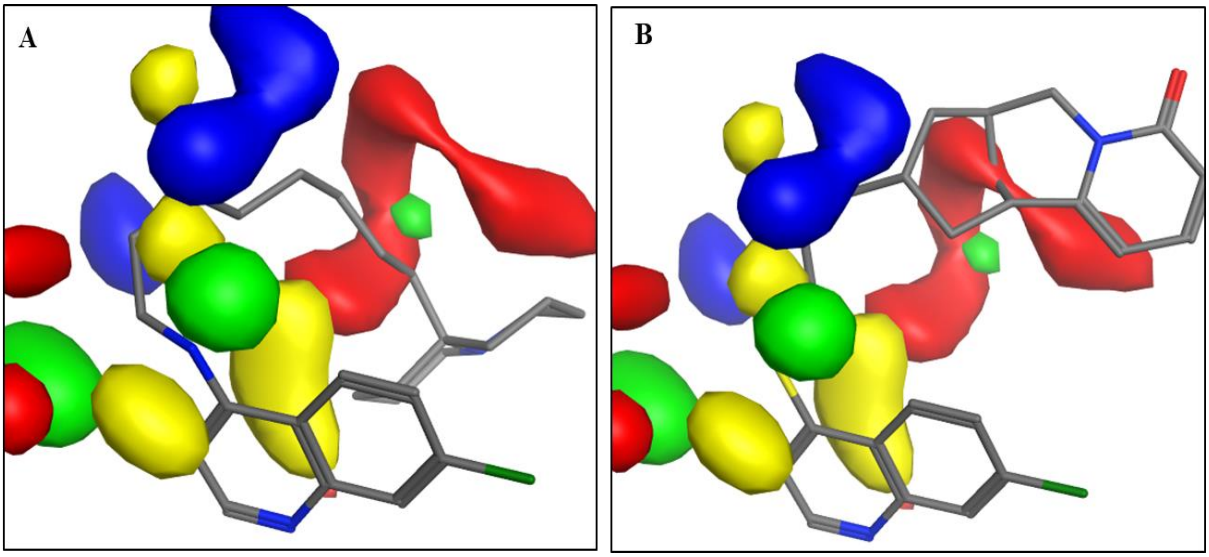


Figure S3. Steric and electrostatic contour maps representing the CoMFA model for activity against *Pf*D-10. *Compounds shown are the strongest (A; **8**) and weakest (B; **23**) anti-P. falciparum activity in the series. Green and yellow indicate regions where steric interactions enhance and reduce activity. Blue and red denote increasing and decreasing electrostatic effects with the positively charged probe.*

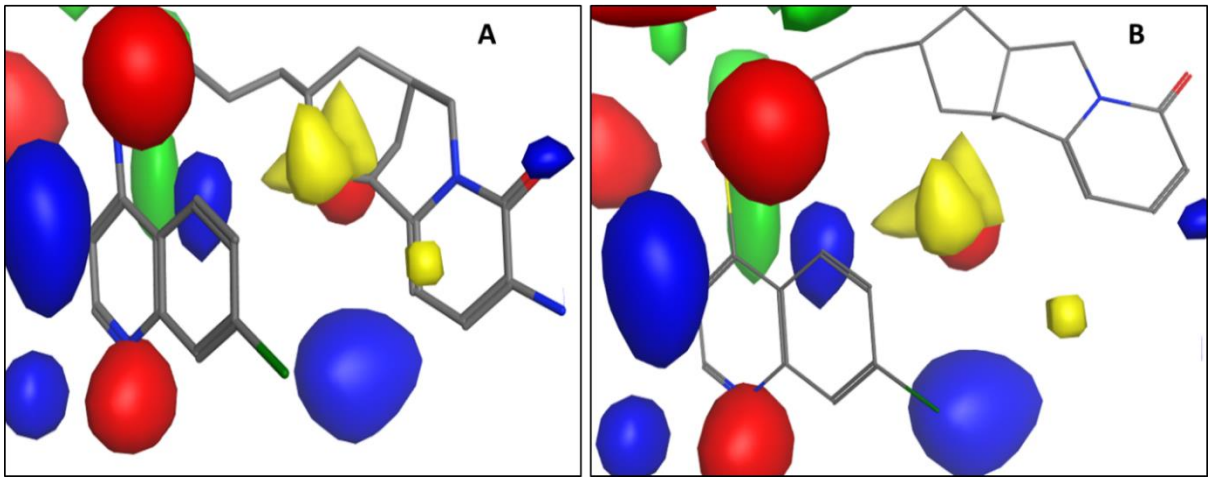


Figure S4. Steric and electrostatic contour maps representing the CoMFA model for *Pf*W-2 *P. falciparum* activity. *Compounds shown are the strongest (A; **18**) and weakest (B; **26**) anti-P. falciparum activity in the series. Green and yellow indicate regions where steric interactions enhance and reduce activity. Blue and red denote increasing and decreasing electrostatic effects with the positively charged probe.*