COMPREHENSIVE REVIEW

Systematic review on the application of machine learning to quantitative structure–activity relationship modeling against *Plasmodium falciparum*

Osondu Everestus Oguike[1,2] · Chikodili Helen Ugwuishiwu[1,2] · Caroline Ngozi Asogwa[1,2] · Charles Okeke Nnadi[1,3] · Wilfred Ofem Obonga[1,3] · Anthony Amaechi Attama[1,4]

Abstract
Malaria accounts for over two million deaths globally. To flatten this curve, there is a need to develop new and high potent drugs against *Plasmodium falciparum*. Some major challenges include the dearth of suitable animal models for anti-*P. falciparum* assays, resistance to first-line drugs, lack of vaccines and the complex life cycle of *Plasmodium*. Gladly, newer approaches to antimalarial drug discovery have emerged due to the release of large datasets by pharmaceutical companies. This review provides insights into these new approaches to drug discovery covering different machine learning tools, which enhance the development of new compounds. It provides a systematic review on the use and prospects of machine learning in predicting, classifying and clustering IC_{50} values of bioactive compounds against *P. falciparum*. The authors identified many machine learning tools yet to be applied for this purpose. However, Random Forest and Support Vector Machines have been extensively applied though on a limited dataset of compounds.

Keywords QSAR · *Plasmodium falciparum* · Machine learning · Drug discovery

Abbreviations

ADME	Absorption, distribution, metabolism, elimination
AI	Artificial intelligence
AM1	Austin model 1
ANN	Artificial neural networks
AUC	Area under the curve
BTR	Boosted trees regression
COVID-19	Coronavirus diseases-2019
DDD	Drug discovery and development
DF	Discriminant functions
DL	Deep learning
DRC	Democratic Republic of Congo
DT	Decision tree
GCNN	Graph convolutional neural networks
GLM	Generalized linear model
GMP	Global Malaria Programme
GRNN	General regression neural network
GSK	Glaxo-Smithkline
GTS	Global technical strategy
HBHI	High burden high impact
HIV/AIDS	Human immunovirus/acquired immunodeficiency syndrome
HTS	High-throughput screening
IRS	Indoor residual spraying
ITN	Insecticide-treated nets
JC	J48 classifier
J–P	Jarvis–Patrick
kNN	K-nearest neighbors
LMD	Low mode dynamic
MDS	Multidimensional scaling
ML	Machine learning
MLR	Multilinear regression
MVA	Multivariate analysis

✉ Charles Okeke Nnadi
charles.nnadi@unn.edu.ng

1 Machine Learning Research Group, University of Nigeria, Nsukka 410001, Enugu State, Nigeria

2 Department of Computer Science, Faculty of Physical Sciences, University of Nigeria, Nsukka 410001, Enugu State, Nigeria

3 Deprtment of Pharmaceutical and Medicinal Chemistry, Faculty of Pharmaceutical Sciences, University of Nigeria, Nsukka 410001, Enugu State, Nigeria

4 Department of Pharmaceutics, Faculty of Pharmaceutical Sciences, University of Nigeria, Nsukka 410001, Enugu State, Nigeria

PDB	Protein data bank
PfENR	*Plasmodium falciparum* Enoyl acyl carrier protein reductase
PLS-DA	Partial least square discriminant analysis
QSAR	Quantitative structure–activity relationship
QSPR	Quantitative structure–property relationship
RF	Random Forest
RMSE	Root mean square error
ROC	Receiver operating characteristic
SAR	Structure–activity relationship
SARS-CoV-2	Severe acute respiratory syndrome–coronavirus 2
SVM	Support vector machine
WHO	World Health Organisation
XGB	XGBoost

Introduction

New drug compounds active against pathogenic organisms or parasites are discovered through a very rigorous process involving many stages and huge human and material resources input. Besides, drug discovery takes a long time (up to 12–15 years), making it difficult to introduce new drugs to combat emerging or resistant strains of existing diseases [1, 2]. This process involves the identification of candidates, synthesis, characterization, validation, optimization, screening and assays for therapeutic efficacy. Since the introduction of Artificial Intelligence (AI), many processes have been made easier and faster than before, because of the ability of the models utilized to handle an unprecedented cache of data within a very short time [3]. Thus, the application of AI to drugs and development is a welcome development as it is expected to shorten the time to market many drug candidates found to be active against parasites and pathogenic organisms. A subunit of AI known as Machine Learning (ML) has been widely applied to Drug Discovery and Development (DDD) [4].

DDD pipelines are long, complex and depend on numerous factors. ML approaches provide a set of tools that can improve discovery and data-driven decision-making for well-specified questions with abundant, high-quality data [5]. The growth of High Throughput Screening (HTS) data has increased the importance of ML tools at virtually all phases of drug discovery. ML has the potential to speed up the process and reduce failure rates in DDD [5]. These patterns form the basis for building models that are effectively applied to prioritize compounds for the subsequent phases. ML techniques can assist in the identification of false leads at an early stage and also facilitate the understanding of structure–activity relationships (SARs) [6].

This paper presents a systematic review on the use and prospects of ML in predicting, classifying and clustering IC_{50} values of compounds active against *P. falciparum.* Fundamentally, ML is the practice of using classification, regression or clustering algorithms to describe data, learn from it and then decide or predict about the future state of any new dataset. Classification is the process of recognizing, understanding and grouping ideas and objects into preset categories or sub-populations [2]. Using pre-categorized input training datasets, ML uses a variety of algorithms to classify future datasets as shown in Fig. 6A. Classification algorithms are predictive calculations used to assign data to preset categories by analyzing sets of training data [7]. Predictive computational models enable one to understand the correlation between descriptors and the biological properties (activities), that is, to computationally screen large molecular datasets thereby offering a possibility to improve the hit rate and thereby reducing the overall costs of drug discovery [8]. Due to the constant emergence of parasitic resistance to the current antimalarial drugs, the discovery of new drug candidates is a major global health priority [9, 10]. Previous works in ML-based tropical diseases research, including malaria and other diseases, have shown effectiveness in drug discovery [11]. In previous studies also, several algorithms have been employed in classifying the IC_{50} value of compounds against *P. falciparum* including Decision Tree (DT), K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), PLS Discriminant Analysis (PLS-DA). These approaches have shown statistical significance in performance [12].

Malaria as a global challenge

Global burden of malaria

Malaria remains one of the most life-threatening diseases caused by the blood-borne protozoan parasites of the genus *Plasmodium*. Five species of *Plasmodium* are known to cause one form of infection or the other to humans across the globe [13, 14]. According to the World Health Organisation (WHO), *P. falciparum* is the most deadly, most common causative agent and also the most prevalent species in sub-Saharan Africa. Southeast Asia, Western Pacific, Eastern Mediterranean and Latin America are currently burdened by *P. vivax* [15]. Malaria due to *P. ovale* leads to life-threatening symptoms but was previously considered benign. *P. malariae*, like *P. ovale*, malaria is not severe in humans, while *P. knowlesi* is the most prevalent species in Southeast Asia. Symptoms of malaria vary from species to species; however, paroxysms, anemia and headaches are common in all cases of human malaria infection. *P. falciparum* results in respiratory distress, deep capillaries blockade, cerebral

malaria and neurological disorders and eventual death, if untreated [16]. Malaria is transmitted through the saliva of the female anopheles mosquitoes. The transmission cascade is complex and involves both sexual and asexual stages in mosquitoes and humans [14].

The global trends in malaria incidences show somewhat dramatic, complex and highly unpredictable episodes. Malaria is still endemic in 87 countries with 29 countries accounting for about 95% of all recorded cases [15]. The leading most endemic countries—Nigeria, the Democratic Republic of Congo (DRC), Uganda, Mozambique and Niger—accounted for about 51% of all cases globally [15]. The WHO estimated that 409,000 malaria deaths occurred globally in 2019; 67% of which were recorded in under 5 year old children, 95% in 31 countries and 23% in Nigeria alone [17]. Despite the global burden of malaria, it is estimated that over a billion cases and millions of death have been averted in the last 20 years. A total of 82% of cases and 94% of deaths were averted in sub-Saharan Africa alone. In most cases, pregnant women and children under 5 years are the worst hit [18, 19]. According to the WHO report, over 12 million pregnancies were exposed to malaria infection during pregnancy; Central Africa, West Africa as well as East and Southern Africa recorded 40, 39 and 24% prevalence of exposure to malaria, respectively, in 2019 [15]. This high prevalence of exposure results in low birth weight in most cases. To save the future, there is a need to save children under 5 years and pregnant women from the menace of malaria.

Malaria prevention and control through investments in research

To prevent malaria and checkmate re-infection, several programs were designed in the past. One of such programs is the High Burden High Impact (HBHI) approach launched by the WHO in 2018 [15]. Though the launching and/or implementation was disrupted in some high burden countries due to the ravaging COVID-19 pandemic, these programs have been fruitful in some cases. For instance, from 2000 to 2019, the prevalence of *P. falciparum* malaria in Cambodia, Myanmar, Vietnam, Thailand and China was reduced by 97%, while countries previously certified malaria-free did not have any transmission or re-infection [17]. However, a global outlook showed that 20 more countries were added to the list of endemic countries within the period under review. Unfortunately, there are also disjointed data on sub-Saharan Africa's improvement, but reports suggest a total of 215 million cases in 2019 up from 204 million in 2000 [20]. However, malaria case incidence per 1000 population at risk reduced from 365 in 2000 to 225 in 2019 further reflecting the complexity in demographic data in such a rapidly growing population. Notwithstanding the effect of

the COVID-19 pandemic, the HBHI approach has kicked off in 10 of 11 malaria-endemic countries in sub-Saharan Africa. The impact, however, is yet to be felt region-wide as the number of cases in the 11 HBHI countries in 2019 (156 million) was similar to 2018 (155 million). Expectedly, the WHO Global Malaria Programme (GMP) foresees positive outcomes from this approach shortly following an aggressive commitment to adhere to the evidence-based recommendations developed by the WHO [15, 21].

