

A Robust SMILES-Based Prediction of Aqueous Solubility of Diverse Antiplasmodial Compounds using Machine Learning Algorithms

Thecla O. Ayoka^{1,4,*} , 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; thecla.ayoka@unn.edu.ng (T.O.A.);

² Department of Pharmaceutical and Medicinal Chemistry, Faculty of Pharmaceutical Sciences, University of Nigeria, Nsukka, 410001, Enugu State, Nigeria; charles.nnadi@unn.edu.ng (C.O.N.);

³ Department of Pharmaceutical Chemistry, Faculty of Pharmacy, Madonna University, PMB 05 Elele Campus, Rivers State, Nigeria

⁴ Department of Biochemistry, School of Biological Sciences, Federal University of Technology, Owerri, 460114, Imo State, Nigeria

* Correspondence: thecla.ayoka@unn.edu.ng (T.O.A.); charles.nnadi@unn.edu.ng (C.O.N.);

Received: 23.03.2023; Accepted: 28.05.2023; Published: 30.09.2025

Abstract: Apart from the pharmacodynamics of drugs and the resistance of the Plasmodium falciparum parasite to existing antimalarial drugs, pharmacokinetic-related properties of drugs also hamper their translation. The need to develop novel drugs with optimum solubility profiles necessitated training an efficient machine learning regression model to predict the solubility of a series of compounds. Four descriptors were used: octanol-water partition coefficient, molecular weight, number of rotatable bonds, and aromatic proportion from the simplified molecular-input line-entry system (SMILES) of 11,478 antiplasmodial molecules. This was trained using five regression models: multiple linear regression, k-nearest neighbors, LASSO regression, support vector regressor, and random forest regressor (RFR) to predict the solubility of molecules. The evaluation metrics (R²), mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) were used to assess the model performance. Of the performed algorithms, the RFR produced a robust model with model statistics of MSE 0.54, R² 0.85, MAE 0.41, and RMSE 0.73. The F-statistic for the model was 7214, showing a strong correlation between the descriptors and the solubility of molecules. This could efficiently predict the antimalarial activity for untested molecules to select promising ligands as leads for further optimization.

Keywords: antimalarial; machine learning; molecule descriptors; regression models; solubility

© 2025 by the authors. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The authors retain copyright of their work, and no permission is required from the authors or the publisher to reuse or distribute this article, as long as proper attribution is given to the original source.

1. Introduction

According to the WHO, Africa is home to 50% of all malaria deaths in children under five worldwide [1]. Despite the organization's ongoing struggles to make drugs more easily accessible, malaria continues to be one of the most severe and threatening diseases because of its negative and damaging impact on world populations [2]. The efficacy of the first-line artemisinin-based combination therapy (ACT), which is threatened constantly by the emergence and spread of drug resistance, is essential for the success of malaria control,

prevention, and treatment [3]. This is so notwithstanding the organization's ongoing elimination campaign, which involves considering all potential control measures [4]. All kinds of antimalarial drugs have developed resistance and so lost their therapeutic efficacy [5]. Resistance to these standard drugs poses a severe threat to the elimination of malaria, resulting in a sharp rise in the number of deaths each year, increased medical expenses, and lost productivity. Due to the challenges that arise during the process, drug development takes approximately 14 years, from necessary pre-clinical testing to regulatory approval [6]. Surprisingly, clinical trials are the most expensive factor during drug development. Given these difficulties, predictive modeling methods are anticipated to become more important in diagnosing hazards and creating tactics for developing new antimalarial drugs [7].

Different strategies have been put forth to discover or design new chemical entities with optimum pharmacokinetic and pharmacodynamic properties. The quantitative structure-activity relationship (QSAR) method uses computational modeling to unravel associations between the biological activities and physicochemical properties of chemical substances to create a robust statistical model to predict the biological activities of novel chemical entities [7]. The basic idea is that changes in structural properties lead to changes in biological activities [8]. Simplified molecular-input line-entry system (SMILES) notations are computer-readable and serve as an effective input for machine learning (ML) models. SMILES offer effective descriptors for QSAR methods, making it an established and effective parameter for QSAR of different chemical entities [9,10]. SMILES is a language with clear symbols (atom and bond symbols) that are compact and take up far less space than other representational structures. The drug discovery field is without a doubt one of the areas that will gain significantly and greatly from the development of ML, as they have become a feasible and valuable tool for data-driven predictions in pharmaceutical research, such as drug repurposing, drug-drug interactions, diagnosis, and pharmacogenomics [11, 12].

Presently, there is voluminous data on the antiplasmodial activity of diverse chemical entities, few of which have been harnessed to develop more effective chemotherapeutic agents. There is also a high rate of attrition of some of these chemical entities that managed to go far in the drug development cycle due to pharmacokinetic-related challenges, which could have been averted at the early stages of the cycle [13]. Apart from efficacy and toxicity-related problems, several drug discovery failures can be attributed to pharmacokinetic and biopharmaceutical profiles of chemical entities [7, 11]. To avert failures in the later stage of drug discovery, SMILES-based models for predicting aqueous solubility of diverse antiplasmodial chemical entities were constructed using machine learning algorithms. The performance of the models' Multiple Linear Regression (MLR), *k*-Nearest Neighbours (*k*-NN), least absolute shrinkage and selection operator (LASSO) regression, Support Vector Regressor (SVR), and Random Forest Regressor (RFR) on the training and test sets was also compared.

This study has shown that robust, reliable, and accurate models can be obtained even if only a small portion of the descriptors is employed. In the study, descriptor values for the compounds were extracted, evaluated to select the more significant descriptors for the solubility, and then trained the models using ML algorithms to identify four implemented models with comparable performance.

2. Materials and Methods

The process flow for training regression models to predict antimalarial activity is shown in Figure 1.

