PLOS ONE

RESEARCH ARTICLE

Predictive classifier models built from natural products with antimalarial bioactivity using machine learning approach

Samuel Egieyeh[1,2]*, James Syce[2☯], Sarel F. Malan[2☯], Alan Christoffels[1]

1 South African Medical Research Council Bioinformatics Unit, South African National Bioinformatics Institute, University of the Western Cape, Cape Town, South Africa, **2** School of Pharmacy, University of the Western Cape, Cape Town, South Africa

☯ These authors contributed equally to this work.
* segieyeh@uwc.ac.za

Citation: Egieyeh S, Syce J, Malan SF, Christoffels A (2018) Predictive classifier models built from natural products with antimalarial bioactivity using machine learning approach. PLoS ONE 13(9): e0204644. https://doi.org/10.1371/journal.pone.0204644

Editor: Dinesh Gupta, International Centre for Genetic Engineering and Biotechnology, INDIA

Received: December 18, 2017

Accepted: September 12, 2018

Published: September 28, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by South African Research Chairs Initiative of the Department of Science and Technology (DST) and the National Research Foundation (NRF) of South Africa http://www.nrf.ac.za. Grant number: UID-64751. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

In view of the vast number of natural products with potential antiplasmodial bioactivity and cost of conducting antiplasmodial bioactivity assays, it may be judicious to learn from previous antiplasmodial bioassays and predict bioactivity of these natural products before experimental bioassays. This study set out to harness antimalarial bioactivity data of natural products to build accurate predictive models, utilizing classical machine learning approaches, which can find potential antimalarial hits from new sets of natural products. Classical machine learning approaches were used to build four classifier models (Naïve Bayesian, Voted Perceptron, Random Forest and Sequence Minimization Optimization of Support Vector Machines) from bioactivity data of natural products with *in-vitro* antiplasmodial activity (NAA) using a combination of the molecular descriptors and two-dimensional molecular fingerprints of the compounds. Models were evaluated with an independent test dataset. Possible chemical features associated with reported antimalarial activities of the compounds were also extracted. From the results, Random Forest (accuracy 82.81%, Kappa statistics 0.65 and Area under Receiver Operating Characteristics curve 0.91) and Sequential Minimization Optimization (accuracy 85.93%, Kappa statistics 0.72 and Area under Receiver Operating Characteristics curve 0.86) showed good predictive performance for the NAA dataset. The amine chemical group (specifically alkyl amines and basic nitrogen) was confirmed to be essential for antimalarial activity in active NAA dataset. This study built and evaluated classifier models that were used to predict the antiplasmodial bioactivity class (active or inactive) of a set of natural products from interBioScreen chemical library.

Introduction

The devastating effect of malaria is evidenced by 584,000 deaths of which 78 percent were children under five years of age in 2013 [1] and thousands of person-hours lost to morbidity [2,3]. Majority of deaths due to malaria are caused by *Plasmodium falciparum*, the most virulent amongst the species that cause the disease [4–6]. The growing resistance and failure of existing

Competing interests: The authors have declared that no competing interests exist.

first-line antimalarial drugs have exacerbated the situation leading to an exigent need to develop novel antimalarial drug candidates [7–9]. Judging by the immense contribution of nature to existing antimalarial drugs [10–13] and the likelihood to encounter novel chemotypes in natural products, *in-vitro* malarial screen data of natural products may be the appropriate starting point for the discovery of new antimalarial drugs.

Recently a number of publications have reported the *in-vitro* antiplasmodial activities of natural products from plants [10–13] and marine life forms [14,15]. In addition, datasets of *in-vitro* antiplasmodial bioassays of natural products and synthetic compounds have been made available in public domain [16–18]. The availability of such data for malaria drug discovery has motivated us to create predictive models based on molecular properties using machine-learning approaches.

Machine Learning, an aspect of artificial intelligence, is the practice of using algorithms to analyze input data (training data), learn from it, and then make a prediction on another set of related or unrelated data. Machine learning approaches may be supervised or unsupervised if the algorithms learned from labelled or unlabeled data [19]. Unsupervised statistical learning allows learning of relationships and structure of input data. Supervised machine learning involves building a model for predicting an output based on one or more sets of input data.

It has been shown that machine learning approaches could accurately predict the activities in assorted sets of compounds with activities as diverse as anti-tubercular [20], antimalarial [21] and RNA-binders [22]. To our knowledge, there has not been any bioactivity predictive model specifically for natural products with antiplasmodial or antimalarial activities. Increasing number of natural products, mostly from ethnomedicine in malaria-endemic regions, show good *in-vitro* and/or *in-vivo* antiplasmodial activities [23–26]. The antiplasmodial bioactivity data for these natural products present a dais to build models that may be used to screen other natural products and predict their potential antiplasmodial activities.

This present study focused on the development of machine learning classification models for natural products with varying *in-vitro* antiplasmodial activities (NAA). Four classification models were built from the bioactivity class (Active or Inactive) and a combination of molecular descriptors (MD) and molecular fingerprints (MF) of the NAA dataset. The performances of the classification models were assessed with standard model evaluation parameters (including accuracy and area under the Receiver Operating Characteristic (ROC) curve). We also analyzed the chemical structures of the datasets to find molecular fragments or chemical features enriched within the active and inactive compounds. Finally, we showed that the machine learning models built in this study might be used to screen large natural compound libraries *in-silico* and identify potential antiplasmodial compounds. This may limit the need for *in-vitro* screening and drastically reduce the expense of finding hits from natural products for antimalarial drug discovery.

Materials and methods

An original Konstanz Information Miner (KNIME) workflow [27,28] was set up (Fig 1) and used for the machine learning from our set of natural products with *in-vitro* antiplasmodial activities (NAA) in order to predict the activity class (active or inactive) of NAA.

Data

The dataset used in this study consist of natural products that have been tested for *in-vitro* antiplasmodial activities (NAA) compiled in-house from literature, PhD and Masters Theses and public chemical databases. The chemical structures of the compounds in NAA were either downloaded in the SMILES format from public chemical databases (ChEMBL or PubChem) or drawn using Chemtool version 1.6.13 (http://ruby.chemie.uni-freiburg.de/~martin/chemtool) running on a Linux platform. The dataset (NAA) was subdivided into two groups

