World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

MACHINE LEARNING-BASED QSAR MODELING OF TRICLOSAN ANALOGS FOR ANTIPLASMODIAL ACTIVITY PREDICTION USING SHRINKAGE REGRESSION MODEL AND TOPOLOGICAL-PHYSICOCHEMICAL DESCRIPTORS

Désiré Mélèdje, Aubin N'guessan*, Kady Silué, Jocelyne Bosson and Logbo Moussé

Fundamental Applied Physics Laboratory (FAPL), Nangui Abrogoua University, Côte d'Ivoire.
* Corresponding Author

ORCID Details
Désiré Mélèdje: https://orcid.org/0009-0008-0089-6595
Aubin N'guessan : https://orcid.org/0000-0002-3043-2985
Jocelyne Bosson: https://orcid.org/0009-0006-6274-3098
Logbo Moussé: https://orcid.org/0009-0002-3024-1457

World Journal of Advanced Research and Reviews, 2026, 31(02), 743–752

Publication history: Received on 29 June 2026; revised on 09 August 2026; accepted on 11 August 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.31.2.2099

ABSTRACT

Drug-resistant *Plasmodium falciparum* continues to threaten malaria control, necessitating new antimalarial discovery strategies. QSAR modeling with machine learning offers a cost-effective approach to relate molecular features to biological activity and prioritize candidate compounds for further development. In particular, topological descriptors (capturing molecular connectivity, branching, and structural arrangement) and physicochemical descriptors (such as hydrophobicity, electronic properties, polarity, and steric effects) provide complementary and mechanistically meaningful information governing ligand–target interactions and biological response. Here, a ridge regression QSAR model was developed to predict the anti-plasmodial activity of triclosan analogs. RFECV (Recursive Feature Elimination with Cross-Validation) was used for optimal descriptor selection. The model achieved an R^2 of 88.83% on the training set and an R^2_test of 80.80% on the external test set, indicating strong predictive performance and robust generalization. These results highlight the importance of carefully selected descriptors in improving both interpretability and accuracy. Overall, the RFECV–Ridge QSAR framework provides a reliable and interpretable approach for virtual screening and rational design of novel *pf*ENR inhibitors.

Keywords: Shrinkage Model; Rigde Regression; Triclosan; Topological-Physicochemical Descriptors; Machine Learning

1 INTRODUCTION

Malaria remains a major parasitic disease that continues to threaten the lives of millions of people in tropical and subtropical regions. According to the World Health Organisation (WHO), there were an estimated 282 million cases of malaria and 610,000 deaths from the disease worldwide in 2024, spread across 80 endemic countries [1]. Children

under five are particularly vulnerable to severe malaria and its complications; in the WHO African Region, they accounted for nearly 75 per cent of malaria-related deaths in 2024.

Malaria is caused by protozoan parasites of the *Plasmodium* genus. Six *Plasmodium* species are known to infect humans (*P. falciparum*, *P. vivax*, *P. malariae*, *P. knowlesi*, *P. ovale curtisi*, and *P. ovale wallikeri*) among which *P. falciparum (Pf)* is responsible for the most severe and virulent form of the disease [2]. A major challenge in malaria control is the emergence and spread of drug resistance in *Plasmodium* parasites, particularly *P. falciparum*. This resistance compromises the efficacy of antimalarial treatment and poses a serious threat to current control and elimination strategies. Several factors contribute to its development, including the circulation of counterfeit or substandard drugs, inadequate therapeutic exposure, the high mutation rate of the parasite, and the selective pressure exerted by improper or prolonged drug use [3]. In addition, the reduced sensitivity of *P. falciparum* to first-line artemisinin-based combination therapies (ACTs), together with the declining effectiveness of insecticides used in vector control, further undermines malaria intervention efforts. The situation is all the more complex given that the protection offered by the recently developed malaria vaccines is limited and variable [3-4].

These alarming figures, together with the continued emergence of drug-resistant *P. falciparum* strains, underscore the urgent need for the design and development of efficient next-generation antimalarial drugs and vaccines, particularly for vulnerable populations such as pregnant women and children. In this context, numerous therapeutic strategies have focused on identifying key enzymes involved in different stages of the *P. falciparum* developmental cycle and developing innovative and selective inhibitors against these targets. Among the strategies explored, triclosan and its derivatives have received significant attention due to their inhibitory action against *P. falciparum* 2-trans-enoyl-ACP reductase (*Pf*ENR) [5].