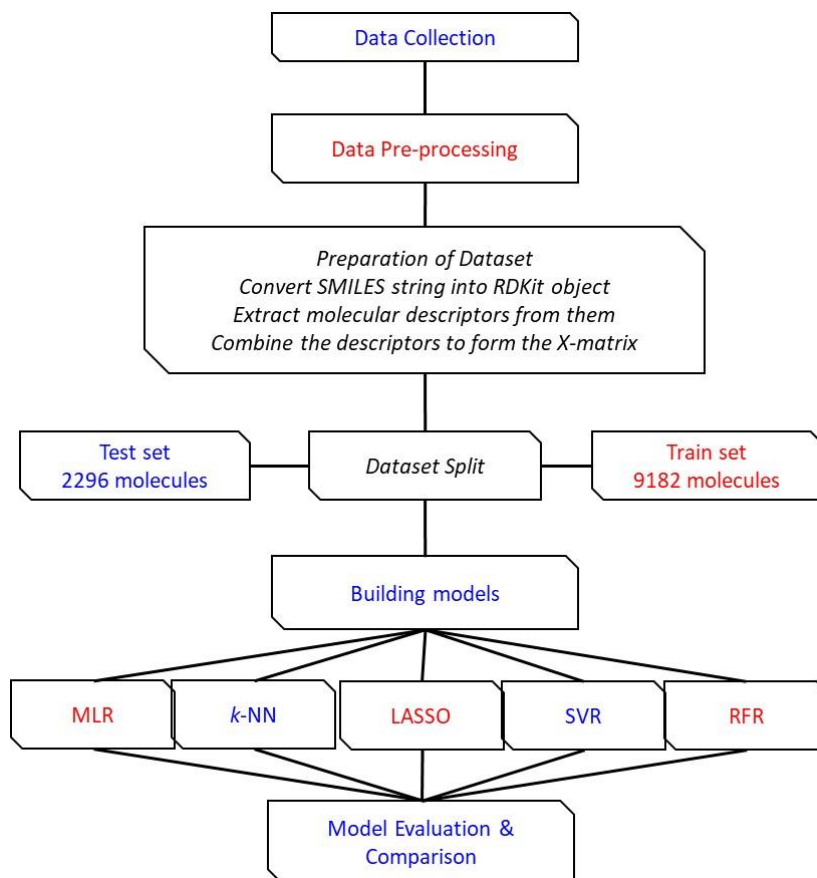


Figure 1. Flowchart for training regression models to predict antiplasmodial activity. RFR = random forest regressor; SVR = support vector regressor; *k*NN = *k*-Nearest neighbors; MLR = multiple linear regression; LASSO = least absolute shrinkage and selection operator.

2.1. Dataset.

The antimalarial molecule list was obtained from literature and the public database ChEMBL. They were converted into their respective and appropriate SMILES using the PubChem Identifier Exchange Service. It consists of 11481 molecules and three (3) features (Table 1). After filtering out missing and duplicate values, a total of 11478 antimalarial molecules were left. The details of the molecules used for the study can be found in the supplementary file, S1.

Table 1. Sample of unprocessed data.

ChEMBL ID	MW	LogS	SMILES
CHEMBL512770	268.75	-4.72644	<chem>C[n+]1cc2[nH]c3ccccc3c2c2ccccc21.[Cl-]</chem>
CHEMBL416376	396.45	-4.41485	<chem>COc1cc(COc2ccc(Cc3cnc(N)nc3N)cc2)cc(OC)c1OC</chem>
CHEMBL416376	396.45	-4.41485	<chem>COc1cc(COc2ccc(Cc3cnc(N)nc3N)cc2)cc(OC)c1OC</chem>
CHEMBL118602	364.45	-4.82925	<chem>CCc1nc(N)nc(N)c1Cc1ccc(OCc2ccccc2)c(OC)c1</chem>
CHEMBL117	254.24	-3.82459	<chem>O=c1cc(-c2ccccc2)oc2cc(O)cc(O)c12</chem>
CHEMBL151	286.24	-3.10069	<chem>O=c1cc(-c2ccc(O)c(O)c2)oc2cc(O)cc(O)c12</chem>
CHEMBL21395	350.42	-4.64130	<chem>CCOc1cc(Cc2cnc(N)nc2N)ccc1OCc1ccccc1</chem>
CHEMBL21395	350.42	-4.64130	<chem>CCOc1cc(Cc2cnc(N)nc2N)ccc1OCc1ccccc1</chem>
CHEMBL564804	500.64	-6.28600	<chem>COc1ccccc1C(=O)N[C@@H](CC(C)C)[C@@H](O)[C@H](O...</chem>
CHEMBL119224	292.34	-4.22814	<chem>Nc1ncc(Cc2ccccc2Oc3ccccc3)c2)c(N)n1</chem>
CHEMBL4228456	391.52	-7.68887	<chem>CCCCC1ccc(-c2ccc3c(=O)n(CC4CCCO4)c(N)nc3c2)cc1</chem>

ChEMBL ID	MW	LogS	SMILES
CHEMBL296588	685.90	-5.95511	<chem>CC(C)CC(=O)N[C@H](C(=O)N[C@H](C(=O)N[C@@H](CC(C)C)[C@@H](O)CC(=O)N[C@@H](C)C(=O)N[C@@H](CC(C)C)[C@@H](O)CC(=O)O)C(C)C)C(C)C</chem>
CHEMBL223243	596.90	-9.51201	<chem>CCCCC1ccc(C(=O)N(CCN(CCCC)CCCC)Cc2ccc(N3CCN(c4cccc4)CC3)cc2)cc1</chem>
CHEMBL219776	507.81	-8.29584	<chem>CCCCC1ccc(C(=O)N(CCN(CCCC)CCCC)Cc2ccc(N(CC)CC)cc2)cc1</chem>
CHEMBL273338	514.76	-7.59291	<chem>CCCCC1ccc(C(=O)N(CCN(CCCC)CCCC)Cc2ccc(-c3ccccc3)nc2)cc1</chem>
CHEMBL387316	586.82	-8.78750	<chem>CCCCC1ccc(C(=O)N(CCN(CCCC)CCCC)Cc2ccc(NC(=O)c3ccnc(OC)c3)cc2)cc1</chem>
CHEMBL385329	562.84	-7.99190	<chem>CCCCC1ccc(C(=O)N(CCN(CCCC)CCCC)Cc2ccc(NC(=O)C3CCNC3)cc2)cc1</chem>
CHEMBL227591	334.25	-5.67179	<chem>CCN1C(=O)/C(=C/c2cc(Cl)c(O)c(Cl)c2)SC1=S</chem>
CHEMBL227191	444.99	-8.67912	<chem>O=C(NC1CCCc2ccccc21)c1cc(Sc2cccc(Cl)c2)nc2ccccc12</chem>
CHEMBL227563	356.25	-7.95927	<chem>O=C1NC(=S)S/C1=C\c1ccc(-c2cccc(Cl)c2Cl)o1</chem>