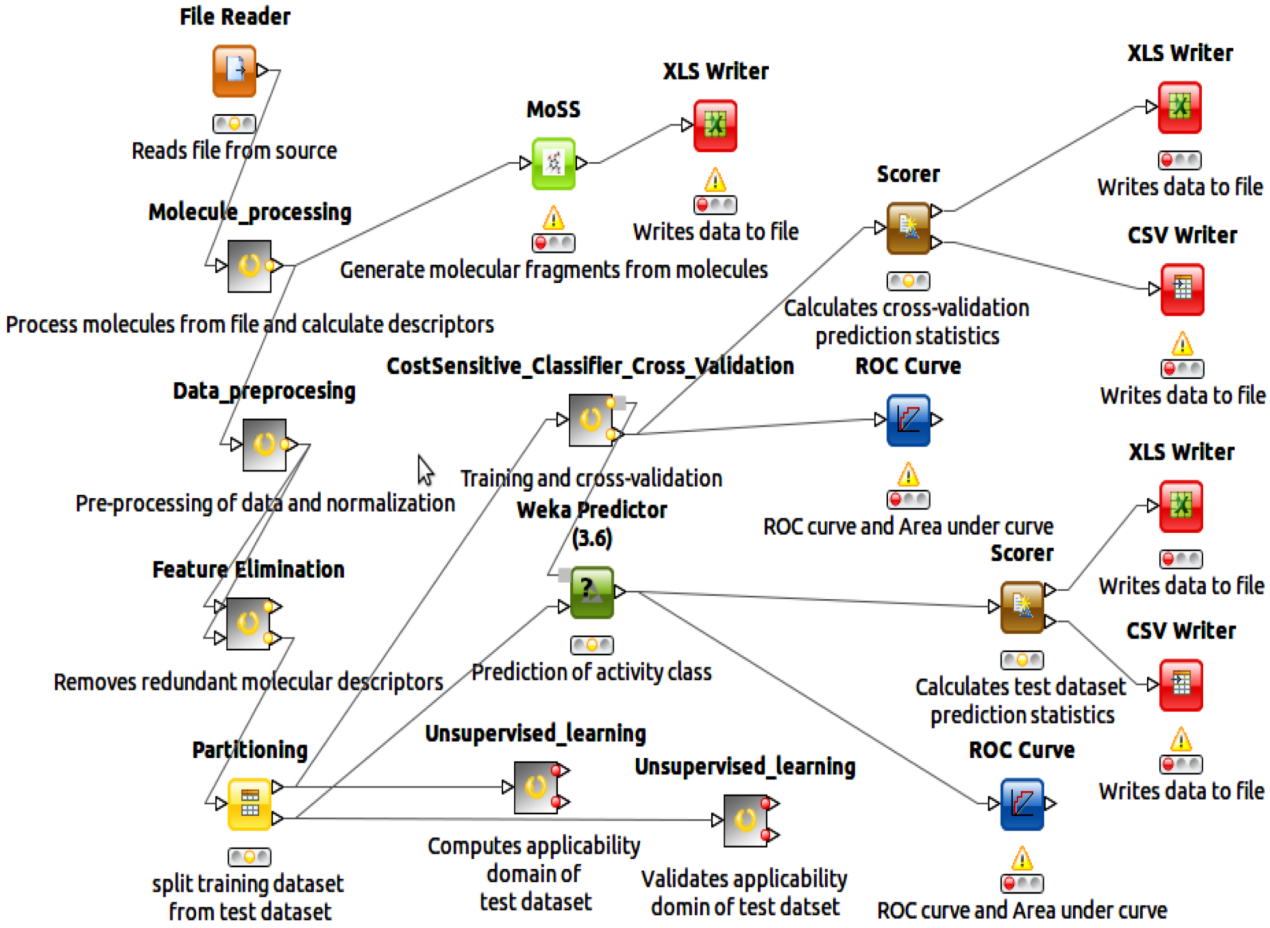

Fig 1. KNIME workflow. Screen-shot of the KNIME workflow used to build the classifier machine-learning models.

https://doi.org/10.1371/journal.pone.0204644.g001

based on their *in-vitro* antiplasmodial activities (IC_{50}): Active (A) (IC_{50}: < 10 μM) and Inactive (A) (IC_{50}: ≥10 μM). A total of 1155 NAA compounds were used in this study, with 70% classified as active and 30% as inactive (S1 Table).

Machine learning algorithms

Four classifier algorithms were used to learn from the dataset: Naïve Bayesian classifier [29,30], Sequential Minimization Optimization (SMO) classifier, a strategy for solving the quadratic problems during training with Support Vector Machine (SVM) [31,32], Random Forest (RF) classifier [33,34] and Voted perceptron (VP) classifier [35,36]. The specific classifiers were chosen in an attempt to represent four major types of classifiers models: Naïve Bayes represents the Bayes classifiers; Random Forest represents the tree-based classifiers; SMO represents the function-based classifier; and the Voted Perceptron represents the neural network classifiers. The classifier algorithms were executed with Waikato Environment for Knowledge Analysis (Weka 3.6) nodes [37] in Konstanz Information Miner (KNIME) [38].

Dataset pre-processing and calculation of molecular descriptors and molecular fingerprints

The "RDKit Descriptor Calculation" and "RDKit Fingerprint" nodes [39] were used to calculate the molecular descriptors and molecular fingerprint. The "Data_preprocessing" node was

then used to normalize the molecular descriptors using a minimum-maximum normalization model. The bit vector representing the molecular fingerprint was expanded into individual columns for each compound.

Selection of descriptors or features

The objective of features selection is three-fold: improving the prediction performance of the predictive model, providing faster and more economical predictive models, and providing a better understanding of the underlying process that generated the data [40]. The "Feature Elimination" (FE) meta-node in KNIME was used to select descriptors that are beneficial to build efficient classifier models.

Training of classifier models

The purpose of the classification algorithm was to build a classifier model that assigns a class (e.g. active/inactive) to molecules defined by a set of attributes (e.g. molecular descriptors). A metanode in the KNIME workflow (Fig 1) was designed to build the various classifier models that were earlier mentioned (i.e. Naïve Bayesian classifier, Sequential Minimization Optimization (SMO) classifier, Random Forest (RF) classifier and Voted perceptron (VP) classifier). The Partitioning node was used to split the data coming from the Feature Elimination meta-node into 80% training cum validation set and 20% independent test set by stratified sampling. The former was then piped into the "CostSensitive_Classifier_Cross_Validation" meta-node while the later was passed to the Weka Predictor (3.6) node. The Weka "Cost Sensitive Classifier" was used to build the classifier models and the Weka Predictor generated predictions from the test data. Regarding the features used to build the model, the molecular descriptors and molecular fingerprints were initially used separately to train the models. In an attempt to improve the accuracy of the model, we combined the molecular descriptors and the molecular fingerprints for each compound and used that combined feature to train the models.

Class Imbalance and cost-sensitive classification

The imbalance bioactivity class (70% active and 30% inactive) was recognized as a major limitation to building a reliable model. Most bioassay datasets are imbalanced where one class is overly represented as observed in our datasets (approximately 70% active class (A) and 30% inactive class (N)). Prior to building the classifier model, the "SMOTE (Synthetic Minority Over-sampling Technique)" node within the KNIME was used to balance the bioactivity classes [38]. This node oversamples the input NAA dataset to enrich the inactive instances in the training dataset.

In addition, cost-sensitivity, which does not assume equality of the *costs* caused by different kinds of errors, was applied to the classifier algorithms used in this study. The Weka "meta-CostSensitiveClassifier" node in KNIME [38] was used to build the classifier models from NAA dataset. The Weka "meta-CostSensitiveClassifier" makes its base classifier cost sensitive and provide it with the capability to predict a class that leads to the lowest expected cost [37,41]. For our datasets that have two class representations (i.e. active (A)/inactive (N)), cost sensitivity was introduced by using a '2 ×2' dimension cost matrix (Table 1).

The four sections of a cost matrix can be read as True Positives (TP)—actives classified as actives; False Positives (FP)–inactives classified as actives; True Negatives (TN)—inactives classified as inactives; False Negatives (FN)—actives classified as inactive. The Weka "meta-CostSensitiveClassifier" enforces a penalty or weight on the base classifier for generating false positives (FP) or false negatives (FN) during learning. By default, the weight on the cost matrix is set to one for FP and FN. However, it has been reported that during the development of the

Table 1. Cost matrix used by weka "meta-cost sensitive classifier".