More readily available options for malaria prevention and eradication are in the form of investments in malaria programs and research as contained in the Global Technical Strategy (GTS). The strategy is aimed at reducing mortality rate and malaria case incidence by 40, 75 and 90% in 2020, 2025 and 2030, respectively, which, at the time of launching in 2015, did not take into consideration the potential disruption due to the COVID-19 pandemic. Several players such as Global Fund as well as Melinda and Gates Foundation had invested immensely in malaria programs for research and development of malaria drugs, vaccines, diagnostic tools and vector control products. Some of the investments had yielded what today have become milestones in malaria treatments and prevention.

Current malaria control and treatments strategies

Chemoprevention and chemotherapy are the two major approaches known to reduce the burden of malaria in humans. Chemoprevention involves vector control (indoor residual spraying and insecticide-treated mosquito nets), which is recommended by the WHO to prevent malaria transmission. Indoor Residual Spraying (IRS) with insecticides is a powerful vector control approach, which involves spraying inside houses with insecticide once or twice a year. Sleeping under Insecticide-Treated Nets (ITN) reduces malaria cases by providing insecticidal effects and physical barriers to mosquitoes [22]. These chemopreventive measures are limited in application, coverage and effectiveness thus the reliance on the chemotherapeutic approach. Ever since the discovery and development of quinine from the Peruvian Amazon *Cinchona* species during the nineteenth century, several antimalarial drugs have come into existence for chemotherapeutic purposes [23]. The use of drugs for this purpose depends on prevalent *Plasmodium* species, demography, age, sex and the affected region. For example, travelers rely on chemoprophylaxis for the prevention of malaria. The WHO has recommended a minimum of three doses of intermittent sulfadoxine/pyrimethamine for pregnant women in endemic regions. For children under 5 years in the endemic region, during the season of high transmission, the administration of monthly courses of amodiaquine in addition to sulfadoxine/pyrimethamine is recommended [15]. Currently, ACTs have remained the first-line treatment

for malaria. These ACTs and other drugs are currently in use across different WHO regions, and their effectiveness has brought malaria prevention and treatment to where we currently are; a lot more still needs to be done to close the gap.

The previous and ongoing antimalarial discovery

Natural-products inspired approach

Plant-based products have shown promising potential as antimalarial agents and are the source of the two most important antimalarial drugs currently in use. Quinine, the first antimalarial agent, was characterized in 1820 by French Chemists. It was isolated from the bark of Peruvian Amazon *Cinchona calisaya* and *C. succirubra* (Rubiaceae) for the treatment of *P. falciparum* malaria [24]. Despite the continual use of quinine in chemotherapy, its effectiveness is hampered by the toxicity when used for a long period. Another plant-based compound still in use is artemisinin, a sesquiterpene endoperoxide from *Artemisia annua* of the Asteraceae family. Artemisinin, an unusual endoperoxide sesquiterpene lactone, was isolated by Chinese Scientists in 1972 and has been in use against chloroquine-resistant *P. falciparum* [24]. Though an alternative to quinine, some problems are associated with artemisinin such as recrudescence and high cost. The search for the ideal antimalarial drugs has continued, and several other compounds with antimalarial activity isolated from plants have been reviewed extensively [25–34].

Synthetic and semi-synthetic approach

Following the characterization of *Cinchona* alkaloid, quinine in 1820 for the treatment of complicated *P. falciparum* malaria, several other 4-aminoquinolines were synthesized based on the quinine ring nucleus. One of the 4-aminoquinolines was chloroquine, which is cheap and less toxic and has been a component of the global malaria eradication campaign. However, *P. falciparum* chloroquine-resistant strains were discovered in Latin America and Southeast Asia and have spread to most of the WHO endemic regions. Like quinine, artemisinin modification has led to the synthesis of several high potent analogues for further development [35]. In a study, some compounds primarily sulfonamides sourced from the Glaxo-Smithkline (GSK) selectively inhibited the in vitro growth of *P. falciparum* at the submicromolar level (IC$_{50}$, μM, 0.16–0.89). The inhibition, however, did not correlate with the known carbonic anhydrase enzyme inhibition by primary sulfonamides [36]. SAR was established for 1,2,3-triazole-naphthoquinone analogues synthesized by a Cu(I)-catalyzed Huisgen 1,3-dipolar cycloaddition reaction against chloroquine-sensitive *P. falciparum* F-32 Tanzania [37]. It was found that the nature of substituents on the

aromatic ring greatly influenced the antiprotozoal activity and further confirmed that the enzyme PfDHODH was the target of these compounds. Violacein, an indole pigment synthetically engineered from *E. coli,* was found to significantly affect the *P. falciparum* actin cytoskeleton [38]. Many traditional methods of antimalarial drug discovery such as optimization of existing therapy, analogue of existing therapy, drug resistance reversers and active compounds against new targets are known. These approaches have been replaced by modern methods such as target and ligand based.

The computer-aided drug design approach

Traditionally, the High Throughput Screening (HTS) method is used in drug discovery and it involves extensive experimental testing of a library of compounds against selected targets. It is a time-consuming and very expensive approach to drug discovery. The computational (virtual) approach has replaced HTS and involves in silico screening of large datasets for hit identification and subsequent design and optimization. This approach also enables the identification of compounds yet to be synthesized or commercially available [34, 39].

(A) Ligand-based approach

The ligand-based approach in drug discovery is designed to retrogressively analyze biological activity data, and different ligand-based approaches have been developed and validated to understand the nature of structural or chemical parameters involved in the antimalarial activity. Previous studies had applied Quantitative Structure–Activity/Property Relationship (QSAR/QSPR) in understanding the contribution of different structural features to the antimalarial activities and further predicted the activities of yet-to-be synthesized molecules [40–42]. Specifically, the applicability of the ligand-based approach has been tested on several synthetic prodiginines, 3-carboxyl-4(1H)-quinolone analogues, sidechain modified 4-amino-7-chloroquinolines, artemisinin derivatives, 7-substituted-4-aminoquinoline derivatives, 4-anilinoquinolines, quinine-based active agents as well as several natural products [34]. The flowchart of building a typical 3D-QSAR model is shown in Fig. 1. Typically, the low-energy conformers of a dataset for building a robust QSAR/QSPR are subjected to alignment based on the biophore hypothesis. This is followed by modeling, internal validation of the model and prediction of untested compounds. The approach (Fig. 1) also provides contouring from the model's coefficient of regression for the futuristic design of potential new bioactive molecules or modification of available molecules for a better activity. This approach

Fig. 1 Workflow for ligand-based drug discovery (LMD— low mode dynamic; AM1— Austin model 1 Hamiltonian). Four major steps: dataset pre-treatment, alignment, modeling and visualization are important here
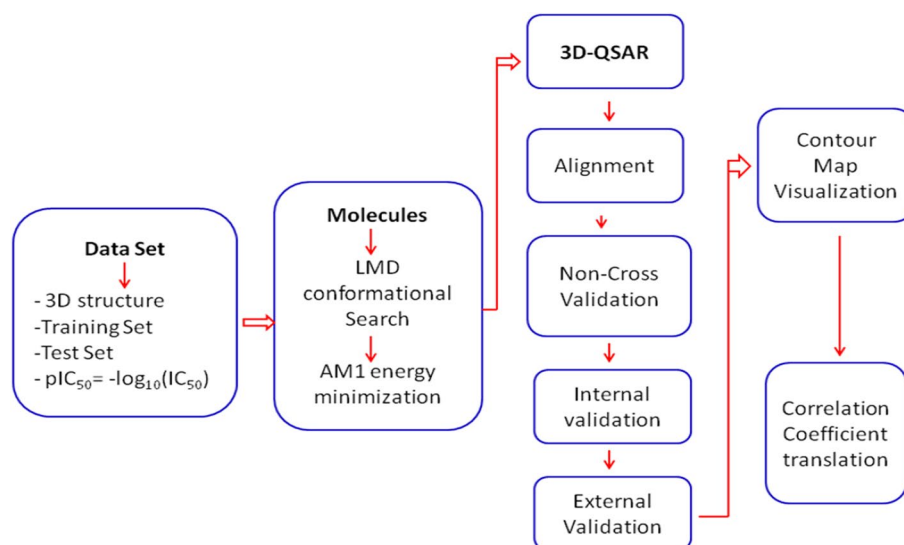
becomes more relevant where drug targets are not available or unknown.

(B) Structure-based approach

The structure-based approach involves a drastic reduction in the number of compounds in chemical space to a few hits having properties suitable for interacting with the target receptor. In this approach, a 3D structure of the target or a homologous protein must be known, several of which are available and freely accessible at the protein data bank (PDB, pdb.org). The workflow for the structure-based approach is shown in Fig. 2.

The first and critical step involved in the structure-based approach is the identification and validation of targets involved in the pathogenesis of malaria. Several targets have been identified in *Plasmodium* species for structure-based drug design as shown in Fig. 2 [43–48]. These targets were routinely used in the identification of single-target therapy where one antimalarial drug is used throughout malaria treatment. This form had led to the emergence of resistance. Recently, a multi-targeting hybrid approach that involves the modulation of several targets by one compound has been developed [49–53]. This involves artemisinin-based hybrid, quinoline-based hybrids, paclitaxel-based hybrids and target-based approaches via HTS in hybrid design [34].