LogS = solubility of molecules

2.2. Data pretreatment.

The SMILES string was converted to rdkit object using the RDKit library. RDKit reads SMILES strings of tested experimental antimalarial molecules that were transmuted from two databases (PubChem and the ChEMBL databases) and extracted descriptor values from them [14]. Thereafter, four 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. The cLogP, MW and nRB were computed using a python package RDKit, while AP was computed manually from the ratio of the number of aromatic atoms to the total number of heavy atoms (details of the descriptors are found in the supplementary file, S2). Machine learning uses molecular descriptors, which are numerical notations linked to chemical structure, to calculate correlations between a compound's physical attributes and biological activity [15]. The relationship between molecular descriptors of antimalarial drugs and their solubility was explained using regression models.

The computed descriptors were combined to form the X-matrix (Table 2), while the LogS column of the dataset forms the Y-matrix.

Table 2. Sample of the computed molecular descriptors (X-matrix).

Molecule Index	cLogP	MW	nRB	ArP
0	0.30280	268.747	0.0	0.894737
1	2.83660	396.447	8.0	0.620690
2	2.83660	396.447	8.0	0.620690
3	3.38180	364.449	7.0	0.666667
4	2.87120	254.241	1.0	0.842105
11473	2.05272	342.348	5.0	0.521739
11474	2.84120	346.402	8.0	0.480000
11475	1.90560	257.293	4.0	0.631579
11476	1.21570	240.270	2.0	0.777778
11477	1.15710	302.378	3.0	0.272727

2.3. Data split.

The X- and Y-matrix were split into train set (9182 molecules) and test set (2296 molecules), using a split ratio of 0.2, where 80 % is assigned to the train set, and 20 % is assigned to the test set [16, 17]. The size of the training dataset was denoted as X-train, Y-train, while the size of the test dataset was X-test, Y-test. The training set was used to train the model,

while 2296 molecules belonging to the test set were used to validate the model. The models were trained on the training set using the fit method. The hyperparameters of the models were adjusted on the test dataset to obtain the best hyperparameter configuration. This was done using a random search because their hyperparameters were continuous.

2.4. Building regression models.

Five (5) machine learning scikit-learn algorithms (Multiple Linear Regression (MLR), *k*-Nearest Neighbours (*k*NN), LASSO regression, Support Vector Regressor (SVR), and Random Forest Regressor (RFR)) were deployed to predict the solubility of molecules, to select further viable lead ligands for creative drug design [9, 16-18]. These learning algorithms combine ensemble learning techniques and conventional learning techniques with linear and nonlinear methods. The goal was to discover the best algorithm capable of predicting the antimalarial activity for untested compounds.

2.4.1. Multiple linear regression.

The dataset includes several features in predicting the aqueous solubility of antimalarial compounds. As a result, the Multiple Linear Regression technique was employed. Each feature was viewed as a dimension, and numerous *x* values were used to predict the result using the multiple linear regression equation [19].

$$Y = C + A1X1 + A2X2 + A3X3 + A4X4 \quad (1)$$

Where *Y* = predicted variable;

C = intercept;

*A*1, *A*2, *A*3, *A*4 = regression coefficients;

*X*1, *X*2, *X*3, *X*4 = molecular descriptors.

Plotting the MLR produces a slope of lines that provides the output variables (*A*1, *A*2, *A*3, *A*4) and *C* as their intercept. The MLR algorithm presumes that the correlation between the input and the output is linear.

2.4.2. *k*-Nearest neighbors.

One of the well-known non-parametric regressors in the field of machine learning is *k*-nearest neighbors [16]. It employs nonlinear regression in machine learning in a simple manner. It places an unclassified sample in the same class as the *k* samples in the training set that is closest to it and compares the new data point to the previous data points. In *k*NN models, the neighbors are assigned a certain weight that specifies that gives a clue of the average value they contribute [20]. It is usual to assign neighbors a weight of 1/*d*, where *d* is the distance between the neighbor and the object whose value will be predicted. The closest neighbors will contribute more than the distant neighbors when determining the value of a new data point using the *k*NN model.

The parameter setting for the model is as follows: `n_neighbours = 10`, `weights = 'distance'`, `metric = 'euclidean'`, `n_jobs = -1`

2.4.3. LASSO regression.

Lasso regression is a regularization technique that can be applied to feature selection, noise reduction, and model regularization [16, 18]. It is used when there is strong multicollinearity in a dataset. A small data change can result in a change in the regression coefficients as a result of the multicollinearity of the dataset. This describes the closeness of the independent variables. It entails adding more data to the unconstrained issue to enhance the quality of the result and is employed as a technique to broaden the generalizability of the trained model. It has the power to lower variability and raise the standard of linear regression techniques.

Lasso regression penalizes the sum of the coefficients to prevent prediction errors. With the use of the shrinkage technique in lasso regression, the coefficients are scaled back toward zero. To achieve a perfect fit with several datasets, it is used to reduce the regression coefficients [21]. The lasso hyperparameter has been set at 0.1 alpha value.