TP (0.0)	FN (2.0)
FP (1.0)	TN (0.0)

The cost values for each possible classification are in brackets. *True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN)*

https://doi.org/10.1371/journal.pone.0204644.t001

classifier models the cost of misclassification may not always be the same [42]. In bioactivity prediction, the cost of FN (misclassification of an active compound as inactive) may be greater than the cost of FP (misclassification of an inactive compound as active) [20]. That is, the cost of missing a potential active compound is greater than the cost of predicting an inactive compound as active [20]. Therefore, the weight or penalty for FN was set to two (Table 1) to minimize the chance of FN misclassification. This cost matrix was used to build the NB, VP, SMO and RF classifier models.

Classifier model performance evaluators

Accuracy statistics and receiver operating characteristic. The performances of the classifier models were assessed by accuracy statistics and Receiver Operating Characteristic graph after a 10-fold cross validation of a training set and prediction of the bioactivity class of an independent test set. In the KNIME workflow (Fig 1), the Scorer node and the ROC node were attached to the output from the Weka predictor nodes (from the cross-validation and the independent test data prediction). The outputs from the Scorer node include a confusion matrix and evaluation statistics (including accuracy of the prediction, Kappa statistic and mean absolute error). Accuracy indicates the proximity of measurement of results to the true value. This can be mathematically expressed as:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} * 100 \quad (1)$$

Where TP is True Positives; FP is False Positives; TN is True Negatives and FN is False Negatives.

The outputs from the ROC node encompass the Receiver Operating Characteristic (ROC) curve, which is a graphical plot of True Positive Rates (TPR) vs. False Positive Rates (FPR) for a binary classification system. The Area under Curve (AUC) value was also computed from the ROC curve and in our case, it denotes the probability that a classifier will rank a randomly chosen active compound higher than a randomly chosen inactive compound.

Applicability domain (AD)

Generally, machine learning models methods are more likely to show good predictive performance for compounds that share similar properties to compounds in the training set. Thus, it is necessary to define the "applicability domain" (i.e. the boundary defined by the chemical space in the training set) of the models and to check if new test compounds fall within such domain [43]. One of the simplest and commonly applied methods used to define AD is based on range-based definition with a preliminary Principal Components (PC) rotation [44]. In the present study, we defined the AD of the models using the training data and evaluated the extent to which the independent test data fit into the AD. This will be helpful to explain the accuracy of prediction from models and assess whether a new compound is inside or outside the AD of the models.

Principal Component Analysis (PCA) of the molecular fingerprints of the compounds in the training/validation data was done with Unsupervised learning metanode in the machine learning KNIME workflow (Fig 1). The PCA was carried out for the training data from the Partition node in the KNIME workflow before cross- validation (Fig 1). PCA of the independent test dataset was also performed in order to validate if the compounds within the test dataset fall within the chemical space or applicability domain (AD) of the compounds in the training dataset.

Enriched molecular fragments in the NAA datasets

The molecular fragments or substructures (or chemical features) enriched within the active and inactive compounds in the NAA dataset was searched with the Molecular Substructure (MoSS) node in KNIME (Fig 1) [27,28]. Minimum and maximum fragment sizes were set to 1 and 100 respectively. Pure carbon fragments were ignored and the ring mining option was enabled (set at 3 to 8 to avoid finding fragments with partial rings). The algorithm used is the Christian Borgelt's MoSS implementation [45].

Results and discussion

In the present study, we have trained and evaluated four antiplasmodial activity classification models based on a combination of molecular descriptors and molecular fingerprints of natural products with antiplasmodial activity (NAA).

Molecular descriptors and molecular fingerprints

A total of 117 molecular descriptors were generated with RDKit Descriptors Calculation node in KNIME [46,47] for the compounds in the NAA dataset. The resultant data was then pre-processed, as described under the method section, before passing on to the "Feature Elimination" meta-node [48] to remove redundant molecular descriptors. Approximately 35% of the molecular descriptors were removed from the NAA dataset. The molecular fingerprints of the compounds in NAA datasets were also generated with RDKit Fingerprint node in KNIME [36,37]. The remaining 76 molecular descriptors were combined with the molecular fingerprints and used to train the classification models.

Training of classifier models and cross-validation

Four classifier models were trained with natural products with in-vitro antiplasmodial activities (NAA) (using Weka version 3.6 node in KNIME): Voted Perceptron (VP), Naïve Bayesian (NB), Random Forest (RF) and Sequential Minimization Optimization (SMO). Running on a Dell Vostro laptop (Intel Core i3-2328M CPU @ 2.20 GHz x 4), SMO was the slowest in terms of program runtime to build one model (2.88 seconds); the NB was the fastest (0.12 seconds) followed by VP (1.05 seconds) and RF (1.31 seconds). A total of 1147 NAA (labelled as 70% active and 30% inactive) was used in this study. This was divided into 917 NAA for training cum 10-fold cross-validation of the classifier model and 230 NAA as the independent test dataset. Misclassification cost was set to two for false negatives (FN).

The values of the accuracy (percentage of correctly classified compounds) of the classifier models over the 10 fold cross-validation are presented in Fig 2. Accuracy may be defined, specifically for this study, as the proportion of compounds that were correctly classified as active and inactive (i.e. the number of compounds correctly classified divided by the total number of compounds classified multiply by 100). From the results (Fig 2), SMO and RF classifier models showed greater predictive accuracies than the NB and VP classifier models for the NAA