Fig. 2 Workflow for a structure-based drug discovery approach (binding modes and mechanisms of molecules to specific amino acid residues of the receptor targets are obtained in this approach)
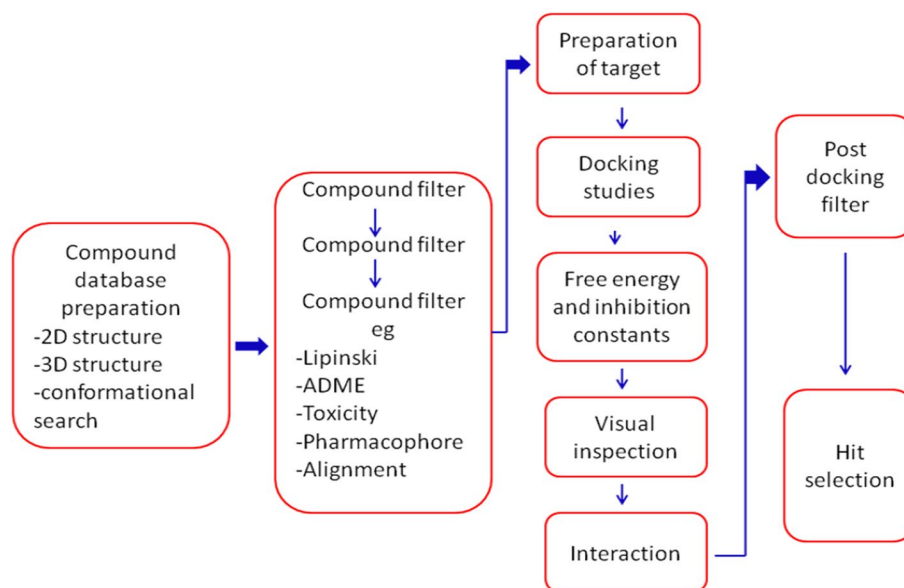
The gap in malaria prevention and treatment

Despite the plethora of approved antimalarial drugs, natural products-inspired and synthetic compounds so far identified as antiplasmodial, malaria still ranks closely with tuberculosis and HIV/AIDS. Various malaria prevention approaches have suffered some setbacks in recent years. Even though over 46% of Africans were protected from malaria by ITN in 2019, ITN coverage was stopped in 2016. More so, IRS protection has consistently declined from 5% in 2010 to 2% in 2019 across many WHO regions. The decline in protection was attributed to resistance developed by *Plasmodium* to pyrethroid IRS, which has forced countries to switch to more expensive insecticides.

Another gap widely documented is the issue of resistance to standard antimalarial drugs. The resistance of *Plasmodium* to chemotherapeutic agents was first observed in the 1950s and 1960s in chloroquine and sulfadoxine/pyrimethamine, thus reversing initial gains made in malaria control efforts [54, 55]. Similarly, partial resistance to artemisinin due to *PfKelch13* mutations has been reported and it is still under study. These gaps have further been widened by the emergence of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV2), causing COVID-19, which had spread to all malaria-endemic countries resulting in over 30 million cases and 1 million deaths as of March 2021 [15].

Worse still, no effective vaccine is available at the moment to fight malaria in humans. Many collaboration among global health funding bodies to develop vaccine has not yielded the desired result. Specifically, a vaccine against *P. falciparum*, RTSS/AS01E is developed and has shown 40% effectiveness in preventing malaria infection. While efforts to develop a vaccine against *Plasmodium* species are ongoing, a new strategy is needed to tackle the ever-changing landscape in malaria infection and treatment. One of such strategies is the use of ML tools.

The role of AI in malaria drug discovery

The need for the discovery of new malaria drugs cannot be overemphasized; this is because the *P. falciparum* parasites have successfully developed resistance against many drugs that are available [54, 55]. Malaria drug discovery involves the following stages: (1) target selection and validation; (2) compound screening and lead optimization; (3) pre-clinical studies and (4) clinical trials. Applying these steps in the traditional approach to drug discovery is very expensive and requires a lot of time; therefore, in recent times, drug discovery steps have focused on computational approaches [56]. The computational approaches to drug discovery use Artificial Intelligence techniques. In this section, we shall explore various Artificial Intelligence techniques in various steps of malaria drug discovery. Table 1 summarizes various AI techniques used in various stages of malaria drug discovery [56].

In another study, other researchers presented various AI and ML techniques used in various stages of drug discovery, together with the methods and level of accuracy [72]. This is summarized in Table 2.

The same study in [72] also summarized various problems at various stages of drug discovery that various novel approaches to AI have been used to solve, and presented the summary (Fig. 3).

Table 1 Different AI tools used in various stages of malaria drug discovery

S/N	AI tools name	Description	References
1	Chemputer	Structured format for procedure documentation for chemical synthesis	[57]
2	DeepChem	Python-based AI platform for the prediction of drug discovery tasks	[58]
3	DeepNeuralNet-QSAR	Use for prediction of molecular activity	[59]
4	DeepTox	Use for prediction of toxicity	[60]
5	DeltaVina	A scoring feature for protein–ligand binding affinity rescoring	[61]
6	Hit Dexter	ML-based molecule prediction models for molecules that could react to biochemical assays	[62]
7	Neural Graph Fingerprints	Predicts properties of novel molecules	[63]
8	NN Score	Used for protein–ligand interaction and neural network-based scoring mechanism	[64]
9	ODDT	Cheminformatic and molecular toolkit	[65]
10	ORGANIC	Method for molecular generation to build molecule with desired characteristics	[66]
11	PotentialNet	Ligand-binding prediction of affinity based on a convolution neural network (CNN)	[67]
12	PPB2	Polypharmacology prediction	[68]
13	REINVENT	Using RNN (recurrent neural network) and RL (reinforcement learning), molecular de novo architecture	[69]
14	SCScore	Scoring feature for the assessment of molecular synthesis complexity	[70]
15	SIEVE-Score	An improved method of virtual screening based on structure via interaction energy-based learning	[71]

Table 2 An overview of some studies that used AI for drug discovery [72]

Technique	Application	Method	Accuracy
Deep learning	Drug screening	Using interactions between proteins and their corresponding ligands as DeepTox input values to predict drug toxicity	Very accurate
Neural networks	Drug design	Using a deep learning neural network to predict the structures of different proteins	Very accurate
Reinforcement learning	Drug screening	Using a machine learning approach to finding an inhibitor molecule for a specific protein (DDR1)	Accurate
Neural networks	Drug design	Used artificial neural networks and deep learning to predict interactions between drugs and their targets	Very accurate
Neural networks	Drug design	Used neural networks which were integrated into a neural computer to design new small organic molecules	Highly accurate

Fig. 3 Outcome of different novel approaches to AI
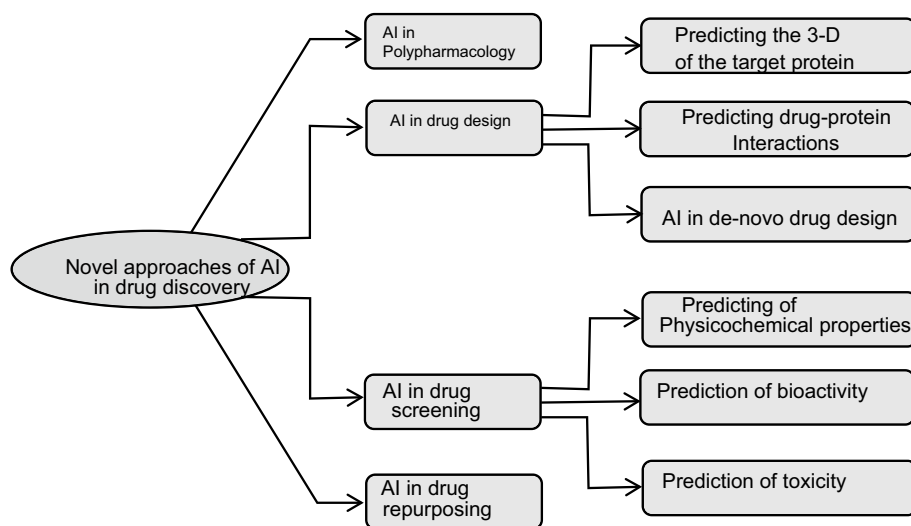
ML approach

ML is an aspect of Artificial Intelligence (AI), which helps you to acquire knowledge through experience. Experience in this context means data, while knowledge in this context means the ability to solve a problem, which can be a prediction of the continuous IC_{50} values of chemical compounds against *P. falciparum*, prediction of the bioactive class of chemical compounds against *P. falciparum,* etc. The experience here is the use of a chemical dataset, with various descriptors to predict the anti-*Plasmodium*. ML tasks can be classified as a prediction of continuous values (regression), prediction of classes (classification) and grouping of similar data items (clustering).

ML as a clustering tool

The clustering algorithm is one of the unsupervised ML algorithms, which identifies groups of similar data in a dataset [73]. In the case of the molecular compounds

dataset, clustering can be used to identify compounds that have similar chemical properties. Clustering dataset into groups of similar items can take any of the following:

(a) Exclusive clustering: The dataset belongs to only one group.
(b) Overlapping clustering: The dataset can belong to more than one group.
(c) Probabilistic clustering: The dataset belongs to any group with a known probability.
(d) Hierarchical clustering: The dataset is split into groups of similar data in a hierarchical manner. For example, the dataset can be split into two main groups, male and female. In each main group, it is refined into subgroups, like age groups, and each age group can be split into smaller subgroups, etc.