2.4.4. Support vector regressor.

The SVR falls under both linear and nonlinear categories of regression in machine learning. The applications of SVR include text categorization, pattern prediction, image processing and segmentation, and more. The SVM algorithm is used to locate the output in a multidimensional space [9]. The data points in a multidimensional space are not depicted as points in a 2D plot. A vector is used to represent the data points in a multidimensional space [22]. This model produces a max-margin hyperplane that divides the classes and gives each class a value. When the dataset has additional noise, the SVM model does not perform to its full potential. To create a linear decision surface, the input data was nonlinearly mapped to a space with higher dimensions [9]. The following hyperparameter values were explored: kernel = 'rbf'

2.4.5. Random forest regressor.

The random forest comprises building several decision trees while learning and generating a class that is a mode, showing the value with the greatest chance of occurrence, the value most commonly occurring in the dataset, or the expected mean of individual trees [17]. To lessen variance, several deep decision trees trained on many subsets of the same training set are averaged using random forests. An ensemble technique called random forest combines various decision trees (parallel trees). Using the classification or regression tree (CART) approach and the decreased Gini impurity (DGI) as the splitting criterion, each tree was created using a bootstrap sample randomly selected from the original dataset. Low bias, little connection between individual trees, and large variance are characteristics of RFR [16]. Using a grid search hyperparameter optimization method, Random Forest models were trained using the Scikit-learn library. The number of estimators was kept at 100, and the random state was at 0.

2.5. Model evaluation.

Different evaluation metrics (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 models [16, 19].

The most often used metric for assessing regression-based models is MSE, which estimates how closely the fitted line matches the actual data points. 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 [18]. 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 experimental values and the predicted values is equivalent to R^2 . The percentage of the outcome's variation that the predictor variables account for is known as R^2 [16]. 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 resemble the regression line. An elevated R^2 indicates a good model fit [23, 24].

2.6. Simplified Molecular-Input Line-Entry System.

The American Standard Code for Information Interchange (ASCII) strings are used to describe different molecules using the simplified molecular-input line-entry system (SMILES). When investigating chemical characteristics with computers and software models, SMILES are helpful. For the depiction of different molecular structure attributes, SMILES has established rules. Because ASCII strings can be processed so easily, SMILES is often used to provide optical descriptors for quantitative structure-activity relationships (QSAR) [10]. Finding ways to enhance the SMILES-based QSPR/QSAR concept is a continuous effort. SMILES is used to model QSAR/QSPR analysis instead of a graph. Additionally, fullerene-based HIV-1 PR inhibitors have been modeled using SMILES-based optical descriptors.

3. Results and Discussion

Machine-learning-based regression models were proposed to predict continuous solubility values of antimalarial molecules. These regression-based techniques give intuition beyond binary classification. The same features were used in the different models.

3.1. Machine Learning Models

A multiple regression model was built to predict continuous solubility values of antimalarial compounds. Four x values were used, and each feature was viewed as a dimension. Only four features were considered to predict the output using the multiple linear regression equation 2. This was derived from the coefficients of the features and the intercept.

$$\text{LogS} = 0.90 - 0.54\text{LogP} - 0.009\text{MW} + 0.038\text{RB} - 1.61\text{AP} \quad (2)$$

where LogS = Solubility of antimalarial compounds;

LogP = Octanol-water partition coefficient;

MW = Molecular weight;

nRB = Number of rotatable bonds;

ArP = Aromatic proportion.

The predicted solubilities for the test compounds are shown in Table 3. To prove further confidence in our predicted solubility values, the predicted solubility scores were plotted against the experimental solubility scores for both the train set and the test set, using different machine learning models (Figure 2). The closeness of the predicted solubility scores and the

experimental scores shows the robustness of our ML models. This shows that the predictive powers of the models are competent.

The correlations of the predicted and experimental solubility values are shown in Figure 2. The R^2 indicates how closely the data resemble the regression line and how well the data fit the regression line. For the MLR model, R^2 values for the train and test sets are 0.71 and 0.72, respectively. The R^2 values for the train and test sets when the RFR model was used were, respectively, 0.96 and 0.85. This is indicative of a good fit. Using the KNN, LASSO, and SVR models, R^2 values for the train and test sets are (0.98 train set, 0.81 test set), (0.70 train set, 0.72 test set), and (0.53 train set, 0.54 test set), respectively.