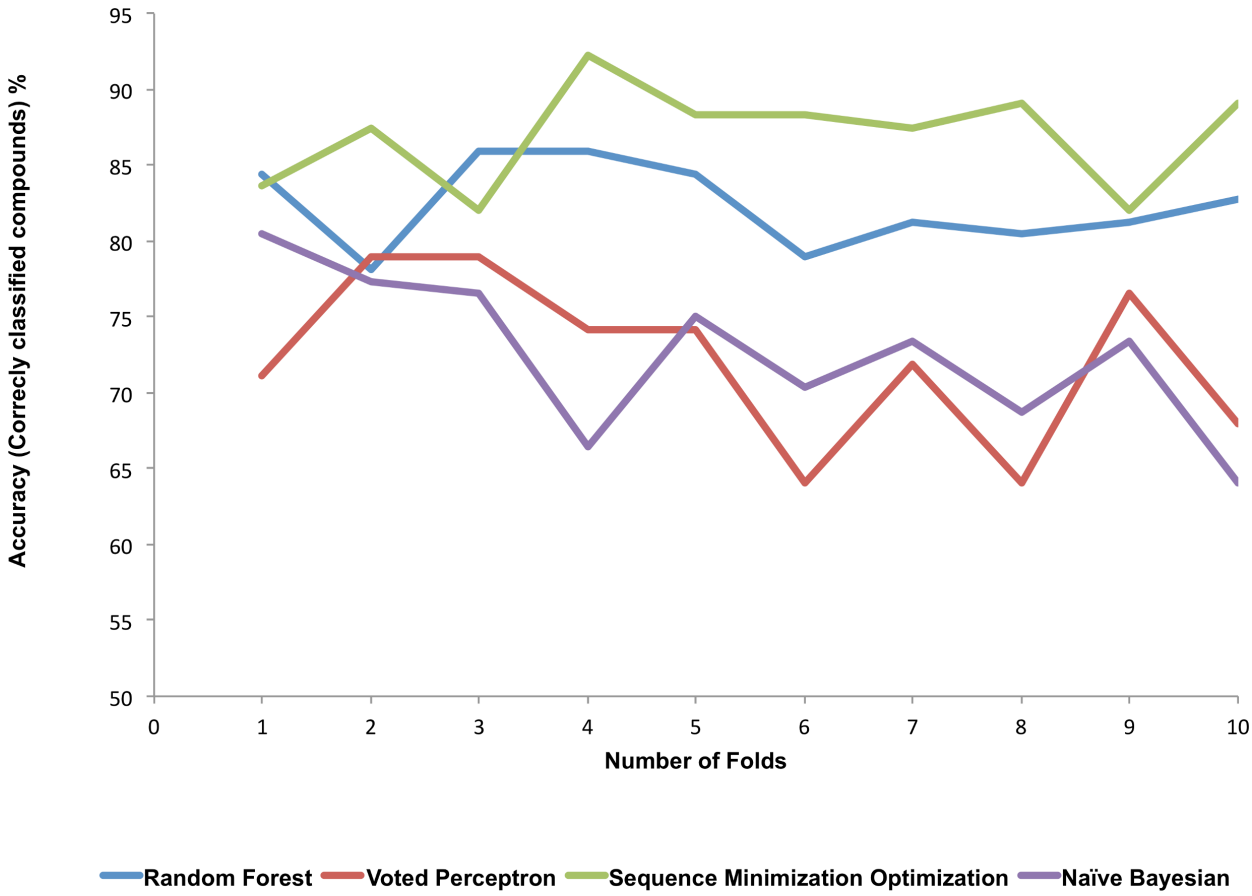

Fig 2. The graph shows values of the accuracy (percentage of correctly classified compounds) of the four classifier models over the 10 folds cross-validation. The sequence minimization optimization (SMO) and Random Forest (RF) classifier models showed greater predictive accuracy than the Naïve Bayesian (NB) and Voted Perceptron (VP) classifier models.

https://doi.org/10.1371/journal.pone.0204644.g002

dataset. Moreover, the values of the predictive accuracy were fairly consistent over the 10 fold cross-validation for SMO and RF judging by the slope of approximately 0.2 for the data points on the graph. This is an indication of the consistency of the predictive abilities of these classifier models.

The goal here is to see the performance of the trained classifier models to predict the activity class of 10 randomly selected test datasets. From the result (Fig 2), we may conclude that SMO and RF classifier models showed good and fairly consistent predictive performance of the 10 randomly selected test datasets. However, we used all classifier models generated to predict the bioactivity class of the independent NAA test dataset that was not included in the training and cross-validation dataset.

Prediction of bioactivity class of an independent NAA test dataset

The classifier models (Sequential Minimization Optimization (SMO), Random Forest (RF), Voted Perceptron (VP) and Naïve Bayesian (NB)), previously trained and cross-validated as described earlier, were used to predict the bioactivity class of an independent test dataset of natural products with *in-vitro* antiplasmodial activities (NAA). The performances of the classifier models were evaluated using accuracy, Kappa statistics and Receiver Operating Characteristic curve.

Accuracy. From the results (Table 2), the Sequence Minimization Optimization (SMO) of Support Vector Machine model showed the highest accuracy (85.94%) followed by the Random Forest (RF), 82.81%. Naïve Bayes (NB) and Voted perceptron (VP) models displayed accuracy just above 70% (73.05% and 71.48% respectively). When we compared the results from the fusion model to the individual models (RF and SMO), we see that the accuracy did not change.

The objective here was to identify the classifier model trained with both molecular descriptors and molecular fingerprints that best predict the bioactivity class of the independent NAA test dataset. From the results, we may conclude that SMO of SVM and RF models are the most suitable classifier models for NAA. Though accuracy provided an overall estimation of the performance of the classifier models, one limitation to the use of accuracy as a metric for assessing predictive performance of classifier models is "accuracy paradox" (i.e. a classifier model with a given level of accuracy may have greater predictive power than models with high accuracy). Therefore less biased metrics like the Kappa statistics and area under Receiver Operating Curve (ROC) were used as a more objective evaluator of the predictive powers of the classifier models.

Kappa statistics. The results (Table 2) showed that the kappa statistics of SMO and RF classifier models (0.72 and 0.65 respectively), like their accuracy values, were higher than that of the other classifier models in this study. The results also showed that the fusion model showed a very slight increase in its Kappa statistics when compared to the individual models (RF and SMO). The value of the Kappa statistics is often used as a measure of consistency or agreement between the *"ground truth"* (the actual class of each compound to be classified) and classifier models' classification (the class assigned to the compounds by the classifier model). It accounts for the chance of random classification of compounds into the two bioactive classes (Active and Inactive). Kappa statistic values of 1 suggest a perfect agreement between the *"ground truth"* and classifier models' classification. Judging by the kappa statistics of SMO and RF models (Table 2), which are closer to 1 than that of NB and VP models, we concluded that SMO and RF classifier models showed the best predictive power as similarly observed with the use of accuracy as the evaluator of the classifier models.

Receiver operating characteristic plot (ROC). Fig 3 shows the Receiver Operating Characteristic (ROC) curve of the classifier models trained and evaluated. Receiver Operating Characteristic (ROC) curve is a graphical plot that shows the performance of a binary classifier model as its discrimination threshold is varied.

ROC is a plot of the true positive rate (Sensitivity) against the false positive rate (1 –Specificity) at various threshold settings. The diagonal grey line represents classifier models that randomly assign compounds to bioactivity class. The blue line shown in the ROC plot of Voted perceptron represents classifier models that perfectly predict bioactivity class of compounds.

Table 2. Evaluation parameters from the prediction of bioactivity class of an independent NAA test dataset by the four classifier models used in this study.

Classifier Models	Accuracy (%)	Kappa Statistics	Area Under Curve (ROC)
Random Forest (RF)	82.81	0.65	0.91
Voted Perceptron (VP)	71.48	0.42	0.72
Sequence Minimization Optimization (SMO) of Support Vector Machine	85.94	0.72	0.86
Naïve Bayesian (NB)	73.05	0.45	0.74
Fused Model (RF and SMO)	82.03	0.68	0.92

Comments: RF: Random forest of 10 trees, each constructed while considering 11 random features. Out of bag error: 0.1797. SMO: The polynomial kernel. Fused Model (RF and SMO): The predictions from RF and SMO were combined (mean) using the "Prediction Fusion" node in KNIME [38].