One classic ML clustering algorithm that is based on Euclidean distance is the K-means clustering algorithm. Figure 4 illustrates the exclusive clustering, while Table 3 shows probability clustering.
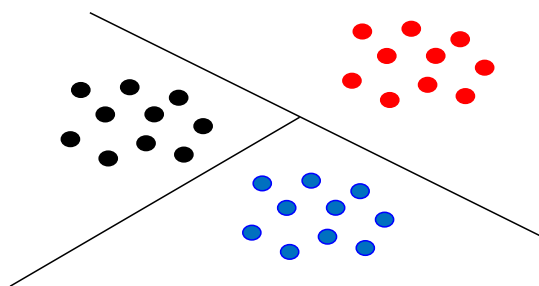
Fig. 4 Example of exclusive clustering

Table 3 Example of probabilistic clustering

Instances	Cluster 1	Cluster 2	Cluster 3
A	0.4	0.1	0.5
B	0.1	0.8	0.1
C	0.3	0.3	0.4
D	0.1	0.1	0.8
E	0.4	0.2	0.4
F	0.1	0.4	0.5

The importance of clustering compounds by structural or property similarity cannot be overemphasized. It provides a powerful approach to correlating compound features with bioactivity [7]. It can also be used for diversity analysis, for identifying compound redundancies and other biases in compound libraries [7]. Clustering has been used as an ML tool in analyzing molecular compound datasets for IC_{50} values against *P. falciparum*. [3], extracted the three most common targets from MacrolactoneDB, which are *P. falciparum* (malaria) [3], Hepatitis C and T-cells. Cheminformatics analysis was conducted on them and an ML workflow was developed. Unsupervised hierarchical clustering was conducted using Euclidean distance. The purpose or basis of the clustering in [3] was to be able to identify compounds that share similar chemical properties but different structural fragments, which resulted in different IC_{50}. Furthermore, clustering could help point to the importance or relevance of descriptors, based on whether they can cluster compounds with similar activities. The result of the clustering in [3] shows that the *P. falciparum* dataset has two clusters, which suggests two groups of compounds; each group shares similar chemical properties, but different structural fragments, which contributed to different IC_{50} values. The relationship between this clustering results in [3] and different structural fragments is the concept of Activity Cliff. This concept of Activity Cliff is very useful as a curative tool when preparing chemical datasets that have activity on P. falciparum. It can be used to separate chemical compounds with similar chemical properties but different IC_{50} values on P. Falciparum,

thereby leading to a high-quality chemical dataset. This was demonstrated in the study, using a clustering metric for similarity measurement (Tanimoto) of 0.87 and IC_{50} difference measure of 11.99 nM. This is called an Activity Cliff, because there is a great disparity between the IC_{50} of the compounds, despite having similar structures [74]. The same clustering metric (Tanimoto) was used to measure the similarity between the training data and test data, to apply a semi-supervised ML framework [75]. However, another study [76] used the similarity measure (Tanimoto coefficient) that is greater than different threshold values for different fingerprint similarity searching methods, to search for compounds whose IC_{50} against *P. falciparum* falls within specific values [76]. On the other hand, to avoid selection bias, clustering was used to establish even the assignment of chemical features into a training set and test set [77, 78]. This was done by dividing the molecules into clusters of ten molecules using hierarchical clustering. However, to select the most appropriate base model to be used to analyze a given chemical dataset, clustering was used to accomplish this [79]. To visualize the result of the clustering, the authors generated a chemical network of the compounds using Gephi. Each node of the chemical network was a micro-lactone ligand.

Activity Cliff

Activity Cliff is related to clustering compounds with similar structural properties. It has been defined as a pair of compounds with similar structural property, but with different potency (activity) against a known target [40]. Activity Cliff plays an important role in medicinal chemistry and chemo-informatics, because, in structure–activity relationship analysis and optimization, small chemical modification can be deduced from cliffs with high value in magnitude [40]. Furthermore, as part of the curative process, an activity cliff has been used to prepare a chemical dataset by removing pairs of compounds with high structural similarity but unexpectedly high activity difference [79]. This is to ensure that pairs of compounds with high structural similarity have similar activity on the target when using the dataset for QSAR. Based on the definition of activity cliff, four key components of activity cliff can be identified, which are: only a pair of compounds is considered, both compounds are active against the same known target, a structural similarity criterion must be specified, and potency difference criterion must be established [40]. Tanimoto value is the commonly used measure for measuring the structural similarity index between two compounds, while IC_{50} or Ki can be used for the potency measure of the two compounds.

Clustering algorithm for chemical compound datasets

Clustering has been used in the pharmaceutical industry to create different training datasets and test datasets as well [80], though the most commonly used clustering algorithm is Jarvis–Patrick's (J-P) clustering algorithm for clustering molecules of a chemical dataset. However, it has its associated problems, which include: it produces clusters that are either too large, in terms of the number of molecules in the clusters, but heterogeneous (small Tanimoto similarity value). It also produces clusters that are too small in terms of the number of molecules in the clusters but homogeneous (high Tanimoto similarity value) [80]. Based on these problems, other researchers [80] developed another clustering algorithm, which was able to create homogeneous clusters (high Tanimoto similarity value) and, at the same time, deal with either too small or too large molecules in each cluster. The clustering algorithm that he developed follows these three steps: (a) generation of daylight fingerprints (ASCII), (b) identification of potential cluster centroid and (c) mutual exclusion clustering.

The first step, Generation of Daylight Fingerprints, generates Fingerprints for each molecule in ASCII format in form of 0 and 1, using Daylight software, while the second step identifies the central molecule in each cluster (centroid). To determine the centroid, step 2 uses the specified Tanimoto similarity value to determine the number of neighbors of each molecule and arranges it in descending order, so that the molecule with the largest number of neighbors will be on top of the list, which will be the first centroid. Finally, step 3 uses different iteration to determine the members of various clusters. It does this by computing the pairwise Tanimoto similarity value of the centroid molecule and other molecules. If the pairwise Tanimoto similarity value is greater than or equal to the Tanimoto value that is used for the clustering, the molecule is taken as a member of the cluster and removed from the list. The next molecule in the list is taken as a centroid, the iteration continues. Any molecule that is still in the list at the end of the process is regarded as a singleton. The result of the clustering algorithm is illustrated in Fig. 5.

In Fig. 5, the members of cluster A have been collected together based on the pairwise Tanimoto value with centroid molecule colored red. Similarly, the members of cluster B have been collected together based on the pairwise Tanimoto value with the centroid molecule colored green. The molecule colored yellow is the singleton [73].

However, most clustering algorithms have been implemented as software tools. One such software tool is Chem-Mine Tools, which is an online portal with the capability for some cheminformatic functions, like search, visualization, clustering, etc. [7]. ChemMine Tools provides five major functionalities, which include: data visualization, structure comparisons, similarity searching, compound clustering and prediction of chemical properties [7]. The similarity toolbox of ChemMine implements an algorithm that uses atom pairs as a structural descriptor and the widely used Tanimoto coefficient as a similarity measure to compute similarity measures among compounds. Another feature of ChemMine is that it allows the use of other similarity coefficients like Tversky or Dice [7]. Furthermore, the clustering toolkit of ChemMine implements three clustering algorithms, which are: hierarchical clustering, Multi-Dimensional Scaling (MDS) and binning clustering [7]. Clustering by structural similarity requires that the similarity measure be computed by first generating the atom pair descriptors (features) for each compound, which is used to calculate the similarity matrix using the Tanimoto coefficient. While hierarchical clustering organizes the compounds by similarity using a tree structure, the MDS outputs the similarity information in a scatter plot. Though both methods do not assign the compounds to discrete similar groups, the assignment to a similar group is done later in the clustering process, using various post-processing approaches, like the tree cutting method [7]. On the other hand, the binning method clustering provides the clustering groups using a user-defined similarity measure cut-off. The method allows the user to choose a similarity cut-off; afterwards, compounds that have a similarity measure that is greater than or equal to the chosen similarity value will be assigned into groups [7].

In addition to ChemMine as a software tool for analyzing chemical compounds dataset, there are other software tools with additional clustering functionalities; one of such is ChemmineR [81].

ML as a classification tool

A review of relevant literature showed some studies that applied ML approaches to predict activity against *P. falciparum*. In this section, relevant classification models were reviewed; six reviews identified SVM as the best classification tool, four report identified Random Forest as the best
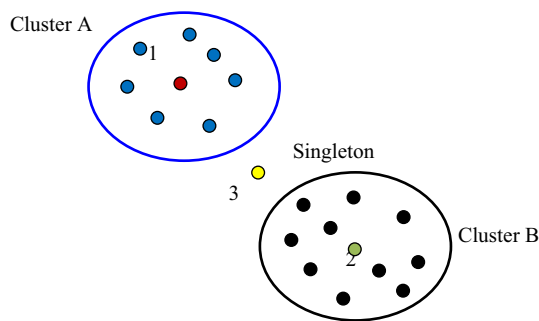
Fig. 5 Result of the clustering algorithm
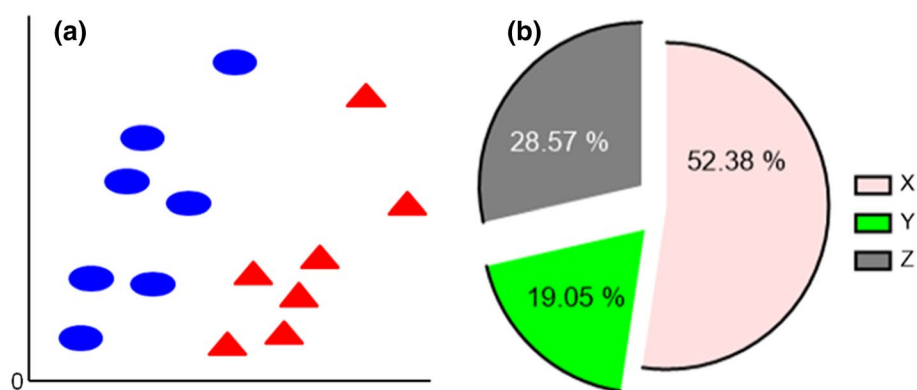
Fig. 6 Classification of objects into two different categories (**A**) and distribution of the best performing ML algorithms based on the relevant articles reviewed (**B**). In (a), datasets are classified based on similarity (blue or red color, circular or triangular shape) or dissimilarity (blue and red colors or circular and triangular shapes); Deep Learning (DL), Boosted Trees Regression (BTR), J48 classifier (JC),

Discriminant Functions (DF), XGBoost (XGB), Graph Convolutional Neural Networks (GCNN), Multilinear Regression (MLR), General Regression Neural Network (GRNN), Multivariate Analysis (MVA), C5.0 and Artificial Neural Networks (ANN) represent (X), Random Forest (RF) represents (Y), and Support Vector Machine (SVM) represents (Z)

modeling tools while the other eleven modeling tools were also identified as shown in Fig. 6B

RF algorithm

A study on varied drug-decorated nanoparticles organic compound/drug complexes used eight ML classifiers to predict activity against *P. falciparum* [8]. The dataset was based on 107 input features and 249,992 compounds, and the best model was RF (27 selected features) with a mean area under the Receiver Operating Characteristic curve (ROC) a value of 0.9921 _ 0.000244 (tenfold cross-validation) which is statistically significant. Janairo et al.introduced a 20 chemical descriptors predictive model (ML) employed to establish a relationship between the mosquito repellent activity of 33 natural compounds using four classifiers. The optimized model through BTR (best performed) demonstrated a good predictive ability (r^2 $train = 0.93$, r^2 $test = 0.66$, r^2 $overall = 0.87$) than other ML applied [82].