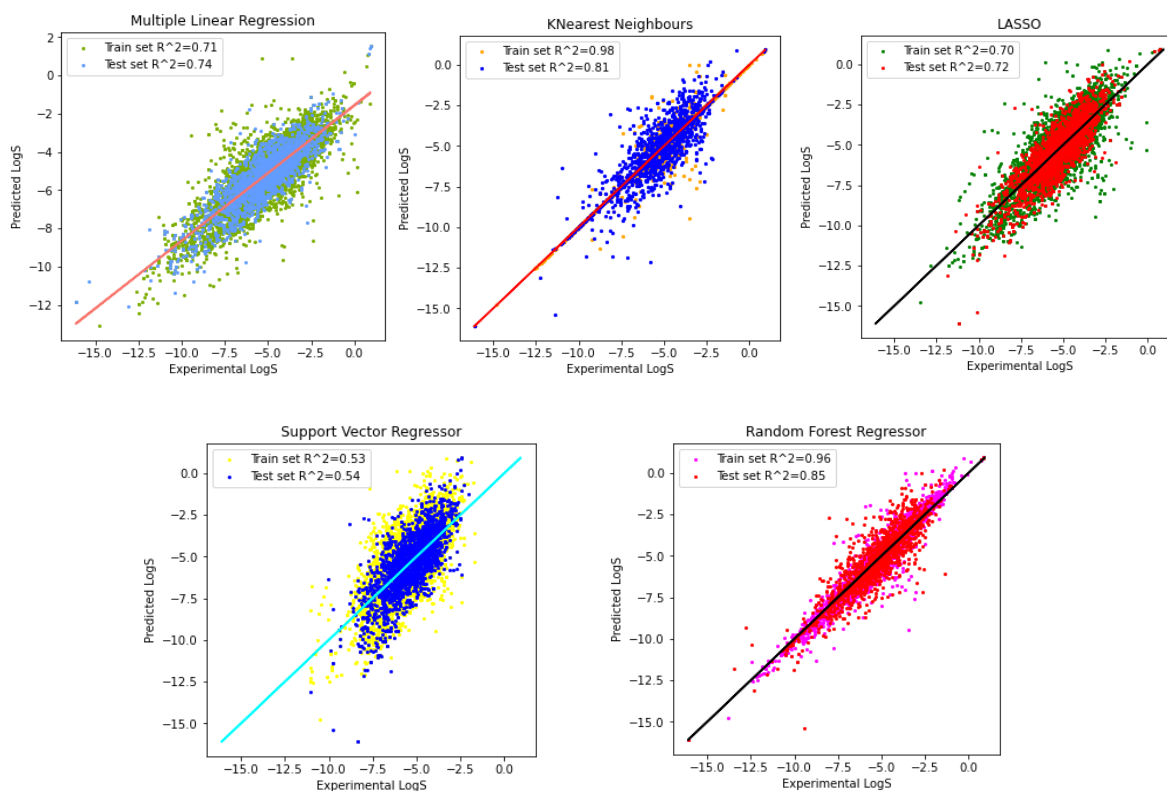


Figure 2. Correlation of the predicted and experimental solubility of antimalarial molecules in both the train and test data using different models.

Table 3. Predicted solubilities for the test molecules.

Molecule index	Experimental solubility	Predicted solubility
0	-5.73262	-4.921481
1	-10.25966	-9.714817
2	-6.83149	-7.638301
3	-4.70568	-4.407898
4	-7.53360	-7.136658
2291	-4.24022	-4.831124
2292	-5.24542	-5.067068
2293	-6.51061	-5.780181
2294	-6.36896	-5.433674
2295	-7.18611	-6.453384

3.2. Model evaluation and comparison.

The summary of the performance of the models is shown in Table 4. To find if the linear model is good, MSE and R^2 statistics are used. The lower the residue error, the better the model fits (the closer the data is to a linear relationship). R^2 measures the percentage of

solubility variation that descriptors can explain. Assuming a linear relationship, the proportion is high, and the R^2 value will be close to 1 if the descriptors X can predict the solubility.

Random forest regression outperformed all other models in all metrics (Table 4). The correlation between the predicted and experimental solubility for the RFR model is high (highest $R^2 = 0.85$; lowest RMSE of 0.73). k NN, MLR, and LASSO models also performed well. Decreased MSE scores show a good model fit (Figure 3). The MSE, MAE, and RMSE values for the test set were, respectively, 0.93, 0.77, and 0.96 when multiple linear regression was used. For the RFR model, MSE, MAE, and RMSE scores for the test set are respectively 0.54, 0.41, and 0.73. In the final five-fold cross-validation test, the highest accuracy of 0.8203 ± 0.0177 was achieved in the random forest regressor, while the lowest accuracy of 0.5143 ± 0.0114 was achieved in the support vector regressor (Table 4).

Table 4. Summary of the ML models' performances.

ML algorithms	MLR	KNN	LASSO	SVR	RFR
Train MSE	1.011191	0.071665	1.071719	1.655333	0.135456
Test MSE	0.931009	0.681645	0.997079	1.621369	0.537102
5-fold cross-validation	1.0192±0.0245	0.8112±0.0528	1.073±0.0301	1.6933±0.0531	0.6260±0.0566
Train R^2	0.710425	0.979477	0.693091	0.525961	0.961209
Test R^2	0.735241	0.806155	0.716453	0.538918	0.847260
5-fold cross-validation	0.7094±0.0085	0.7673±0.0154	0.6922±0.0067	0.5143±0.0114	0.8203±0.0177
Train MAE	0.76617 0	0.586540	0.792787	0.977027	0.187781
Test MAE	0.750951	0.415008	0.776314	0.965481	0.411517
5-fold cross-validation	0.7666±0.0112	0.4669±0.0145	0.7932±0.0114	0.9895±0.0195	0.4578±0.0158
Train RMSE	1.00558	0.267704	1.035239	1.286598	0.368043
Test RMSE	0.964888	0.825618	0.998539	1.273330	0.732872
5-fold cross-validation	1.0063±0.0121	0.9002±0.0296	1.0359±0.0145	1.3011±0.0204	0.7904±0.0357

RFR = random forest regressor; SVR = support vector regressor; KNN = K Nearest neighbors; MLR = multiple linear regression; MSE = mean square error; R^2 = coefficient of determination; MAE = mean absolute error; RMSE = root mean square error. Five-fold cross-validation values were expressed as mean±SD

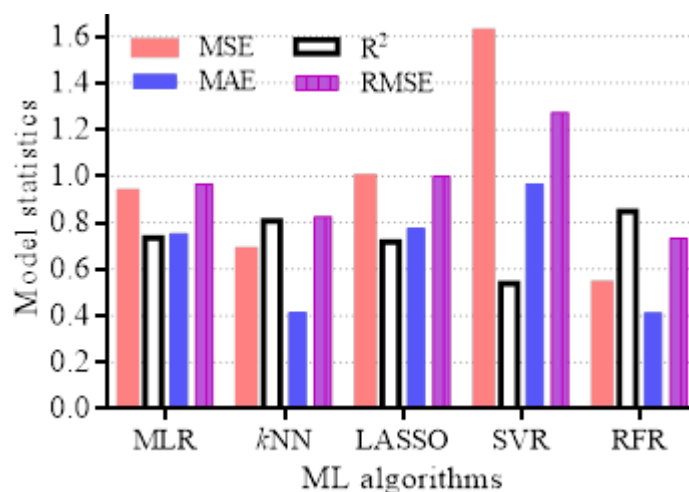


Figure 3. Grouped plots showing the models' statistics values of the test set. RFR = random forest regressor; SVR = support vector regressor; k -NN = K Nearest Neighbours; MLR = multiple linear regression.