https://doi.org/10.1371/journal.pone.0204644.t002

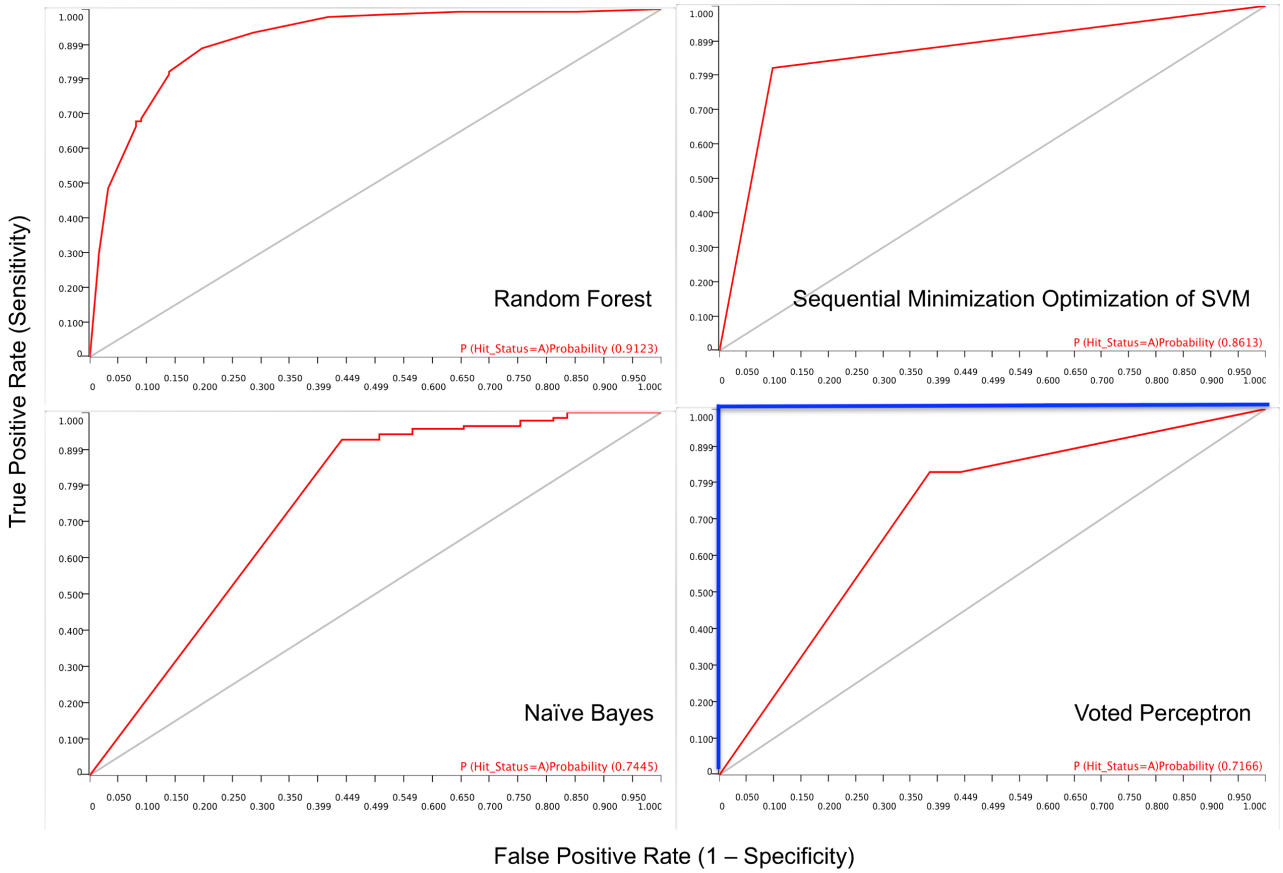

Fig 3. The Receiver operating characteristics (ROC) curve for the four classifier models. The diagonal grey line represents classifier models that randomly assign compounds to bioactivity class (and will have an area under the curve (AUC) of 0.5). The blue line shown in the ROC curve of Voted perceptron (will have an AUC of 1.0) represents classifier models that perfectly predict bioactivity class of compounds. The red line is the ROC curve from the predictions by the four classifier models. The area under the ROC curve (AUC), a measure of bioactivity class discriminatory power of a classifier model, is shown on each ROC curve.

https://doi.org/10.1371/journal.pone.0204644.g003

The red line is the ROC curve from the predictions by the four classifier models in this study. In contrast to NB and VP classifier models, SMO, RF and the fusion classifier models showed ROC curves that were initially very close to the true positive rate axis (i.e. minimizing false positive rate (maximizing specificity) and maximizing true positive rates (maximizing sensitivity)). An optimum prediction aims to maximize sensitivity and specificity. However in all the models, as the threshold changes the false positive rate increases (i.e. the specificity decreases) and the true positive rate (i.e. sensitivity) approaches its maximum value. Hence low specificity values may lead to high incidence of false positives (i.e. detecting inactive compounds as active). It is therefore expedient to choose the threshold that will have good specificity and thus avoid investing resources to synthesize and conduct bioassays for compounds that may not be active and fail along the drug development pipeline [49–52].

Area under the receiver operating characteristic curve (AUC). The values of the area under the ROC curve (AUC) are shown in Table 2. The RF had the highest value of AUC of 0.91 followed by SMO with an AUC value of 0.86. The VP and NB classifier models showed AUC values of 0.72 and 0.74 respectively. The fusion model showed AUC that is higher than AUC of SMO but similar to the AUC seen for RF. The ROC The area under the ROC curve (AUC) is a measure of how well a model can discriminate between two classes in a dataset (e.g.

active and inactive compounds) [53]. In this study, AUC depicts the probability that the active class predicted by the classifier models for a randomly selected compound will exceed that of a randomly selected non-active class [54]. Where the prediction of the bioactivity class of compounds is purely random, the AUC will be equal to 0.5 (i.e. the ROC curve will coincide with the diagonal line). When the prediction results in perfect separation of the bioactivity class of the compounds, i.e. where there is no overlapping of the distributions of the bioactivity classes, the area under the ROC curve will be one.

The values of the AUC for the four classifier models indicate that the discriminatory or predictive power (separation of the bioactivity class of the compounds) of the models range from fair (0.7–0.8) to excellent (> 0.9). The discriminating powers of the classifier models, judging by the AUC, were thus: RF, SMO, NB and VP in decreasing order of discriminating power to predict a bioactive class of the compounds in the NAA dataset. Overall, the nature of the ROC plot and the higher AUC values of RF and SMO suggest their suitability as good classifier models for the NAA dataset used in this study.

Applicability domain (AD) of the classifier models

Applicability Domain (AD) of the classifier models refers to the chemical space, defined by the training set, within which a test compound should be in order for its bioactivity class to be reliably predicted. In this present study, the AD of the models was defined with the training and cross-validation dataset and its validity evaluated on the independent test dataset. Principal Component Analysis (PCA) was used to define the AD of the models and to map the test dataset (active and inactive compounds) in their respective chemical spaces.

Fig 4 is the visualization of the first three principal components of the compounds in the training and cross-validation dataset (Fig 4 (X)) and compounds in the independent test dataset (Fig 4 (Y)) for the NAA dataset. From Fig 4(Y) and 4(X), we observed that almost all compounds in the test dataset fell within the chemical space or AD of the training dataset used to build the classifier models.

The results also revealed no clear boundary between the active and inactive compounds in the training dataset and the independent test dataset for NAA. This implies some level of chemical structural similarity amid the active and inactive compounds in the datasets, which may pose a restriction on the discriminatory ability of the models. Overall, this analysis enabled the identification of the AD for the models built in this study. Therefore the models may reliably predict new compounds that fall within this AD.

Enriched molecular fragments in the NAA datasets

We sought to understand the molecular substructures (or chemical features) associated with antiplasmodial activity and inactivity of compounds in the NAA (natural products with *in-vitro* antiplasmodial activities) dataset. To this end, we used the Molecular Substructure (MoSS) node in KNIME to search for most common molecular substructures in the active and inactive compounds in the NAA dataset.

A total of 52 most common molecular substructures from active compounds (717 compounds) and 48 most common molecular substructures from inactive compounds (323) were identified. The molecular similarities amongst the substructures from the active and inactive compounds were estimated and projected in a three dimensional (3D) space (Fig 5).