The RF algorithm showed a lower overall accuracy of 0.75 in a QSAR study involving 323,201 compounds to identify the biological activity of new antimalarial against the apicoplast in *P. falciparum* with 179 descriptors [83]. The regression analysis showed an AUC of 70%, specificity of 80% and a sensitivity of 40–50%. Egieyeh et al. [84] applied four ML algorithms (Optimization of SVMs, Naïve Bayesian, Voted Perceptron, Sequence Minimization and RF) on QSAR of 1155 natural products with an in vitro antiplasmodial activity using 76 descriptors. With an accuracy of 82.8% and an AUC of 0.91, this study appeared to be better predictive than the previous study [83] and could be attributed to the outrageous number of descriptors or poor correlation used in the former [83]. A study developed and evaluated a 97 QSAR model of 16 datasets to generate a

predicted profile in bioactivity and cytotoxicity using different approaches (e.g., conformal prediction framework) to improve the prediction accuracy of models [85]. The result was evaluated by modeling the dataset with and without the addition of the predicted continuous bioactivity profiles; the efficacies of the final models improved with the addition of the predicted continuous bioactivity profiles.

SVM algorithm

The SVM is another ML algorithm used in support vector classification to find a hyperplane in both classification and regression analyses [86–88]. This algorithm has been applied in regression analysis for the prediction of biological activity against *P. falciparum*. In a study, both linear and nonlinear SVM algorithms were built to classify 999 compounds (inhibitors and non-inhibitors) for anti-proliferative activity against *P. falciparum* using 383 descriptors [12]. The statistical validation showed performance with an accuracy of 83% and an AUC of 0.88. The predictive power of the optimized model shows that it may be effective in selecting potential hits in screening large libraries. A dataset of ~ 4750 compounds with activity against *P. falciparum* was subjected to four ML algorithms (SVM, RF, kNN and XGB) with 98 descriptors [89]. Both SVM and XGB performed better with ~ 85% on the independent test set. This finding further supported the work of [12] that the built models are efficient and may be potentially useful for facilitating the discovery of antimalarial agents [12]. With a slightly higher SVM prediction accuracy (R^2 training 8.95 and R^2 test 8.73), a study discovered a good 2D-QSAR model in a study involving 4750 compounds to identify antimalarial activities against *P. falciparum* using 15 descriptors [90]. The study also showed that GRNN prediction accuracies of 99.7% for

the training set (3887 compounds) and 88.9% for the test set (863 compounds). Similarly, a study evaluated 116,987 antimalarial compounds against apicoplast formation using 173 descriptors [91]. The R-caret package employed different algorithms for the predictive model building including Generalized Linear Model (GLM), kNN), SVM, RF and C5.0 decision tree. The model validation showed that C5.0 and SVM and RBF outperform others. The modeling of 277 *P. falciparum* proliferation inhibitors and non-inhibitors with SVM using various descriptors showed 87% overall accuracy and an AUC of 0.73 [92].

QSAR ML algorithms

A structural descriptors-based QSAR model of anti-*Plasmodium* liver stages bioactivity and prediction of physicochemical parameters influencing intestinal absorption for 127 compounds have been reported [93]. Seventeen drugs that were predicted to be active or inactive were selected for testing against the hepatic stage of *P. yoelii* in vitro. Antiretroviral, antifungal and cardiotonic drugs were found to be highly active (nanomolar 50% inhibitory concentration values), and two ionophores completely inhibited parasite development. The most active compounds against the hepatic stages of *P. yoelii yoelii* and *P. falciparum* were monensin and nigericin, with IC$_{50}$ of 10.3 nM, and the analysis was used to categorize the compounds into highly active, active and inactive groups according to their 50% inhibitory concentrations (IC$_{50}$). A more comprehensive MLR 2D-QSAR model to predict anti-*P. falciparum* activity of two datasets of organic compounds, each with an R^2 of 0.84 and 0.89, has been demonstrated [94]. In addition to MLR, Santos et al. [95] had used 230 descriptors, PLS and PCR analysis to describe the QSAR of artemisinin and 20 derivatives and further predicted the antimalarial activity of 30 new artemisinin compounds unknown activity showing high statistical significance. A higher dataset of 72 compounds with lower descriptors (39) was applied to build a QSAR model against the 3D7 *P. falciparum* strain, which identified 31 potential antimalarial compounds [6]. Interestingly, another study demonstrated a 2D-QSAR model of 3133 compounds using 929 descriptors in which the study showed abysmal 14.2% accuracy [96]. A similar study applied Artificial Neural Networks with Levenberg–Marquardt algorithm (nonlinear approach) on the anti-malarial activity of a set of 33 imidazolopiperazine compounds against 3D7 and W2 strains [97]. Results showed the potential of the suggested model for the prediction of 3D7 activity and more acceptable than W2 strain with $R^2_{train} = 0.947$, $R^2_{val} = 0.959$, $R^2_{test} = 0.920$. The results of R^2, MSE and leverage value showed that the prediction ability of the ANN method for estimation of the antimalarial activity in imidazolopiperazine compounds is good and can be used as a virtual tool molecule to design more

efficient compounds with activity against malaria (3D7 and W2 strains). An integrated application of ML algorithms, CoMFA analyses and molecular docking methods on a set of 228 known triclosan and rhodanine inhibitors of *P. falciparum* enoyl acyl carrier protein reductase (PfENR) of potential antimalarial agents targeted to PfENR yielded accuracies for the training set and evaluation set are 94.18 and 57.14% for IB1 and 92.80 and 68.57% for Kstar, respectively [77]. Neves et al. [78] adopted deep learning to build binary and continuous 2D RDKit descriptors QSAR models based on large datasets for predicting the antiplasmodial activity and cytotoxicity of 413,855 untested compounds. The developed computational models were used to prioritize novel, active, and nontoxic compounds from virtual chemical libraries for experimental evaluation. Similarly, a researcher had developed an ML-based QSAR model to predict which molecules will block the malaria parasite's ion pump, PfATP4 [98]. The model was then employed to screen and classify the DrugBank database molecules and compounds coming from a proprietary marine molecules library.

Other ML algorithms

A deep learning-based algorithm (DeepMalaria) for anti-*P. falciparum* activities features of 13,446 compounds and 23 descriptors was demonstrated using their SMILES [99]. The algorithm predicted 72.3% of active compounds from the validation dataset and 87.8% of that of the test dataset with acceptable accuracy in an imbalanced setting showing significant predictive potentials to improve drug design and development. A study] reported a systematic review on the green synthesis of metal nanoparticles as a potential source of new antiplasmodial drugs [100]. Seven electronic databases and 17 papers were included in the review. A very high proportion of the studies (82.4%) used plant leaves to produce nanoparticles (NPs) while three studies used microorganisms, including bacteria and fungi.

ML as a regression tool

There have been many reports on the use of regression analysis as an ML tool using different ML algorithms such as deep learning, Random Forest (RF), Boosted Trees Regression (BT), J48 classifier, DF2, SVM, XG boost, GCNN, Multilinear Regression (MR), GRNN, C5.0, ANN among others. In this regard, the regression tool is a predictive computational model that enables one to understand the correlation between chemical properties (descriptors) and their activities, i.e., to computationally screen large molecular datasets thereby offering a possibility to improve the hit rate and thereby reduce the overall costs of drug discovery. This has been applied extensively in the drug discovery of anti-*P. falciparum* drug. Careful analysis of the result reported

by different authors shows that efficient computational predictive models help to screen large datasets in silico and could be potentially used to prioritize molecules for high-throughput screens.

In a multi-parametric QSAR study to predict IC_{50} and Log P for 5-N-acetyl-β-D-neuraminic acid, consisting of 110 training sets and 50 test sets of compounds structurally related to 5-N-acetyl-β-D-neuraminic acid, polyAnalyst was used to develop the linear model using a stepwise linear regression algorithm [101]. The predicted IC_{50} values provide good statistical measures for the correlation coefficient, standard deviation and standard error as 0.8545, 0.2932 and 0.3815, respectively. Importantly, the model showed that a strong correlation exists between Log P and IC_{50} of drug compounds. A dataset of 34 compounds and 8 descriptors was subjected to MLR analysis to construct a QSAR. This method produced higher R^2 (0.9714–0.9909) and RMSEP of 0.0938 and 0.1819 compared with the method of Pushpa and co-workers [101]. In drug discovery of the transmission-blocking potential of 44 anti-malarial compounds in the mosquito feeding assay using *P. falciparum* male gamete inhibition assay, [102] applied regression tool. Root Mean Square Error (RMSE) of 22.51% was obtained from the measured relationship between exflagellation inhibition (EI) and oocyst reduction [102]. The model provided pIC_{50} predictions in SMFA with high accuracy, and IC_{50} values for 11 compounds obtained in the exflagellation inhibition assay were correlated with IC_{50} values in SMFA. Significantly, the result of the regression models gave IC_{50} predictions results in SMFA that had high accuracy. However, it was stated that the small dataset ($n = 44$) used to build the model may render the result unreliable.