The *Pf*ENR protein plays a crucial role in the type II fatty acid synthesis (FAS-II) pathway of *P. falciparum*, which is essential for the survival of parasite. Long-chain fatty acids are key metabolic precursors for the formation of biological membranes, apicoplast development and energy metabolism [6-7]. Given that the parasite relies on the FAS-II pathway whereas human cells primarily utilize the type I fatty acid biosynthesis (FAS-I) system, which has a different structure, inhibition of *Pf*ENR has emerged as a promising therapeutic strategy. Triclosan (Figure 1) is one of the most extensively studied inhibitors of *Pf*ENR. This widely used antibacterial agent has been shown to inhibit the *in vitro* growth of *P. falciparum*. Biochemical studies have demonstrated that triclosan acts as an uncompetitive inhibitor of purified *Pf*ENR resulting in significant anti-plasmodial activity against cultured parasites.

Identifying the structural determinants underlying the anti-plasmodial activity of triclosan analogs remains a significant challenge, particularly when available datasets are relatively small and exhibit substantial descriptor redundancy. To overcome these limitations and improve the predictive performance of QSAR models, machine learning (ML) techniques have been increasingly adopted [5]. However, ML methods relying on polynomial regression or high-degree multilinear models (HDMLM) face several inherent limitations that can compromise predictive performance. These limitations include (i) overfitting, which becomes more severe as the polynomial degree increases; (ii) multi-collinearity arising from polynomial feature expansion; (iii) reduced model interpretability, making it difficult to evaluate the contribution of individual descriptors, contrary to the OECD principles for QSAR model development; and (iv) numerical instability, as high-order polynomial terms may produce excessively large values, leading to numerical conditioning problems and unreliable coefficient estimates.

Accordingly, this study investigated the structural determinants of the anti-plasmodial activity of triclosan analogs by developing QSAR models based on shrinkage regression techniques. Shrinkage regression encompasses a family of penalized regression techniques that improve model estimation and predictive performance by regularizing regression coefficients. Unlike ordinary least squares (OLS) and high-degree multilinear (HDML) models, which are often prone to overfitting, multi-collinearity, and poor predictive performance, shrinkage methods impose a penalty on the regression coefficients to produce more robust and parsimonious models [7].

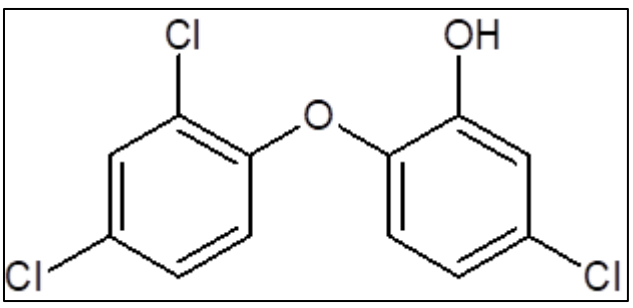

Figure 1: Triclosan

2 MATERIAL AND METHODS

2.1 Dataset preparation

A dataset comprising 108 triclosan analogs was used to develop QSAR model based on shrinkage (penalized) regression techniques [5]. The complete dataset was obtained from the study of Guimarães *et al.*, who calculated 777 molecular descriptors for each compound using the freely available Mold² software [9]. Consequently, the final matrix consisted of 108 compounds described by 777 molecular descriptors. The biological activity of each compound, expressed as the half-maximal effective concentration (EC_{50}) against the 3D7 strain of *Plasmodium falciparum*, was retrieved from PubChem and transformed into its logarithmic form, ($pEC_{50} = -\log10(EC_{50})$), which was used as the dependent variable. The 777 molecular descriptors served as the independent variables in the QSAR modeling.

2.2 Feature ranking and selection

At this stage, recursive feature elimination with cross-validation (RFECV) was employed to identify and remove redundant or weakly informative descriptors, thereby retaining only the variables that contributed significantly to the QSAR models [10]. RFECV is a wrapper-based feature selection method that evaluates subsets of descriptors by iteratively training and validating a machine learning model using Leave N-Out (LNO) with N=5 cross-validation procedure. In the present study, shrinkage regression model was adopted as the wrapper estimator within the RFECV framework because of its robustness in handling multi-collinearity among molecular descriptors and its ability to provide stable estimates of descriptor importance. The RFECV algorithm begins by fitting the ridge regression model using the complete set of molecular descriptors. The importance of each descriptor is then assessed according to the fitted model, after which the least informative descriptor is eliminated. The model is subsequently rebuilt using the remaining descriptors, and the descriptor importance is re-evaluated. This recursive elimination process continues until the optimal subset of descriptors is identified. At each iteration, model performance is assessed by LNO cross-validation procedure, and the subset yielding the highest cross-validated predictive performance is selected as the final descriptor set for QSAR model development [11].

2.3 Development of shrinkage regression method

In usual multiple linear regression model, given the data (x^i, y_i), $i = 1, \ldots, N$, where $x^i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$ are the independent variables and y_i are dependent variables, the multiple linear regression model has the form:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \varepsilon_i \tag{1}$$

Where β_j are the parameters and ε_i, the error term.