From these results, most of the substructures from the active and inactive compounds overlapped in the 3D space indicating their high molecular similarity. However, some of the substructures from the active and inactive compounds occupy a distinct region of the 3D space (Fig 5). These include hydroxyisoquinoline and isoquinoline substructures from active

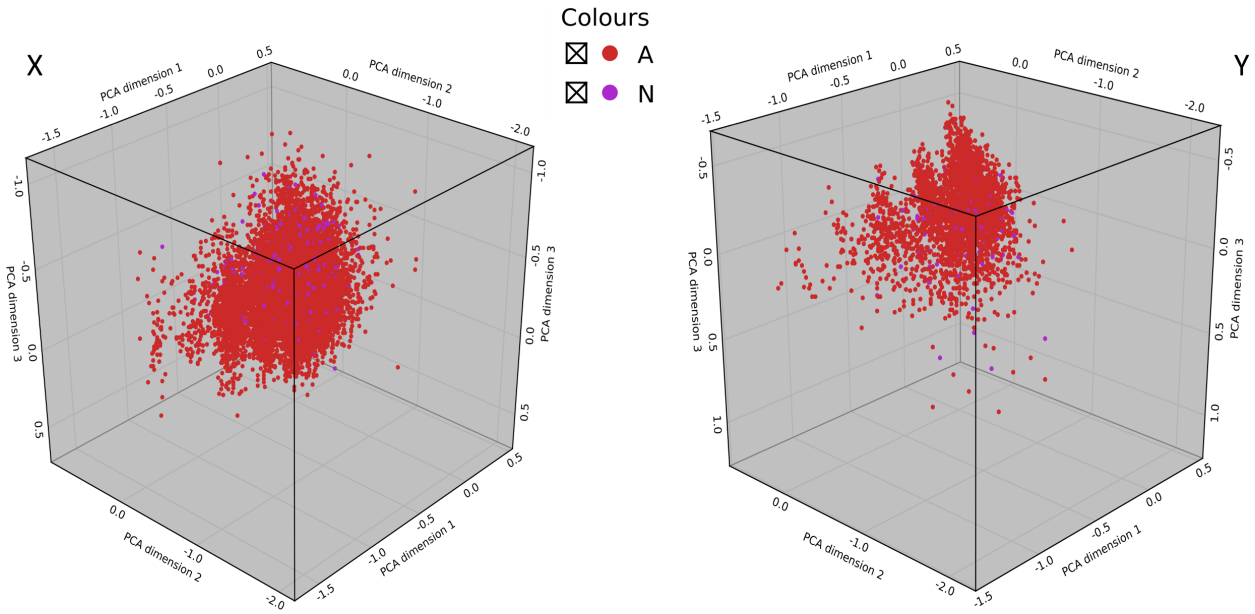

Fig 4. A visualization of the applicability domain (chemical space) of the of classifier models built in this study. Active compounds (red dots) and inactive compounds (purple dots) are represented using the first three Principal Components. Panel X depicts the range of Principal Components of compounds in the training set that define the applicability domain (AD). Panel Y shows that almost all compounds in the test set fell within the AD of the defined by the training set. Therefore, classifier models generated in this study can reliably predict the bioactivity class of new compounds that fall within this AD. *NAA: natural products with in-vitro antiplasmodial activity.*

https://doi.org/10.1371/journal.pone.0204644.g004

compounds and hydroxyflavone from inactive compounds. These substructures may be determinants of antiplasmodial activities and may guide rational selection and design of active antiplasmodial compounds. A closer at difference in the functional groups between the active NAA dataset (A) and inactive NAA dataset (N) revealed the following: Akylamine (29% in A, 13% in N); Aromatic amine (1.5% in A, 3% in N); Basic nitrogen (36% in A, 14% in N); Acidic oxygen (4% in A, 9% in N). In all, the amine chemical group (specifically alkyl amines and basic nitrogen) was confirmed to be essential for antimalarial activity in active NAA dataset.

Benefits of models from machine learning (*in-silico* compound screening)

To illustrate the benefit of the machine learning and the resultant classifier models, the Sequential Minimization Optimisation (SMO) and Random Forest classifier models, adjudged the top classifier models in this study, were used to screen 450 natural compounds of a private natural product chemical library from InterBioScreen (http://www.ibscreen.com). The results (S2 Table) showed that the SMO classifier model predicted that 39% of the compounds will possess active antiplasmodial activities while Random Forest predicted a higher proportion of the natural product chemical library as active (87%). Although there was a significant difference in the proportion of compounds predicted as active by the classifier models the output from the RF classifier model may be less reliable due to the tendency for RF models to overfitting data [33,34]. The two classifier models showed consistent antiplasmodial bioactivity class prediction for 54% of the compounds in the natural product chemical library.

The natural compounds predicted as active, which are readily available from InterBioScreen and other chemical libraries, may be prioritized and readily purchased for *in-vitro* antiplasmodial screening. Overall, these results attest to the importance of bioactivity predictive models