ML has also been successfully applied in epidemiological studies of malaria [103, 104]. The outbreak of malaria using six observed variables; a dataset of thirty-eight compounds collected from malaria samples of Maharashtra State with eight descriptors was used [105]. To determine the performance of the model, logistic regression, random decision trees and Gaussian processes were used. The regression model as well as the Decision tree and Gaussian models was able to give 100% accuracy in predicting malaria outbreaks [106]. A combination of 8 ML algorithms (KNN, SVM, SVM linear, linear regression, linear discriminant analysis, DT and RF classifiers) to predict the effect of compound/drug reactions that have antimalarial activity against *Plasmodium* has been documented [8]. Findings showed that Random Forest classifiers gave a more accurate result than other learning algorithms.

Importantly, the top six ML algorithms—simple linear regression model, lasso, logistic regression, Support Vector Machines, multivariate regression algorithm and multiple regression algorithm—are commonly used in data mining and their applications in industry are well known.

Expert opinion and prospects

The available epidemiological data show that malaria, no doubt, is a disease of today and the future despite huge investment toward vaccine development and drug discovery [107]. Available drugs have been overwhelmed by *Plasmodium* resistance and poor pharmacokinetic-related limitations; this calls for an urgent need to explore more approaches. Natural products-derived compounds, synthetic and several modification attempts on existing drugs have not yielded the desired products. Despite huge deposits of potential antimalarial compounds in various databases, none has been transformed from such virtual spaces to the bedside; perhaps the goldmine strategies are yet to be exploited. The target of pharmaceutical and drug discovery scientists has always been to discover and develop new drugs that will ultimately benefit the patient within the shortest possible time and at an affordable cost. ML is a developing trend in the drug discovery industry. It is expected to revolutionize the drug discovery process by introducing efficiency that will lead to the discovery of new drugs at a shorter time and at a lower cost.

ML in drug discovery has come to stay and its application in the discovery of anti-Plasmodium species drugs is emerging. The quagmire now is whether a game-changing ML approach has been explored, exploited or adopted. That is the crux of this systematic review. Several known ML algorithms have been applied in anti-*Plasmodium* species drug discovery which resulted in acceptable statistical significance measures. With ML, the biodiversity which has been under threat because of several drug discovery programs can be conserved or handled with a more precise approach. There is a need for scaling down the ML technology to early-career DDD scientists so that soon, the tools used by ML specialists will become a norm in laboratories involved in drug discovery and development. ML and its tools could also find use in downstream processing of pharmaceuticals where current good manufacturing practices are expected to be religiously followed to ensure the production of consistently high-quality medicines that will meet regulators' specifications.

Several ML tools were reviewed in this paper. Careful analysis of the literature reviewed in this paper indicated that Support Vector Machine (SVM) was the most highly favored tool followed by Random Forest. SVM has been widely applied in biological and other sciences with high accuracy. However, other machine learning tools identified in this study have been sparingly used and could serve as a good starting point for the discovery of game-changing antimalarial drugs. It is thus expected that the application of this ML tool or its modification in the discovery of antimalarial and other drugs will progress rapidly in

the coming years considering the urgency needed for the discovery of new anti-infectives required to meet the healthcare needs of countries in the endemic regions of the world.

Conclusion

Malaria can be eradicated in sub-Saharan Africa by the combination of chemotherapy and chemoprevention. The emergence of resistance has continued to hamper chemotherapeutic approaches. However, emerging drug discovery methods such as ML have continued to show potential for new molecules capable of circumventing many known challenges. Until total eradication is achieved, the search for vaccines and cures will continue to receive attention. Our team is currently working on the application of various ML algorithms to the discovery of potent, safe, affordable and deliverable molecules against *P. falciparum*.

Declarations

Competing interest The authors declare that they have no competing interests.

References

1. Cheoymang A, Na-Bangchang K (2018) A systematic review: application of in silico models for antimalarial drug discovery. Afr J Pharm Pharmacol 12(13):159–167. https://doi.org/10.5897/AJPP2018.4904

2. Ojha PK, Kumar V, Roy J, Roy K (2021) Recent advances in quantitative structure–activity relationship models of antimalarial drugs. Expert Opin Drug Discov 16(6):659–685. https://doi.org/10.1080/17460441.2021.1866535

3. Zin PPK, Williams GJ, Ekins S (2020) Cheminformatics analysis and modeling with MacrolactoneDB. Sci Rep 10(6):8284

4. Sellwood MA, Ahmed M, Segler MH, Brown N (2018) Artificial intelligence in drug discovery. Fut Sci. https://doi.org/10.4155/fmc-2018-0212

5. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M, Zhao S (2019) Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 18(6):463–477. https://doi.org/10.1038/s41573-019-0024-5

6. Viira B, Gendron T, Lanfranchi DA, Cojean S, Horvath D, Marcou G, Varnek A, Maes L, Maran U, Loiseau PM, Davioud-Charvet E (2016) In silico mining for antimalarial structure–activity knowledge and discovery of novel antimalarial curcuminoids. Molecules 21(7):853. https://doi.org/10.3390/molecules21070853

7. Backman TW, Cao Y, Girke T (2011) ChemMine tools: an online service for analyzing and clustering small molecules. Nucleic Acids Res 39(2):W486–W491. https://doi.org/10.1093/nar/gkr320

8. Urista DV, Carrué DB, Otero I, Arrasate S, Quevedo-Tumailli VF, Gestal M, González-Díaz H, Munteanu CR (2020) Prediction of antimalarial drug-decorated nanoparticle delivery systems with random forest models. Biol 9(8):198. https://doi.org/10.3390/biology9080198

9. Neves BJ, Moreira-Filho JT, Silva AC, Borba JV, Mottin M, Alves VM, Braga RC, Muratov EN, Andrade CH (2021) Automated framework for developing predictive machine learning models for data-driven drug discovery. J Braz Chem Soc 32(1):110–122. https://doi.org/10.26434/chemrxiv.12250046.v1

10. Bagavan A, Rahuman AA, Kamaraj C, Kaushik NK, Mohanakrishnan D, Sahal D (2011) Antiplasmodial activity of botanical extracts against *Plasmodium falciparum*. Parasitol Res 108(5):1099–1109. https://doi.org/10.1007/s00436-010-2151-0

11. Ford CT, Janies D (2020) Ensemble machine learning modeling for the prediction of artemisinin resistance in malaria. F1000Res 9(62):62. https://doi.org/10.12688/f1000research.21539.5

12. Subramaniam S, Mehrotra M, Gupta D (2011) Support vector machine-based classification model for screening Plasmodium falciparum proliferation inhibitors and non-inhibitors. Biomed Engi Comput Biol 3:13–24 BECB-S7503. https://doi.org/10.4137/BECB.S7503

13. Upadhyay C, Chaudhary M, De Oliveira RN, Borbas A, Kempaiah P, Rathi B (2020) Fluorinated scaffolds for antimalarial drug discovery. Expert Opin Drug Discov 15(6):705–718. https://doi.org/10.1080/17460441.2020.1740203

14. Andrews KA, Wesche D, McCarthy J, Mohrle JJ, Tarning J, Phillips L, Kern S, Grasela T (2018) Model-informed drug development for malaria therapeutics. Annu Rev Pharmacol Toxicol 58:567–582. https://doi.org/10.1146/annurev-pharmtox-010715-103429

15. World Health Organization. World malaria report 2020: 20 years of global progress and challenges. 2020

16. Diagana TT (2015) Supporting malaria elimination with 21st Century antimalarial agent drug discovery. Drug Discov Today 20(10):1265–1270. https://doi.org/10.1016/j.drudis.2015.06.009

17. Badger-Emeka LI (2020) The malaria burden: a look at 3 years outpatient malaria clinic visits in a university community town in Southeast of Nigeria. Nig J Clin Pract 23(5):711–719. https://doi.org/10.4103/njcp.njcp_218_19

18. Snow RW, Sartorius B, Kyalo D, Maina J, Amratia P, Mundia CW, Bejon P, Noor AM (2017) The prevalence of *Plasmodium falciparum* in sub-Saharan Africa since 1900. Nature 550(7677):515–518. https://doi.org/10.1038/nature24059

19. Winstanley PA (2000) Chemotherapy for *falciparum* malaria: the armoury, the problems and the prospects. Parasitol Today 16(4):146–153. https://doi.org/10.1016/S0169-4758(99)01622-1

20. Alegana VA, Okiro EA, Snow RW (2020) Routine data for malaria morbidity estimation in Africa: challenges and prospects. BMC Med 18:1–13. https://doi.org/10.1186/s12916-020-01593-y

21. Roll Back Malaria Partnership Secretariat. Action and investment to defeat malaria 2016–2030. For a malaria-free world. Geneva: World Health Organization; 2015 https://endmalaria.org/sites/default/files/RBM_AIM_Report_0.pdf

22. Lengeler C (2004) Insecticide-treated bed nets and curtains for preventing malaria. Cochrane Database Syst Rev (2)

23. Dondorp AM, Yeung S, White L, Nguon C, Day NP, Socheat D, Von Seidlein L (2010) Artemisinin resistance: current status and scenarios for containment. Nat Rev Microbiol 8(4):272–280. https://doi.org/10.1038/nrmicro2331

24. Pan WH, Xu XY, Shi N, Tsang SW, Zhang HJ (2018) Antimalarial activity of plant metabolites. Int J Mol Sci 19:1382. https://doi.org/10.3390/ijms19051382