In matrix notation**:**

$$Y = X\beta + \epsilon \tag{2}$$

The model parameters are estimated from the training data, resulting in the regression coefficient vector $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ with the vector error $\epsilon = (\varepsilon_1, \ldots, \varepsilon_N)^{\mathrm{T}}$. The most popular estimation procedure is the Ordinary Least Squares (OLS) method, which determines the coefficient estimates by minimizing the residual sum of squares (RSS), defined as the squared differences between the observed and predicted response values (Eq. 3).

$$RSS(\beta) = \sum_{i=1}^{N} \left(y_i - \hat{y}_i(\beta)\right) \tag{3}$$

$$= \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 \tag{4}$$

In matrix form:

$$RSS(\beta) = \|Y - X\beta\|^2 \tag{5}$$

Multiple linear regression (MLR) is a statistical technique used to model the relationship between a dependent variable (response) and several independent variables (predictors). By incorporating all predictors into a single model, MLR

estimates the independent effect of each predictor whilst taking into account the influence of the others. This approach provides a more accurate and comprehensive understanding of the relationship between individual predictors and the response variable, whilst quantifying their combined effect and highlighting potential associations between the predictors themselves [12]. Despite its widespread application, the Ordinary Least Squares (OLS) estimator is often inadequate for QSAR datasets, which are typically characterized by high-dimensional and highly correlated molecular descriptors. In such settings, OLS may produce unstable regression coefficients and poor predictive performance. These limitations have led to the development of shrinkage (penalized) regression methods, which regularize coefficient estimates by imposing a penalty on their magnitude, thereby improving model stability, reducing overfitting, and enhancing predictive performance.

In ridge regression, a shrinkage estimator of the regression coefficients is obtained by solving the following penalized optimization problem:

$$min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \alpha \sum_{j=1}^{p} \|\beta_j\|^2 \qquad (6)$$

The tuning parameter α ($\alpha \geq 0$) governs the strength of the regularization penalty and consequently the extent of coefficient shrinkage. Larger values of α produce greater coefficient shrinkage whereas smaller values yield estimates closer to those obtained by ordinary least squares (OLS) [8-9].

The robustness and predictive performance of the shrinkage regression models were evaluated using complementary internal and external validation procedures. Internal validation was conducted on the training set to assess model stability and fitting accuracy whereas external validation was performed on the independent test set to evaluate the predictive ability and generalization performance of the model. In this study, the dataset was randomly partitioned into training and test sets using the *train_test_split* function from the Scikit-learn library implemented in Python (version 3.9.2). The *test_size* parameter was set to 0.20, resulting in an 80:20 split, with 80% of the data used for model training and internal validation and the remaining 20% reserved as an independent test set for external validation [13]. Data shuffling was enabled prior to partitioning to ensure that the samples were distributed randomly between the two subsets.

2.4 Model fitting and coefficient analysis

To assess the stability of the regression coefficients, the complete ridge coefficient path was examined as a function of the regularization parameter α. The optimal value of α was determined using a five-fold cross-validation procedure, ensuring an appropriate balance between model complexity and predictive performance. The coefficient path provides a graphical representation of how the estimated regression coefficients evolve with increasing regularization strength thereby allowing the identification of stable molecular descriptors [7,12].

2.5 Regression model assessment and validation

The performance assessment of the QSAR models developed for triclosan analogs using shrinkage regression method was carried out using a comprehensive set of statistical metrics under both internal and external validation schemes. In this work, the goodness-of-fit (GOF) of the developed QSAR models was evaluated using the coefficient of multiple determination (R^2), which quantifies the proportion of the variance in the response variable explained by the regression model (Eq.7) . The R^2 value was computed by comparing the observed and predicted response values for the training dataset. Although a high R^2 value (approaching 1.0) indicates that the model adequately fits the training data, it does not guarantee robust predictive performance on unseen data. Therefore, external validation is essential to assess the model's predictive ability. To summarize the overall errors of the developed QSAR models under both internal and external validation schemes, two widely used error metrics (Eq. 8-9), namely, the root mean square error (RMSE) and the mean absolute error (MAE) were calculated.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2} \qquad (7)$$

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \qquad (8)$$

$$RMSE = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)}{N}} \qquad (9)$$

In this study, cross-validation (CV) was employed as a robust internal validation strategy to evaluate the stability of the developed QSAR regression models. Leave-One-Out cross-validation (LOO-CV) is an internal validation procedure in which N regression models are successively constructed by excluding one molecule at a time from the training dataset [14]. At each iteration, the observed response of excluded molecule is predicted using the model developed from the remaining N−1 molecules, and the resulting prediction errors are aggregated to assess the robustness and internal predictive performance of the QSAR model. The predictive performance obtained through LOO-CV is commonly quantified using the cross-validated coefficient of determination Q^2_{LOO}, the root mean square error of cross-validation ($RMSE_{CV}$), and the mean absolute error of cross-validation (MAE_{CV})[15].