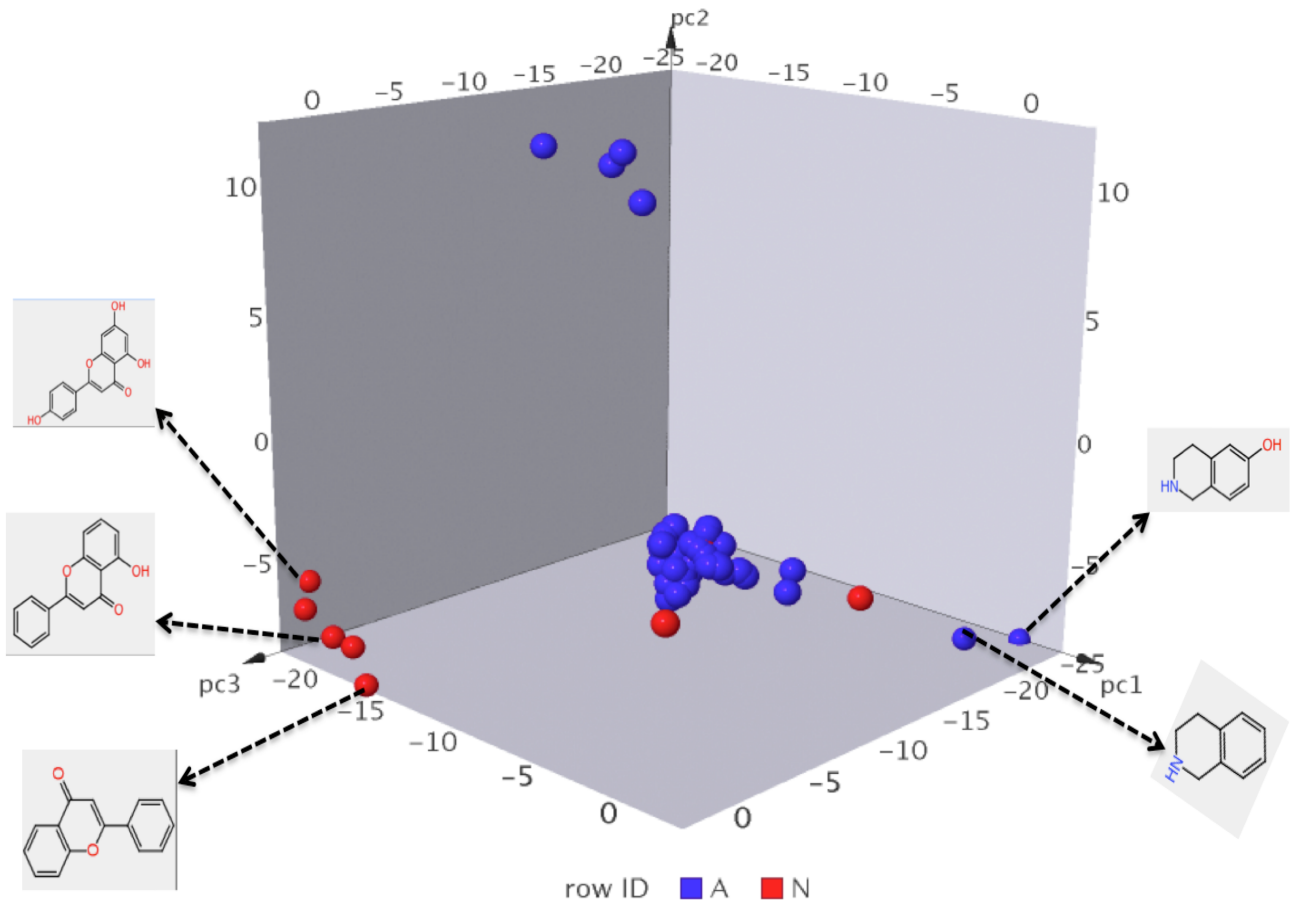

Fig 5. Chemical features from active and inactive compounds from NAA dataset. The blue markers represent most common substructures from active compounds (IC50 $\leq$ 10 μM) while the red markers represents most common substructures from inactive compounds (IC50 $>$ 10 μM). The most common substructures were projected in a three-dimension (3D) space based on molecular similarity. Some of the most common substructures that are peculiar to the active and inactive compounds are highlighted. This may guide rational selection and design of active antiplasmodial compounds. *NAA*: *natural products with in-vitro antiplasmodial activity*.

https://doi.org/10.1371/journal.pone.0204644.g005

built with machine learning algorithms that are capable of effective and efficient learning from existing bioactivity data and predicting biological activities *in-silico* to modern drug discovery.

Conclusions

In this study, we used machine learning as a method to build various antimalarial predictive models that can predict the bioactivity class of natural products. The classifier models that were most suitable for the dataset (natural products with *in-vitro* antiplasmodial activities) were identified. These models were used, *in-silico*, to annotate potential antimalarial compounds in a large natural product library. Such compounds may be prioritized for the more expensive *in-vitro* bioactivity screening. In addition, we generated a pool of chemical features that were present within active and inactive natural products with *in-vitro* antiplasmodial activities (NAA) used in this study. Such chemical features from active NAA in conjunction with the molecular scaffolds that may be identified from the active NAA could be valuable in designing antimalarial specific virtual compound library.

The knowledge of the classifier models that provide the most accurate prediction of the desired bioactivity for a particular class of compounds will enable medicinal chemist to pre-screen

compounds prior to the expensive step of synthesis and *in-vitro* assay. Accurate prediction of bioactivity class of compounds will improve decision-making processes in antimalarial drug design and development to achieve better and cost-effective outcomes (i.e. drug candidate for malaria). Overall, knowledge provided by this study could contribute significantly to and accelerate the ongoing efforts for antimalarial drug discovery, especially from natural products.

Supporting information

S1 Table. Natural products that have been tested for *in-vitro* antiplasmodial activities (NAA) compiled in-house from literature, PhD and Masters Theses and public chemical databases.
(PDF)

S2 Table. The antiplasmodial bioactivity class predictions of 450 natural compounds from a private natural product chemical library from InterBioScreen (http://www.ibscreen.com) by sequential minimization optimisation (SMO) and random forest classifier models.
(PDF)

Acknowledgments

The authors wish to acknowledge William Jose, Godinez Navarro and Azzaoui Kamal all of Novartis Institute for Biomedical Research (NIBR), Basel Switzerland for reading this paper and making valuable contributions. The South African Research Chairs Initiative of the Department of Science and Technology (DST) and the National Research Foundation (NRF) of South Africa supported this work.

Author Contributions

Conceptualization: Samuel Egieyeh, Alan Christoffels.

Data curation: Samuel Egieyeh.

Funding acquisition: Alan Christoffels.

Methodology: Samuel Egieyeh.

Supervision: James Syce, Sarel F. Malan, Alan Christoffels.

Writing – original draft: Samuel Egieyeh.

Writing – review & editing: Samuel Egieyeh, James Syce, Sarel F. Malan, Alan Christoffels.

References

1. Sharma I, Sullivan M, McCutchan TF. The in vitro anti-malarial activity of novel semi synthetic nocathiacin I antibiotics. Antimicrob Agents Chemother 2015:AAC. 04294–14.
2. El Tahir MN. The impact of malaria on labour use and efficiency in the Sudan. Soc Sci Med 1993; 37 (9):1115–1119. PMID: 8235750
3. Russell S. The economic burden of illness for households in developing countries: a review of studies focusing on malaria, tuberculosis, and human immunodeficiency virus/acquired immunodeficiency syndrome. Am J Trop Med Hyg 2004 Aug; 71(2 Suppl):147–155. PMID: 15331831
4. Gupta S, Hill AV, Kwiatkowski D, Greenwood AM, Greenwood BM, Day KP. Parasite virulence and disease patterns in Plasmodium falciparum malaria. Proc Natl Acad Sci U S A 1994 Apr 26; 91(9):3715–3719. PMID: 8170975
5. Bull PC, Marsh K. The role of antibodies to Plasmodium falciparum-infected-erythrocyte surface antigens in naturally acquired immunity to malaria. Trends Microbiol 2002; 10(2):55–58. PMID: 11827798