The size of train data is correlated to how well the learning algorithm functions. Since the random forest regressor algorithm performed the best in the final test, its learning curves were examined (Figure 4). The line graphs illustrated the learning curves for RFR models trained using training sizes of 1000, 2000, 3000, 4000, 5000, and 9182 on solubility datasets. The method of five-fold cross-validation was used. In the dataset, 80% of the data was used as training data, and the remaining 20% as test data.

Figure 4 displays the model performance and accuracy variations between the training and test datasets as the data volume varied. The learning curves' results revealed a consistent pattern in the four metrics. As the data volume increased, the model's performance on the test dataset gradually improved.

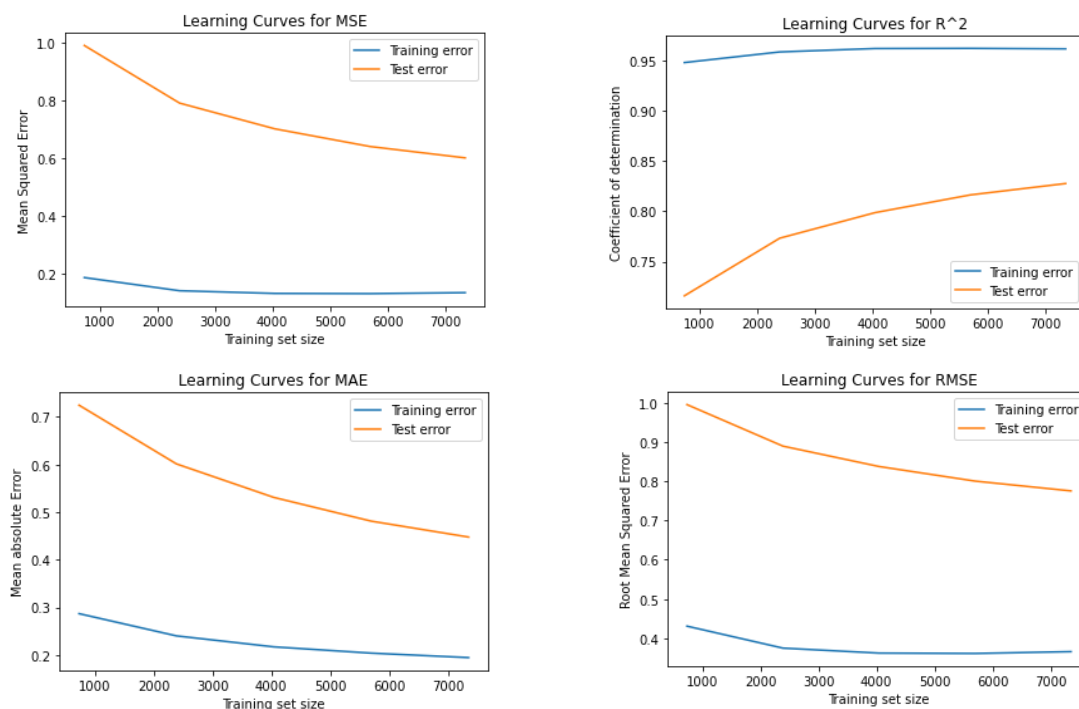


Figure 4. Learning curves of RFR models trained on the solubility dataset using the training sizes of 1000, 2000, 3000, 4000, 5000, and 9182.

3.3. Discussion.

The solubility of chemical entities is an important parameter to consider at all drug discovery and development stages. Machine learning is a capability that is used in finding useful information, such as solubility, from data; it is widely used in several fields in the sciences [25]. The goal of this study's machine learning model development is to build a robust predictive model for predicting the continuous solubility values of antimalarial compounds. Regression-based models were previously used to predict anti-inflammatory binding activity in a set of compounds [26]. Its use in predicting tuberculosis prevalence rates achieved a goodness-of-fit with an adjusted R^2 of 0.96 [27]. Support vector regression was employed to predict skin aging [27]. The authors evaluated the algorithm on test compounds to verify the accuracy of the generated predictive model. The shallow neural network was developed using computed molecular descriptors in DMSO to classify molecules into soluble and not soluble [28]. The model achieved 93% accuracy. In a similar work, the shallow neural network was compared with SVM using 65 structural group descriptors to develop models for the prediction of solubility in both organic solvents and water [20]. In another work, seven different algorithms were built to predict solubility in water, ethanol, acetone, and benzene. In this work, 14 calculated descriptors and the melting point were chosen by human experts [22]. Though classic machine learning algorithms heavily rely on feature engineering, high-dimensional and sparse data are frequently used in machine learning applications.

Machine learning algorithms are proving to be useful in areas of medicine where more precise prognostic models are needed, such as the prediction of patients' outcomes and the diagnosis of diseases. Here, several models to predict continuous solubility values were

developed, and the best models were selected from the several models generated. As shown in Table 4, the best models achieved in each evaluation metric in the Random Forest Regressor were MSE 0.54, R^2 0.85, MAE 0.41, and RMSE 0.73. An R^2 of 0.85 means about 85% of the solubility is explained by the descriptors. The RFR is always known for its high performance and ease of use [17]. Additionally, it is robust to adjustable parameters. RFR is a nonlinear regressor, and the association between molecular structure and the solubility of molecules is nonlinear [20]. However, a slight decrease in prediction performance was achieved in the k Nearest Neighbor model (MSE 0.68, R^2 0.81, MAE 0.42, and RMSE 0.83). RFR proved its robustness by producing the least error, that is, reducing the MSE to 0.54, the most important metric in assessing predictive models. The worst prediction performance was observed in the Support Vector Regressor model (MSE 1.62, R^2 0.54, MAE 0.97, RMSE 1.27). Meanwhile, both the Multiple Linear Regression and LASSO models had good prediction performance, as their R^2 scores were greater than 0.70. This shows that more than 70% of their solubility is described by the descriptors.