$$Q^2_{LOO} = 1 - \frac{\sum_i \left(y_i - \hat{y}_{i/i}\right)^2}{\sum_i (y_i - \bar{y}_i)^2} \quad (10)$$

$$MAE_{CV} = \frac{1}{N}\sum_{i=1}^{N} \left|y_i - \hat{y}_{i/i}\right| \quad (11)$$

$$RMSE_{CV} = \sqrt{\frac{\sum_i \left(y_i - \hat{y}_{i/i}\right)}{N}} \quad (12)$$

where $\hat{y}_{i/i}$ are the predicted response of the ith molecular observation estimated using a model calibrated on the remaining N−1 molecular observations. Conversely, when more than one molecule is recursively excluded at each iteration from the training dataset, the procedure is referred to as Leave-N-Out (LNO) cross-validation.

3 RESULTS AND DISCUSSION

3.1 Feature selected

Feature selection algorithm was employed to rank molecular descriptors based on their predictive relationship with the response variable (pEC_{50}), thereby identifying the most informative descriptors for subsequent QSAR model development. At this stage, a ridge regression model, with a regularization parameter of α = 1.0, was used as the wrapper estimator within the RFECV framework because of its robustness in handling multi-collinearity among molecular descriptors and its ability to provide stable estimates of descriptor importance. Among all evaluated descriptor subsets, the 19-descriptor model exhibited the best predictive performance achieving a cross-validated coefficient of determination $Q^2_{LNO} = 0.56$. Table 1 summarizes the complete set of molecular descriptors obtained and used in the present study.

Table 1: List of Mold2 descriptors used in this study

N°	Code	Mold2 descriptors
1	D170	First No-Zero eigenvalue of Laplacian matrix
2	D195	Maximal valence vertex electrotopological positive variation
3	D274	Information content order-5 index
4	D294	Bond information content order-1 index
5	D361	Ratio of convention bonds with total path counts
6	D367	Sum of topological distance between the vertices N and P
7	D375	Sum of topological distance between the vertices O and S
8	D451	Geary topological structure autocorrelation length-5 weighted by atomic masses
9	D483	Moran topological structure autocorrelation length-5 weighted by atomic masses
10	D486	Moran topological structure autocorrelation length-8 weighted by atomic masses
11	D496	Moran topological structure autocorrelation length-2 weighted by atomic Sanderson electronegativities
12	D519	Molecular topological order-9 charge index
13	D582	Hightest eignevalue from Burdex matrix weighteed by electronegativities Sanderson-Scale order-3

14	D594	Hightest eignevalue from Burdex matrix weighteed by polarizabilities order-7
15	D607	Number of tertiary C-sp2
16	D650	Number of group secondary amines (aromatic)
17	D662	Number of group ammonium groups (aromatic)
18	D738	Number of group R~CR~R
19	D748	Number of group H attached to C0(sp3) with 1X attached to next C

Because molecular structures of triclosan analogs cannot be directly incorporated into conventional regression models, they are represented by molecular descriptors that quantitatively encode their structural and physicochemical properties. These descriptors serve as numerical input variables for the development of QSAR model. As shown in Table 1, topological and physicochemical descriptors constitute the most informative subset of descriptors for modeling the anti-plasmodial activity of triclosan analogs. Topological descriptors are two-dimensional (2D) molecular descriptors derived from the molecular graph that encode atomic connectivity and bond-related information, such as the number of single, double, aromatic, and rotatable bonds etc. In the present study, the selected topological descriptors were **D274, D294, D361, D367, D375, D451, D483, D486, D497, D519, D607, D650, D662, D738** and **D748** indicating that topological characteristics play a major role in determining the anti-plasmodial activity of triclosan analogs. The second category comprises physicochemical descriptors, which are calculated from the two-dimensional (2D) molecular structure of a compound. Unlike one-dimensional (1D) descriptors, which are derived solely from the molecular formula, physicochemical descriptors encode molecular properties related to the chemical environment and atomic interactions. In the present study, the selected physicochemical descriptors were **D582, D594,** and **D946**, which are associated with molecular electronegativity and polarizability [9]. These properties play a key role in governing intermolecular interactions and may therefore contribute significantly to the anti-plasmodial activity of triclosan analogs.

3.2 Stability of the regression coefficients

Since the solution of equation (6) depends on the regularization parameter α, this parameter must be optimized through a hyper-parameter tuning procedure. In the present study, the optimal value of α was determined using five-fold cross-validation. Figure 2 illustrates the evolution of the non-zero regression coefficients over the large range of alpha values. As illustrated in Figure 2, five-fold cross-validation identified α = 1.459 as the optimal regularization parameter, corresponding to the best compromise between model complexity and predictive performance.