25. Kaya GI, Sarıkaya B, Onur MA, Somer NU, Viladomat F, Codina C, Bastida J, Lauinger IL, Kaiser M, Tasdemir D (2011) Antiprotozoal alkaloids from *Galanthus trojanus*. Phytochem Lett 4(3):301–305. https://doi.org/10.1016/j.phytol.2011.05.008

26. Bringmann G, Messer K, Schwöbel B, Brun R, Assi LA (2003) Habropetaline A, an antimalarial naphthylisoquinoline alkaloid from *Triphyophyllum peltatum*. Phytochem 62(3):345–349. https://doi.org/10.1016/S0031-9422(02)00547-2

27. Graziose R, Rathinasabapathy T, Lategan C, Poulev A, Smith PJ, Grace M (2011) Antiplasmodial activity of aporphine alkaloids and sesquiterpene lactones from *Liriodendron tulipifera* L. J Ethnopharmacol 133(1):26–30. https://doi.org/10.1016/j.jep.2010.08.059

28. Fernandez LS, Sykes ML, Andrews KT, Avery VM (2010) Antiparasitic activity of alkaloids from plant species of Papua New Guinea and Australia. Int J Antimicrob Agents 36(3):275–279. https://doi.org/10.1016/j.ijantimicag.2010.05.008

29. Toriizuka Y, Kinoshita E, Kogure N, Kitajima M, Ishiyama A, Otoguro K, Yamada H, Ōmura S, Takayama H (2008) New lycorine-type alkaloid from *Lycoris traubii* and evaluation of antitrypanosomal and antimalarial activities of lycorine derivatives. Bioorg Med Chem 16(24):10182–10189. https://doi.org/10.1016/j.bmc.2008.10.061

30. Osorio EJ, Berkov S, Brun R, Codina C, Viladomat F, Cabezas F, Bastida J (2010) In vitro antiprotozoal activity of alkaloids from *Phaedranassa dubia* (Amaryllidaceae). Phytochem Lett 3(3):161–163. https://doi.org/10.1016/j.phytol.2010.06.004

31. Fournet A, Barrios AA, Muñoz V, Hocquemiller R, Roblot F, Cavé A, Richomme P, Bruneton J (1994) Antiprotozoal activity of quinoline alkaloids isolated from *Galipea longiflora*, a Bolivian plant used as a treatment for cutaneous leishmaniasis. Phytother Res 8(3):174–178. https://doi.org/10.1002/ptr.2650080312

32. Li J, Seupel R, Feineis D, Mudogo V, Kaiser M, Brun R, Brünnert D, Chatterjee M, Seo EJ, Efferth T, Bringmann G (2017) Dioncophyllines C2, D2, and F and related naphthylisoquinoline alkaloids from the Congolese liana *Ancistrocladus ileboensis* with potent activities against *Plasmodium falciparum* and against multiple myeloma and leukemia cell lines. J Nat Prod 80(2):443–458. https://doi.org/10.1021/acs.jnatprod.6b00967

33. Wright CW (2005) Plant-derived antimalarial agents: new leads and challenges. Phytochem Rev 4(1):55–61. https://doi.org/10.1007/s11101-005-3261-7

34. Tibon NS, Ng CH, Cheong SL (2020) Current progress in antimalarial pharmacotherapy and multi-target drug discovery. Eur J Med Chem 188:111983. https://doi.org/10.1016/j.ejmech.2019.111983

35. Meshnick SR (2001) Artemisinin and its derivatives. In: Rosenthal PJ (ed) Antimalarial chemotherapy: mechanisms of action, resistance, and new directions in drug discovery. Humana Press Totowa, NJ, pp 191–201

36. Fisher GM, Bua S, Del Prete S, Arnold MS, Capasso C, Supuran CT, Andrews KT, Poulsen SA (2017) Investigating the antiplasmodial activity of primary sulfonamide compounds identified in open source malaria data. Int J Parasitol Drugs Drug Resistance 7(1):61–70. https://doi.org/10.1016/j.ijpddr.2017.01.003

37. Oramas-Royo S, López-Rojas P, Amesty Á, Gutiérrez D, Flores N, Martín-Rodríguez P, Fernández-Pérez L, Estévez-Braun A (2019) Synthesis and antiplasmodial activity of 1, 2, 3-triazolenaphthoquinone conjugates. Molecules 24(2):3917. https://doi.org/10.3390/molecules24213917

38. Wilkinson MD, Lai HE, Freemont PS, Baum J (2020) A biosynthetic platform for antimalarial drug discovery. Antimicrob

39. Ramsay RR, Popovic-Nikolic MR, Nikolic K, Uliassi E, Bolognesi ML (2018) A perspective on multi-target drug discovery and design for complex diseases. Clin Transl Med 7(1):1–14. https://doi.org/10.1186/s40169-017-0181-2

40. Hu Y, Stumpfe D, Bajorath J (2013) Advancing the activity cliff concept. F1000Res 2:199. https://doi.org/10.12688/f1000research.2-199.v1

41. Dimova D, Stumpfe D, Bajorath J (2015) Systematic assessment of coordinated activity cliffs formed by kinase inhibitors and detailed characterization of activity cliff clusters and associated SAR information. Eur J Med Chem 90:414–427. https://doi.org/10.1016/j.ejmech.2014.11.058

42. Ojeda-Montes MJ, Gimeno A, Tomas-Hernández S, Cereto-Massagué A, Beltrán-Debón R, Valls C, Mulero M, Pujadas G, Garcia-Vallvé S (2018) Activity and selectivity cliffs for DPP-IV inhibitors: Lessons we can learn from SAR studies and their application to virtual screening. Med Res Rev 38(6):1874–1915. https://doi.org/10.1002/med.21499

43. Coronado L, Nadovich C, Spadafora C (2014) Malarial hemozoin: from target to tool. Biochim Biophys Acta 1840(6):e2032–e2041. https://doi.org/10.1016/j.bbagen.2014.02.009

44. Chen GQ, Benthani FA, Wu J, Liang D, Bian ZX, Jiang X (2020) Artemisinin compounds sensitize cancer cells to ferroptosis by regulating iron homeostasis. Cell Death Differ 27(1):242–254. https://doi.org/10.1038/s41418-019-0352-3

45. Moras M, Lefevre SD, Ostuni M (2017) From erythroblasts to mature red blood cells: organelle clearance in mammals. Front Physiol 8:1076. https://doi.org/10.3389/fphys.2017.01076

46. Kumar S, Bhardwaj TR, Prasad DN, Singh RK (2018) Drug targets for resistant malaria: historic to future perspectives. Biomed Pharmacother 104:8–27. https://doi.org/10.1016/j.biopha.2018.05.009

47. Hikosaka K, Komatsuya K, Suzuki K, Kita K (2015) Mitochondria of malaria parasites as a drug target. In: Samie A (Ed.), an overview of tropical diseases, IntechOpen, pp 17e37. https://doi.org/10.5772/61283

48. Fidock DA, Eastman RT, Ward SA, Meshnick SR (2008) Recent highlights in antimalarial drug resistance and chemotherapy research. Trends Parasitol 24(12):537–544. https://doi.org/10.1016/j.pt.2008.09.005

49. Morphy R, Rankovic Z (2005) Designed multiple ligands. An emerging drug discovery paradigm. J Med Chem 48(21):6523–6543. https://doi.org/10.1021/jm058225d

50. Walsh JJ, Coughlan D, Heneghan N, Gaynor C, Bell A (2007) A novel artemisinin–quinine hybrid with potent antimalarial activity. Bioorg Med Chem Lett 17(13):3599–3602. https://doi.org/10.1016/j.bmcl.2007.04.054

51. Agarwal D, Gupta D, Awasthi SK (2017) Are antimalarial hybrid molecules a close reality or a distant dream? Antimicrob Agents Chemother. https://doi.org/10.1128/AAC.00249-17

52. Schellenberg D, Abdulla S, Roper C (2006) Current issues for anti-malarial drugs to control *P. falciparum* malaria. Curr Mol Med 6(2):253–260. https://doi.org/10.2174/156652406776055168

53. Srivastava V, Lee H (2015) Chloroquine-based hybrid molecules as promising novel chemotherapeutic agents. Eur J Pharmacol 762:472–486. https://doi.org/10.1016/j.ejphar.2015.04.048

54. Alonso P, Noor AM (2017) The global fight against malaria is at crossroads. The Lancet 390(10112):2532–2534. https://doi.org/10.1016/S0140-6736(17)33080-5

55. Trape JF (2001) The public health impact of chloroquine resistance in Africa. Am J Trop Med Hygiene 64(1_suppl): 12–17

56. Ghosh B Choudhuri S (2021) Drug design for malaria with artificial intelligence (AI). In: Tyagi RK (Ed) Plasmodium specie and

Agents Chemother 64(5):e02129-e2219. https://doi.org/10.1128/AAC.02129-19

drug resistance. IntechOpen https://doi.org/10.5772/intechopen.98695

57. Steiner S, Wolf J, Glatzel S, Andreou A, Granda JM, Keenan G, Hinkley T, Aragon-Camarasa G, Kitson PJ, Angelone D, Cronin L (2019) Organic synthesis in a modular robotic system driven by a chemical programming language. Science 363:eaav2211. https://doi.org/10.1126/science.aav2211

58. Lavecchia A (2019) Deep learning in drug discovery: opportunities, challenges and future prospects. Drug Discov Today 24(10):2017–2032. https://doi.org/10.1016/j.drudis.2019.07.006

59. Xu Y, Ma J, Liaw A, Sheridan RP, Svetnik V (2017) Demystifying multitasks deep neural networks for quantitative structure–activity relationships. J Chem Inf Model 57(10):2490–2504. https://doi.org/10.1021/acs.jcim.7b00087