Prediction of solubility values for antimalarial compounds is an important task in drug discovery research. Solubility is a key property of drug molecules, as it can affect their absorption, distribution, metabolism, and excretion in the body [28, 29]. In the case of antimalarial compounds, solubility can also affect their effectiveness in treating malaria. To prove further confidence in our predicted solubility values, the predicted solubility scores were plotted against the experimental solubility scores for both the train set and the test set, using different machine learning models (Figure 2). The closeness of the predicted solubility scores and the experimental scores shows the robustness of our models. This shows that the predictive powers of the models are competent.

In addition to how well machine learning models performed when applied to the total amount of solubility data, the accuracy of the models under various conditions was also thoroughly explored, demonstrating how the learning tasks affected the performance of the models. To describe those coefficients that are relevant to predict solubility, p -values were measured. When calculating statistical significance, the p -value is used to indicate whether or not the null hypothesis should be rejected [30]. If a variable is statistically significant for predicting the target, the p -value for each coefficient will indicate this. The p -values for the descriptors are MolLogP 0.00, MolWt 0.00, NumRotatableBonds 2.19×10^{-16} , and AromaticProportion 5.79×10^{-106} . Since the p -values are less than 0.05, it shows there is a strong correlation/relationship between the descriptors and the solubility.

For multiple linear regression, the p -value is unique to each predictor, whereas the F -statistic is determined for the entire model [19]. The F -statistic for the model is 7214. Since the F -statistic is much larger than 1, it suggests that there is a strong relationship between the descriptors and solubility. To indicate a strong relationship for a small dataset, the F value must be significantly more than 1. The model with the lowest Akaike Information Criterion (AIC) offers the best fit.

Solubility is one of the essential features of a drug candidate in medicinal chemistry and pharmacology [6]. Regression-based predictions of the solubility of drug candidates can be compared to their human predictions. Several researchers have previously developed models for the prediction of solubility based on either property or activity [7]. COSMO-RSol considered the energy of fusion in the prediction of equilibrium constants for liquid-liquid, liquid-vapor, and solid-solid states. Another work predicted solubility with temperature. It is more informative to predict the solubility values for antimalarial compounds rather than just

indicating the binary classification of either being soluble or not soluble. It is imperative to point out here that further studies will explore a variety of embedding strategies (sequence embeddings, knowledge graph embeddings, and graph embeddings) to enhance the prediction performance of our models. Our goal is to further apply this study to a real-life drug design and repurposing and then test our model's performance to experimentally evaluate our findings' clinical applicability.

4. Conclusions

The study demonstrated that RFR is a powerful predictive supervised learning model with reproducible outcomes and the lowest model error when compared to other ML techniques. This model allows for the prediction of drug solubility, which can be utilized in the design of new bioactive chemical entities using artificial intelligence qualities. The authors concluded that the increasingly accurate methods of predicting solubility might eliminate the need for animal testing.

Author Contributions

Conceptualization, T.O.A. and C.O.N.; methodology, T.O.A.; software, T.O.A.; validation, T.O.A. and C.O.N.; formal analysis, T.O.A.; resources, T.O.A. and C.O.N.; data curation, C.O.N.; writing – original draft preparation, T.O.A.; writing – review and editing, T.O.A. and C.O.N.; supervision, C.O.N. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Data supporting the findings of this study are available upon reasonable request from the corresponding author.

Funding

This research received no external funding.

Acknowledgments

None.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Chen, S.; Hong, J.; Wang, G.; Liu, Y. Move More, Sit Less and Sleep Well: An analysis of WHO movement guidelines for children under 5 years of age. *Sports Med Health Sci* **2021**, *3*, 54–57. <https://doi.org/10.1016/j.smhs.2021.02.003>.
2. Rathmes, G.; Rumisha, S.F.; Lucas, T.C.D.; Twohig, K.A.; Python, A.; Nguyen, M.; Nandi, A.K.; Keddie, S.H.; Collins, E.L.; Rozier, J.A.; Gibson, H.S.; Chestnutt, E.G.; Battle, K.E.; Humphreys, G.S.; Amratia, P.; Arambepola, R.; Villa, A.B.; Hancock, P.; Millar, J.J., Weiss, D.J. Global estimation of antimalarial drug effectiveness for the treatment of uncomplicated *Plasmodium falciparum* malaria 1991 – 2019. *Malaria J* **2020**, *19*, 1–15. <https://doi.org/10.1186/s12936-020-03446-8>.
3. Ikwuka, C.E.; Asogwa, C.M.; Ikwuka, O.J.; Ogonna, 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>.
4. 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>.
5. Manirakiza, G.; Kassaza, K.; Taremwa, I.M.; Bazira, J.; Byarugaba, F. Molecular identification and anti - malarial 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>.
6. Charman, S.A.; Andreu, A.; Barker, H.; Blundell, S.; Campbell, A.; Campbell, M.; Chen, G.; Chiu, F.C.K.; Crighton, E.; Katneni, K.; Morizzi, J.; Patil, R.; Pham, T.; Ryan, E.; Saunders, J.; Shackleford, D.M.; White, K. L.; Almond, L.; Dickins, M.; ... Abla, N. An in vitro toolbox to accelerate anti - malarial drug discovery and development. *Malaria J* **2020**, 1–27, <https://malariajournal.biomedcentral.com/articles/10.1186/s12936-019-3075-5>.
7. Arabi, A.A. Artificial intelligence in drug design: algorithms, applications, challenges and ethics. *Future Drug Discov* **2021**, *3*, 2631–3316. <https://doi.org/10.4155/fdd-2020-0028>.
8. 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>.
9. 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>.
10. Costa, A.S.; Martins, J.P.A.; de Melo, E.B. SMILES-based 2D-QSAR and similarity search for identification of potential new scaffolds for development of SARS-CoV-2 MPRO inhibitors. *Struct Chem* **2022**, *33*, 1691–1706. <https://doi.org/10.1007/s11224-022-02008-9>.
11. 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*. *Biomolecules* **2021**, *11*, 1750–1770. <https://doi.org/10.3390/biom11121750>.
12. 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>.
13. Nnadi, C.O.; Ozioko, L.U.; Eneje, G.C.; Onah, C.M.; Obonga, W.O. In-vivo antitrypanosomal effect and *in-silico* prediction of chronic toxicity of N-methylholaphyllamine in rats. *Trop J Pharm Res* **2020**, *19*, 2369–2375, <https://www.ajol.info/index.php/tjpr/article/view/216415>.
14. Comensana, A.E.; Huntington, T.T.; Scown, C.D.; Niemeyer, K.D.; Rapp, V.H. A systematic method for selecting molecular descriptors as features when training models for predicting physicochemical properties. *Fuel* **2022**, *321*, 123836. <https://doi.org/10.1016/j.fuel.2022.123836>.
15. 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>.
16. Du, Z.; Wang, D.; Li, Y. Comprehensive evaluation and comparison of machine learning methods in QSAR modeling of antioxidant tripeptides. *ACS Omega* **2022**, *7*, 25760–25771 <https://doi.org/10.1021/acsomega.2c03062>.