The ridge coefficient path provides a graphical representation of the influence of regularization on the estimated coefficients [16]. On the left-hand side of the plot, α is close to zero, indicating that the corresponding ridge regression coefficient estimates are nearly identical to the ordinary least-squares (OLS) estimates. As α increases toward the right-hand side of the plot, the ridge regression coefficient estimates progressively shrink toward zero. Overall, the regression coefficients remain relatively stable across the range of regularization parameters, as none of them rapidly shrinks toward zero, indicating sustained contributions to the prediction of the anti-plasmodial activity of triclosan analogs. However, four descriptors, namely **D270, D274, D582** and **D594** exhibit changes in the sign of their regression coefficients as the regularization parameter increases. Such sign reversals, together with the presence of crossing coefficient paths, may reflect multi-collinearity among descriptors and competition between variables encoding similar structural or physicochemical information. For example, **D270** and **D274** are information-content descriptors derived from molecular graph theory, where the molecular structure is represented as a graph and characterized by probability distributions. Information-theoretic concepts are then applied to quantify the structural complexity and information content of the molecule [10,17]. Similarly, descriptors **D582** and **D594**, representing electronegativity and polarizability, show sign reversals as regularization increases, indicating overlapping information and redistribution of their contributions under stronger regularization.

The **D195** and **D519** standardized coefficients follow almost identical trajectories, indicating a high degree of collinearity between these two descriptors, which encode similar physicochemical information relating to the charge topology of triclosan analogs [10]. Beyond the optimal regularization parameter, **D519**, which converges toward zero more rapidly, appears to be less stable under regularization suggesting that it is largely redundant with **D195** and contributes progressively less to the QSAR model as the regularization strength increases.

3.3 Internal validation statistical metrics

In this study, ridge regression model was evaluated for predicting the anti-plasmodial activity of triclosan analogs. Prior to model development, the dataset was randomly partitioned into training and test subsets using the *train–test_split* algorithm. The training subset was used to build the regression model, whereas the test subset was reserved for external

validation. Of the 108 molecular observations, 86 triclosan analogs (80% of the dataset) were assigned to the training set, while the remaining 22 compounds (20%) constituted the test set. The ridge regression model, optimized using the best regularization parameter, achieved satisfactory statistical performance and exhibited good predictive ability for the anti-plasmodial activity of triclosan analogs. Internal validation statistical parameters Q^2_{LOO} and Q^2_{LNO} of 78.01% and 74.19% respectively, confirm the robustness and predictive stability of the developed ridge regression model.