6. Kaestli M, Cockburn IA, Cortes A, Baea K, Rowe JA, Beck HP. Virulence of malaria is associated with differential expression of Plasmodium falciparum var gene subgroups in a case-control study. J Infect Dis 2006 Jun 1; 193(11):1567–1574. https://doi.org/10.1086/503776 PMID: 16652286
7. Klein E. Antimalarial drug resistance: a review of the biology and strategies to delay emergence and spread. Int J Antimicrob Agents 2013; 41(4):311–317. https://doi.org/10.1016/j.ijantimicag.2012.12.007 PMID: 23394809
8. Burgess DJ. Evolution: Taking advantage of drug resistance. Nature Reviews Genetics 2014; 15 (3):147–147.
9. Severini C, Menegon M. Resistance to antimalarial drugs: An endless world war against Plasmodium that we risk losing. Journal of Global Antimicrobial Resistance 2015.
10. Christensen SB, Kharazmi A. Antimalarial natural products. Bioactive Compounds from Natural Sources 2001:379.
11. Batista R, Silva Ade J, De Oliveira AB. Plant-derived antimalarial agents: new leads and efficient phytomedicines. Part II. Non-alkaloidal natural products. Molecules 2009; 14(8):3037–3072. https://doi.org/10.3390/molecules14083037 PMID: 19701144
12. Xu Y, Pieters L. Recent developments in antimalarial natural products isolated from medicinal plants. Mini reviews in medicinal chemistry 2013; 13(7):1056–1072. PMID: 22974400
13. Mojab F. Antimalarial natural products: a review. Avicenna Journal of Phytomedicine 2012; 2(2):52. PMID: 25050231
14. Davis RA, Buchanan MS, Duffy S, Avery VM, Charman SA, Charman WN, et al. Antimalarial activity of pyrroloiminoquinones from the Australian marine sponge Zyzzya sp. J Med Chem 2012; 55(12):5851–5858. https://doi.org/10.1021/jm3002795 PMID: 22686608
15. Mayer AM, Rodríguez AD, Berlinck RG, Fusetani N. Marine pharmacology in 2007–8: Marine compounds with antibacterial, anticoagulant, antifungal, anti-inflammatory, antimalarial, antiprotozoal, antituberculosis, and antiviral activities; affecting the immune and nervous system, and other miscellaneous mechanisms of action. Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology 2011; 153(2):191–222.
16. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 2009 Jul; 37(Web Server issue):W623–33. https://doi.org/10.1093/nar/gkp456 PMID: 19498078
17. Spangenberg T, Burrows JN, Kowalczyk P, McDonald S, Wells TN, Willis P. The open access malaria box: a drug discovery catalyst for neglected diseases. PloS one 2013; 8(6):e62906. https://doi.org/10.1371/journal.pone.0062906 PMID: 23798988
18. Bathurst I, Hentschel C. Medicines for Malaria Venture: sustaining antimalarial drug development. Trends Parasitol 2006; 22(7):301–307. https://doi.org/10.1016/j.pt.2006.05.011 PMID: 16757213
19. James G, Witten D, Hastie T. An Introduction to Statistical Learning: With Applications in R. 2014.
20. Periwal V, Rajappan JK, Open Source Drug Discovery Consortium, Jaleel AU, Scaria V. Predictive models for anti-tubercular molecules using machine learning on high-throughput biological screening datasets. BMC Res Notes 2011 Nov 18; 4:504-0500-4-504.
21. Jamal S, Periwal V, Open Source Drug Discovery Consortium, Scaria V. Predictive modeling of anti-malarial molecules inhibiting apicoplast formation. BMC Bioinformatics 2013 Feb 15; 14:55-2105-14-55.
22. Jamal S, Scaria V. Cheminformatic models based on machine learning for pyruvate kinase inhibitors of Leishmania mexicana. BMC Bioinformatics 2013 Nov 19; 14:329-2105-14-329.
23. Batista R, De Jesus Silva Júnior, Ademir, De Oliveira AB. Plant-derived antimalarial agents: new leads and efficient phytomedicines. Part II. Non-alkaloidal natural products. Molecules 2009; 14(8):3037–3072. https://doi.org/10.3390/molecules14083037 PMID: 19701144
24. Kaur K, Jain M, Kaur T, Jain R. Antimalarials from nature. Bioorg Med Chem 2009; 17(9):3229–3256. https://doi.org/10.1016/j.bmc.2009.02.050 PMID: 19299148
25. Frederich M, Tits M, Angenot L. Potential antimalarial activity of indole alkaloids. Trans R Soc Trop Med Hyg 2008 Jan; 102(1):11–19. https://doi.org/10.1016/j.trstmh.2007.10.002 PMID: 18035385
26. Nogueira CR, Lopes LM. Antiplasmodial natural products. Molecules 2011; 16(3):2146–2190.
27. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. KNIME: The Konstanz information miner.: Springer; 2008.
28. Meinl T, Cebron N, Gabriel TR, Dill F, Kötter T, Ohl P, et al. The Konstanz Information Miner 2.0. 2009.
29. Nidhi, Glick M, Davies JW, Jenkins JL. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. Journal of chemical information and modeling 2006; 46(3):1124–1133. https://doi.org/10.1021/ci060003g PMID: 16711732

30. Zhang H, Yu P, Xiang M, Li X, Kong W, Ma J, et al. Prediction of drug-induced eosinophilia adverse effect by using SVM and naïve Bayesian approaches. Med Biol Eng Comput 2015:1–9. https://doi.org/10.1007/s11517-014-1207-1
31. Chu W, Keerthi SS. Support vector ordinal regression. Neural Comput 2007; 19(3):792–815. https://doi.org/10.1162/neco.2007.19.3.792 PMID: 17298234
32. Chang C, Lin C. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2011; 2(3):27.
33. Sheridan RP. Three useful dimensions for domain applicability in QSAR models using random forest. Journal of chemical information and modeling 2012; 52(3):814–823. https://doi.org/10.1021/ci300004n PMID: 22385389
34. Singh H, Singh S, Singla D, Agarwal SM, Raghava GP. QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest. Biol Direct 2015 Mar 25; 10:10-015-0046-9.
35. Martišius I, Šidlauskas K, Damaševičius R. Real-Time Training of Voted Perceptron for Classification of EEG Data. International Journal of Artificial Intelligence 2013; 10(S13):41–50.
36. Loukeris N, Eleftheriadis I. Further Higher Moments in Portfolio Selection and A Priori Detection of Bankruptcy, Under Multi-layer Perceptron Neural Networks, Hybrid Neuro-genetic MLPs, and the Voted Perceptron. International Journal of Finance & Economics 2015.
37. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter 2009; 11(1):10–18.
38. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. KNIME: The Konstanz information miner.: Springer; 2008.
39. Landrum G. RDKit: Open-source cheminformatics, http 2014.
40. Guyon I, Elisseeff A. An introduction to variable and feature selection. The Journal of Machine Learning Research 2003; 3:1157–1182.
41. Ji S, Carin L. Cost-sensitive feature acquisition and classification. Pattern Recognit 2007; 40(5):1474–1485.
42. Drummond C, Holte RC. Cost curves: An improved method for visualizing classifier performance. 2006.
43. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R. Comparison of different approaches to define the applicability domain of QSAR models. Molecules 2012; 17(5):4791–4810. https://doi.org/10.3390/molecules17054791 PMID: 22534664
44. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. QSAR applicability domain estimation by projection of the training set descriptor space: a review. ATLA-NOTTINGHAM- 2005; 33(5):445.
45. Moss: a program for molecular substructure mining. Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations: ACM; 2005.
46. Landrum G. RDKit Documentation. Release 2013; 1:1–79.
47. M P Mazanetz, R J Marmon, C BT Reisser, Morao I. Drug discovery applications for KNIME: an open source data mining platform. Current topics in medicinal chemistry 2012; 12(18):1965–1979. PMID: 23110532
48. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. KNIME-the Konstanz information miner: version 2.0 and beyond. AcM SIGKDD explorations Newsletter 2009; 11(1):26–31.
49. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nature reviews Drug discovery 2004; 3(8):711–716. https://doi.org/10.1038/nrd1470 PMID: 15286737
50. Ferri N, Siegl P, Corsini A, Herrmann J, Lerman A, Benghozi R. Drug attrition during pre-clinical and clinical development: understanding and managing drug-induced cardiotoxicity. Pharmacol Ther 2013; 138(3):470–484. https://doi.org/10.1016/j.pharmthera.2013.03.005 PMID: 23507039
51. Roberts RA, Kavanagh SL, Mellor HR, Pollard CE, Robinson S, Platz SJ. Reducing attrition in drug development: smart loading preclinical safety assessment. Drug Discov Today 2014; 19(3):341–347. https://doi.org/10.1016/j.drudis.2013.11.014 PMID: 24269835
52. Barnes PJ, Bonini S, Seeger W, Belvisi MG, Ward B, Holmes A. Barriers to new drug development in respiratory disease. Eur Respir J 2015 May; 45(5):1197–1207. https://doi.org/10.1183/09031936.00007915 PMID: 25931481
53. Jiménez-Valverde A. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Global Ecol Biogeogr 2012; 21(4):498–507.
54. Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian J Intern Med 2013 Spring; 4(2):627–635. PMID: 24009950