17. Urista, D.V.; Carru, D.B.; Otero, I.; Arrasate, S.; Quevedo-tumaili, V.F.; Gestal, M.; Gonz, H.; Munteanu, C.R. Prediction of Antimalarial Drug-Decorated Nanoparticle Delivery Systems with Random Forest Models. *Biol* **2020**, *9*, 198–223. <https://doi.org/10.3390/biology9080198>.
18. 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>.
19. 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>.
20. Ye, Z.; Ouyang, D. Prediction of small-molecule compound solubility in organic solvents by machine learning algorithms. *J Cheminform* **2021**, *13*, 1–13. <https://doi.org/10.1186/s13321-021-00575-3>.
21. Lee, S.; Lee, M.; Gyak, K.W.; Kim, S.D.; Kim, M.J.; Min, K. Novel Solubility Prediction Models: Molecular Fingerprints and Physicochemical Features vs Graph Convolutional Neural Networks. *ACS Omega* **2022**, *7*, 12268–12277. <https://doi.org/10.1021/acsomega.2c00697>.
22. Alvarsson, J.; Lampa, S.; Schaal, W.; Andersson, C.; Wikberg, J.E.S.; Spjuth, O. Large-scale ligand-based predictive modelling using support vector machines. *J Cheminform* **2016**, *8*, 1–10. <https://doi.org/10.1186/s13321-016-0151-5>.
23. Gao, Y.; Xiong, Z.; Wang, X.; Ren, H.; Liu, R.; Bai, B.; Zhang, L.; Li, D. Abnormal degree centrality as a potential imaging biomarker for right temporal lobe epilepsy: a resting-state functional magnetic resonance imaging study and support vector machine analysis. *Neuroscience* **2022**, *487*, 198–206. <https://doi.org/10.1016/j.neuroscience.2022.02.004>.
24. Messenlehner, J.; Hetman, M.; Tripp, A.; Wallner, S.; Macheroux, P.; Gruber, K.; Daniel, B. The catalytic machinery of the FAD-dependent AtBBE-like protein 15 for alcohol oxidation: Y193 and Y479 form a catalytic base, Q438 and R292 an alkoxide binding site. *Arch Biochem Biophys* **2021**, *700*, 1–10. <https://doi.org/10.1016/j.abb.2021.108766>.
25. Galetsi, P.; Katsaliaki, K.; Kumar, S. Big data analytics in health sector: Theoretical framework, techniques and prospects. *Int J Info Manag* **2020**, *50*, 206–216. <https://doi.org/10.1016/j.ijinfomgt.2019.05.003>.
26. Staszak, M.; Staszak, K.; Wieszczycka, K.; Bajek, A.; Roszkowski, K.; Tylkowski, B. Machine learning in drug design: Use of artificial intelligence to explore the chemical structure-biological activity relationship. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2022**, *12*, e1568. <https://doi.org/10.1002/wcms.1568>.
27. Lytras, M.D.; Chui, K.T.; Visvizi, A. Data Analytics in Smart Healthcare : The Recent Developments and Beyond. *Appl Sci* **2019**, *9*, 2812. <https://doi.org/10.3390/app9142812>.
28. Fleuren, L.M.; Klausch, T.L.; Zwager, C.L.; Schoonmade, L.J.; Guo, T.; Roggeveen, L.F.; Swart, E.L.; Girbes, A.R.; Thorat, P.; Ercole, A.; Hoogendoorn, M. Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med* **2020**, *46*, 383–400. <https://doi.org/10.1007/s00134-019-05872-y>.
29. Nnadi, C.O.; Ayoka, T.O.; Okorie, N.H. 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>.
30. 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>.

Publisher's Note & Disclaimer

The statements, opinions, and data presented in this publication are solely those of the individual author(s) and contributor(s) and do not necessarily reflect the views of the publisher and/or the editor(s). The publisher and/or the editor(s) disclaim any responsibility for the accuracy, completeness, or reliability of the content. Neither the publisher nor the editor(s) assume any legal liability for any errors, omissions, or consequences arising from the use of the information presented in this publication. Furthermore, the publisher and/or the editor(s) disclaim any liability for any injury, damage, or loss to persons or property that may result from the use of any ideas, methods, instructions, or products mentioned in the content. Readers are encouraged to independently verify any information before relying on it, and the publisher assumes no responsibility for any consequences arising from the use of materials contained in this publication.