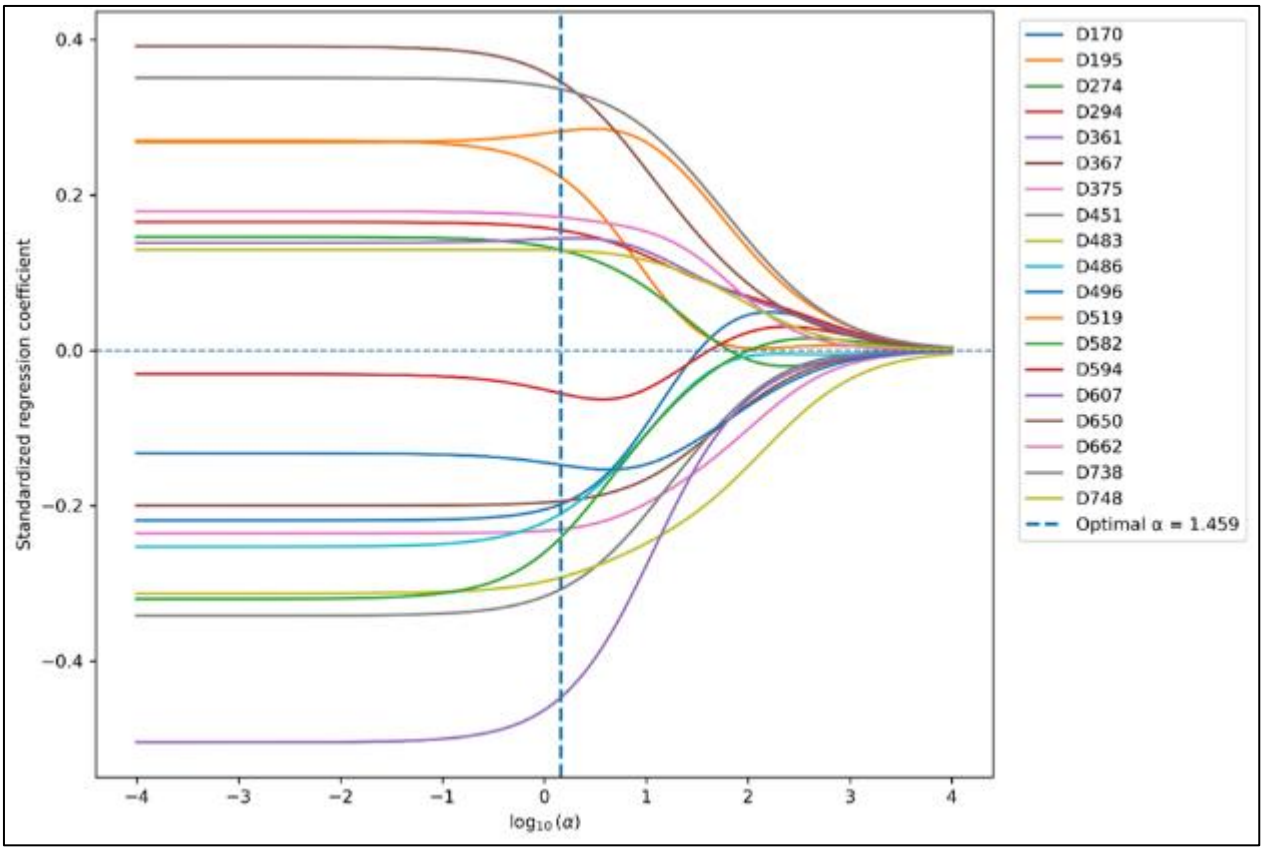

Figure 2: Standardized ridge regression coefficients displayed as function of α.

Table 2: Statistical parameters of machine learning based ridge regression model

Metrics	Values
R^2(%)	88.72%
Q^2_{LOO}(%)	78.01%
Q^2_{LNO}(%)	74.19%
MAE	0.248
RMSE	0.302
MAE^{LOO}_{CV}	0.334
$RMSE^{LOO}_{CV}$	0.422
MAE^{LNO}_{CV}	0.345
$RMSE^{LNO}_{CV}$	0.437

The relatively small difference between these two validation metrics indicates that the model is stable with respect to the validation procedure and suggests that the estimated regression coefficients are robust, exhibiting limited sensitivity to variations in the training data. The ridge regression model developed for predicting the pEC_{50} values of triclosan analogs achieved a coefficient of determination of (R^2 = 88.72%), indicating a strong agreement between the

experimental and predicted activities. The model also exhibited low prediction errors, with a mean absolute error (MAE) of 0.248 and a root mean square error (RMSE) of 0.302. In addition to these global error estimates, cross-validated metrics were also computed, namely the mean absolute error of cross-validation (MAE_{CV}) and the root mean square error of cross-validation ($RMSE_{CV}$), which provide a more stringent assessment of the model's predictive performance on unseen data. The comparison between MAE/RMSE and MAE_{CV}/$RMSE_{CV}$ allows for evaluating the stability and generalization ability of the model, with only limited deviations observed, further confirming its robustness. These error metrics, computed through the internal validation procedure, demonstrate the good predictive accuracy of the developed QSAR model. Furthermore, the relatively low values of both standard and cross-validated errors indicate that the model generates reliable and consistent predictions during cross-validation. Table 2 summarizes the internal validation metrics obtained for the developed shrinkage regression QSAR model.

3.4 External validation of regression model

The predictive performance of the developed regression model was assessed using an external test set by comparing the predicted and experimental anti-plasmodial activities of triclosan analogs that were not included in the model development (Figure 3). In the present study, the external test set comprised 20% of the complete dataset. The optimal ridge regression model, obtained with a regularization parameter of α = 1.459, achieved an external coefficient of determination of $R^2_{test} = 80.53\,\%$ with $MAE_{test} = 0.331$ and $RMSE_{test} = 0.424$. These external validation metrics demonstrate the good predictive performance and generalization ability of the developed QSAR model. The practical value of predictive QSAR model lies primarily in their ability to accurately predict the biological activities of untested compounds. As we can see, the coefficient of determination obtained for the external test set is high and close to unity, indicating that the developed ridge regression model possesses good predictive performance. Furthermore, the low values of the average prediction error metrics, particularly in the test set, indicate a close agreement between the predicted and experimental pEC_{50} values, confirming the high predictive accuracy and strong generalization ability of the developed QSAR model [18]. These results demonstrate the good generalization ability of the model and support its applicability for predicting the anti-plasmodial activity of structurally related triclosan analogs.

3.5 Mechanistic interpretation

The analytic solution that represents the model equation is presented:

$$pEC_{50} = 5.273 - 0.198D170 + 0.223D195 + 0.129D274 + 0.155D294 + 0.144D361 - 0.194D367 - 0.230D375 - 0.307D451 - 0.292D483 - 0.210D486 - 0.147D496 + 0.282D519 - 0.241D582 - 0.055D594 - 0.447D607 + 0.346D650 + 0.171D662 + 0.336D738 + 0.129D748 \quad [13]$$

Descriptors **D650** and **D738** exhibit the strongest positive contributions to the anti-plasmodial activity of triclosan analogs. **D738** corresponds to the number of R~CR~R groups, representing an aliphatic carbon atom substituted by two carbon or hydrocarbon groups (i.e., a branched carbon center or related sub-structural fragment) [9]. The positive contribution of this descriptor suggests that hydrophobic substructures and branched aliphatic carbon environments enhance the anti-plasmodial activity of triclosan analogs, likely by promoting favorable hydrophobic interactions with the *pf*ENR binding site [7]. The second descriptor exhibiting a strong positive contribution to the response variable is **D650**, which represents nitrogen-containing aliphatic and aromatic functional groups. Its positive coefficient suggests that the presence of nitrogen-containing moieties, particularly protonatable amine groups, may enhance the anti-plasmodial activity of triclosan analogs by promoting favorable interactions with the NAD+ cofactor and/or surrounding amino acid residues or solvent molecules. These interactions may contribute to stabilizing the ligand–enzyme complex and improving binding affinity. Physicochemical descriptors **D496**, **D582**, and **D594**, which encode the distribution of molecular electronegativity and polarizability in triclosan analogs, exhibit negative regression coefficients. This finding suggests that increasing the contribution of polar or highly polarizable substituents is associated with a decrease in the predicted pEC_{50} values. Consequently, excessive molecular polarity may be unfavorable for anti-plasmodial activity, possibly because it weakens hydrophobic interactions within the enzyme binding pocket or reduces the affinity of triclosan analogs for the target protein.

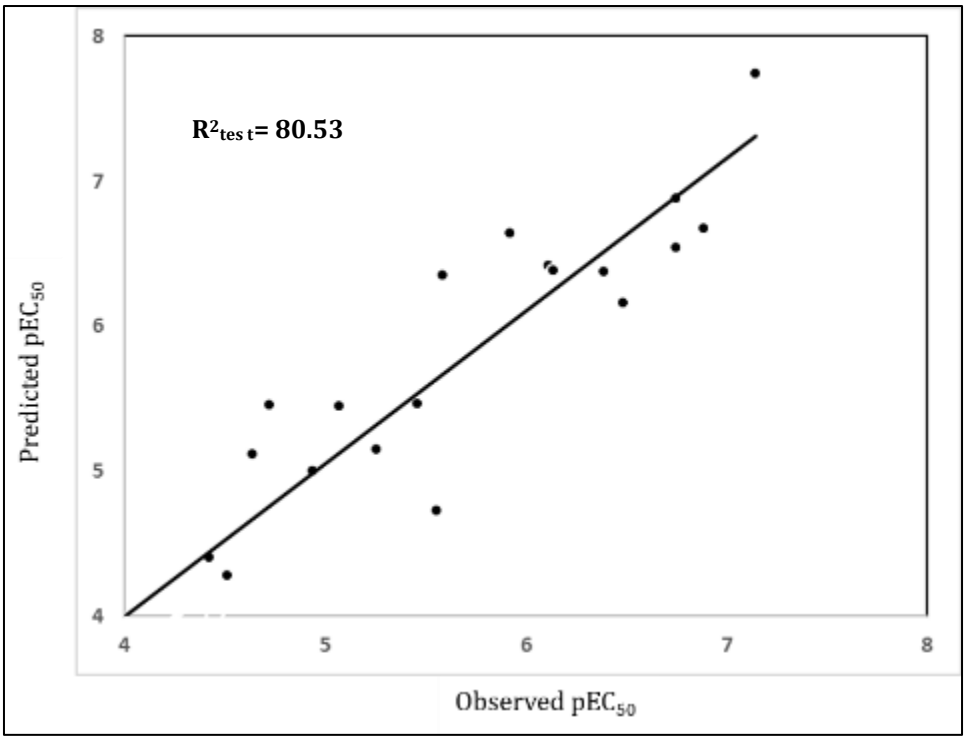

Figure 3: Correlation plot between the observed and predicted values on test set

4 CONCLUSION

In this study, a QSAR model was developed for triclosan analogs to predict the anti-plasmodial activity of potential *pf*ENR inhibitors and thereby facilitate the filling of experimental data gaps. *Pf*ENR is an attractive therapeutic target because its inhibition represents a promising strategy for overcoming drug resistance in *Plasmodium falciparum*, the most virulent human malaria parasite. A shrinkage regression approach based on ridge regression was employed to construct the QSAR model. Using the optimal regularization parameter selected by cross-validation, a final model incorporating 19 molecular descriptors was obtained. The developed model exhibited good predictive performance and strong generalization ability on unseen compounds. Compared with the regression model previously reported by Guimarães and co-workers, the proposed model provides greater numerical stability and more reliable coefficient estimates owing to the use of regularization. Overall, this study demonstrates that combining recursive feature elimination with regularized QSAR modeling constitutes a robust and interpretable framework for predicting the anti-plasmodial activity of triclosan analogs and supports its application in the virtual screening and rational design of novel *pf*ENR inhibitors.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Disclosure of conflict of interest

No conflict of interest to be disclosed.

Statement of informed consent

Informed consent was obtained from all individual participants included in the study.

REFERENCES

[1] World malaria report 2025. Available https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2025

[2] Shekhar, Chowdhary S., Mosnier J., Fonta I., Pradines B and Kumar V. Design, synthesis and mechanistic insights into triclosan derived dimers as potential anti-plasmodials.RSC Medicinal Chemistry. 2025;16(2): 437-972.doi:10.1039/d4md00494a

[3] Pandey V., J. F. Kennedy, and Neera Raghav. Falcipain-2: A review on structurally diverse non-peptide inhibitors. International Journal of Biological Macromolecules. 2025; 309: 142817. doi: 10.1016/j.ijbiomac.2025.142817

[4] Alberca L. N., Chuguransky S. R., Álvarez C.L., Talevi A., Salas-Sarduy E. . In silico Guided Drug Repurposing: Discovery of New Competitive and Non-competitive Inhibitors of Falcipain-2. Frontiers in Chemistry. 2019 ;7, doi:10.3389/fchem.2019.00534

[5] de Souza Guimarães R.M., Vieira I.H.P., Zanchi F.B., Caceres R.A., Zanchi, F.B. Application of Machine Learning in the Development of Fourth Degree Quantitative Structure–Activity Relationship Model for Triclosan Analogs Tested against Plasmodium falciparum 3D7. 2024; 9(44): 44436-44447. doi:10.1021/acsomega.4c05768

[6] Ralph Stuart A., et al. Metabolic maps and functions of the Plasmodium falciparum apicoplast. Nature Reviews Microbiology. 2004, 2(3):203-216. doi:10.1038/nrmicro843

[7] Bieri C., et al. Structure-based design and pharmacophore-based virtual screening of combinatorial library of triclosan analogues active against enoyl-acyl carrier protein reductase of Plasmodium falciparum with favourable ADME profiles. Int. J. Mol. Sci. 2023; 24(8):6916. doi:10.3390/ijms24086916

[8] Hastie T., Robert T., Jerome F.. The elements of statistical learning. Springer series in statistics New-York. 2009. doi:10.1007/978-0-387-84858-7

[9] Hong, H., Xie, Q., Ge, W., Qian, F., Fang, F., Shi, L., Su, Z., Perkins, R., Tong, W. Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. Journal of Chemical Information and Modeling. 2008.48(7):1337–1344. doi:10.1021/ci800038f

[10] Kuhn M., Kjell J., Aughan, D. Applied predictive modeling in R. Springer New York. 2013, 26. doi:10.1007/978-1-4614-6849-3

[11] Misra, Puneet, and Arun Singh Yadav. Improving the classification accuracy using recursive feature elimination with cross-validation. International Journal on Emerging Technologies. 2020, 11(3): 659-665.

[12] Emmert-Streib Frank, and Matthias Dehmer. High-dimensional LASSO-based computational regression models: regularization, shrinkage, and selection. Machine Learning and Knowledge Extraction. 20191,1(1): 359-383. doi:10.3390/make1010021

[13] Pedregosa F. Scikit-learn: Machine learning in python. J Mac Learn Res. 2011, 12: 2825-2830.

[14] Veerasamy R., Rajak H., Jain A., Sivadasan S., Varghese C.P., Agrawal, R. K. Validation of QSAR Models - Strategies and Importance. International Journal of Drug Design and Discovery. 2011 2(3): 511–519. doi:10.37285/ijddd.2.3.

[15] Gramatica, Paola. On the development and validation of QSAR models. Computational Toxicology: 2012, 2:499-526. doi:10.1007/978-1-62703-059-5_21

[16] James G., Witten D., Hastie T., Tibshirani RAn introduction to statistical learning: with applications in R. Springer.New York. 2021, 2. Doi:10.1007/978-1-4614-7138-7

[17] Leonis G., Melagraki G., Afantitis, A. Open source chemoinformatics software including KNIME analytics. Handbook of computational chemistry. 2016. 9:2201–2230. doi: 10.1007/978-3-319-27282-5_57

[18] Willmott C.J., Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research. 2005.30:79-82. doi:10.3354/cr030079.