60. Mayr A, Klambauer G, Unterthiner T, Hochreiyer S. (2016) DeepTox: toxicity prediction using deep learning. Front Environ Sci 3:80. https://doi.org/10.3389/fenvs.2015.00080

61. Wang C, Zhang Y (2017) Improving scoring-docking-screening powers of protein-ligand scoring functions using random forest. J Comput Chem 38:169–177. https://doi.org/10.1002/jcc.24667

62. Stork C, Chen Y, Sicho M, Kirchmair J (2019) Hit Dexter 2.0: machine-learning models for the prediction of frequent hitters. J Chem Inf Model 59:1030–1043. https://doi.org/10.1021/acs.jcim.8b006777

63. Duvenaud DK, Maclaurin D, Aguilera-Iparraguirre J, Gómez-Bombarelli R, Hirzel T, Aspuru-Guzik A, Adams RP (2015) Convolutional networks on graphs for learning molecular fingerprints. Preprint. arXIv: 1509.09292v2

64. Durrant JD, McCammon JA (2011) NNScore 2.0: a neural-network receptor-ligand scoring function. J Chem Inf Model 51:2897–2903. https://doi.org/10.1021/ci2003889

65. Wojcikowski M, Zielenkiewicz P, Siedlecki P (2015) Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. J Cheminform 7:26. https://doi.org/10.1186/s13321-015-0078-2

66. Benjamin SL, Outeiral C, Guimaraes GL, Aspuru-Guzik A (2017) Optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC). ChemRxiv. https://doi.org/10.26434/chemrxiv5309668.v3

67. Feinberg EN, Sur D, Wu Z, Husic BE, Mai H, Li Y, Sun S, Yang J, Ramsundar B, Pande VS (2018) PotentialNet for molecular property prediction. ACS Cent Sci 4(11):1520–1530. https://doi.org/10.1021/acscentsci.8b00507

68. Awale M, Reymond JL (2019) Polypharmacology browser PPB2: target prediction combining nearest neighbors with machine learning. J Chem Inf Model 59(1):10–17. https://doi.org/10.1021/acs.jcim.8b00524

69. Olivecrona M, Blaschke T, Engkvist O, Chen H (2017) Molecular de-novo design through deep reinforcement learning. J Cheminform 9(1):1–14. https://doi.org/10.1186/s13321-017-0235-x

70. Coley CW, Rogers L, Green WH, Jensen KF (2018) SCScore: synthetic complexity learned from a reaction corpus. J Chem Inf Model 58(2):252–261. https://doi.org/10.1021/acs.jcim.7b00622

71. Yasuo N, Sekijima M (2019) Improved method of structure-based virtual screening via interaction-energy-based learning. J Chem Inf Model 59(3):1050–1061. https://doi.org/10.1021/acs.jcim.8b00673

72. Patel V, Shah M (2021) A comprehensive study on artificial intelligence and machine learning in drug discovery and drug development. Intell Med. https://doi.org/10.1016/j.imed.2021.10.001

73. Butina D (1999) Unsupervised database clustering based on daylight's fingerprint and Tanimoto similarity: a fast and automated way to cluster small and large data sets. J Chem Inf Comput Sci 39(4):747–750. https://doi.org/10.1021/ci9803381

74. Stumpfe D, Bajorath J (2012) Exploring activity cliffs in medicinal chemistry: mini perspective. J Med Chem 55(7):2932–2942. https://doi.org/10.1021/jm201706b

75. Watson OP, Cortes-Ciriano I, Watson JA (2021) A semi-supervised learning framework for quantitative structure-activity regression modeling. Bioinformatics 37(3):342–350. https://doi.org/10.1093/bioinformatics/btaa711

76. Sharma R, Lawrenson AS, Fisher NE, Warman AJ, Shone AE, Hill A, Mbekeani A, Pidathala C, Amewu RK, Leung S, Gibbons P (2012) Identification of novel antimalarial chemotypes via chemoinformatic compound selection methods for a high-throughput screening program against the novel malarial target, PfNDH2: increasing hit rate via virtual screening methods. J Med Chem 55(7):3144–3154. https://doi.org/10.1021/jm3001482

77. Shah P, Tiwari S, Siddiqi MI (2014) Integrating molecular docking, CoMFA analysis, and machine-learning classification with virtual screening toward the identification of novel scaffolds as *Plasmodium falciparum* enoyl acyl carrier protein reductase inhibitor. Med Chem Res 23(7):3308–3326. https://doi.org/10.1007/s00044-014-0910-7

78. Neves BJ, Braga RC, Alves VM, Lima MN, Cassiano GC, Muratov EN, Costa FT, Andrade CH (2020) Deep Learning-driven research for drug discovery: Tackling Malaria. PLoS Comput Biol 16(2):e1007025. https://doi.org/10.1371/journal.pcbi.1007025

79. Caballero-Alfonso AY, Cruz-Monteagudo M, Tejera E, Benfenati E, Borges F, Cordeiro MND, Jaramillo VA, Castillo YP (2019) Ensemble-based modeling of chemical compounds with antimalarial activity. Curr Top Med Chem 19(11):957–969. https://doi.org/10.2174/1568026619666190510100313

80. Butina D (1999) Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: a fast and automated way to cluster small and large data sets. J Chem Inf Comput Sci 39:747–750. https://doi.org/10.1021/ci9803381

81. Cao Y, Charisi A, Cheng LC, Jiang T, Girke T (2008) ChemmineR: a compound mining framework for R. Bioinform 24(15):1733–1734. https://doi.org/10.1093/bioinformatics/btn307

82. Janairo JIB, Janairo GC (2018) A machine learning approach in predicting mosquito repellency of plant-derived compounds. Nova Biotechnologica et Chimica 17(1):58–65

83. Jamal S, Periwal V, Scaria V (2013) Predictive modeling of anti-malarial molecules inhibiting apicoplast formation. BMC Bioinform 14(1):1–8. https://doi.org/10.1186/1471-2105-14-55

84. 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

85. Norinder U, Spjuth O, Svensson F (2020) Using predicted bioactivity profiles to improve predictive modelling. J Chem Inf Model. https://doi.org/10.1021/acs.jcim.0c00250

86. Lu WC, Chen NY, Ye CZ, Li GZ (2002) Introduction to the algorithm of support vector machine and the software ChemSVM. Comput Appl Chem 19(6):697–702

87. Yu X (2019) Prediction of depuration rate constants for polychlorinated biphenyl congeners. ACS Omega 4(13):15615–15620. https://doi.org/10.1021/acsomega.9b02072

88. Yu X, Xu L, Zhu Y, Lu S, Dang L (2019) Correlation between ^{13}C NMR chemical shifts and complete sets of descriptors of natural coumarin derivatives. Chemom Intell Lab Sys 184: 167–174. https://doi.org/10.1016/j.chemolab.2018.12.006

89. Danishuddin MG, Malik MZ, Subbarao N (2019) Development and rigorous validation of antimalarial predictive models using machine learning approaches. SAR QSAR Environ Res 30(8):543–560. https://doi.org/10.1080/1062936X.2019.1635526

90. Liu Q, Deng J, Liu M (2020) Classification models for predicting the antimalarial activity against *Plasmodium falciparum*. SAR QSAR Environ Res 31(4):313–324. https://doi.org/10.1080/1062936X.2020.1740890

91. Bharti DR, Lynn AM (2017) QSAR based predictive modeling for anti-malarial molecules. Bioinfo 13(5):154–159. https://doi.org/10.6026/97320630013154

92. Shah P, Tiwari S, Siddiqi MI (2014) Integrating molecular docking, CoMFA analysis, and machine learning classification with virtual screening toward the identification of novel scaffolds as *Plasmodium falciparum* enoyl acyl carrier protein reductase inhibitor. Med Chem Res 23(2):3308–3326. https://doi.org/10.1007/s00044-014-0910-7

93. Mahmoudi N, Garcia-Domenech R, Galvez J, Farhati K, Franetich JF, Sauerwein R, Hannoun L, Derouin F, Danis M, Mazier D (2008) New active drugs against liver stages of *Plasmodium* predicted by molecular topology. Antimicrob Agents Chemother 52(4):1215–1220. https://doi.org/10.1128/AAC.01043-07

94. Katritzky AR, Kulshyn OV, Stoyanova-Slavova I, Dobchev DA, Kuanar M, Fara DC, Karelson M (2006) Antimalarial activity: a QSAR modeling using CODESSA PRO software. Bioorg Med Chem 14(7):2333–2357. https://doi.org/10.1016/j.bmc.2005.11.015

95. Santos CB, Vieira JB, Lobato CC, Hage-Melim LI, Souto RN, Lima CS, Costa EV, Brasil DS, Macêdo WJ, Carvalho JC (2014) A SAR and QSAR study of new artemisinin compounds with antimalarial activity. Molecules 19(1):367–399. https://doi.org/10.3390/molecules19010367

96. Zhang L, Fourches D, Sedykh A, Zhu H, Golbraikh A, Ekins S, Clark J, Connelly MC, Sigal M, Hodges D, Guiguemde A (2013) Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. J Chem Inf Model 53(2):475–492. https://doi.org/10.1021/ci300421n

97. Yousefinejad S, Mahboubifar M, Eskandari R (2019) Quantitative structure-activity relationship to predict the anti-malarial activity in a set of new imidazolopiperazines based on artificial neural networks. Malar J 18:310. https://doi.org/10.1186/s12936-019-2941-5

98. Rio ALD, Llorach-Parés L, Perera-Lluna A, Avila C, Nonell-Canals A, Sanchez-Martinez M (2017) Machine-learning QSAR model for predicting activity against malaria parasite's ion pump PfATP4 and in silico binding assay validation. MDPI Proc 1(6):652. https://doi.org/10.3390/proceedings1060652

99. Keshavarzi AA, Salem MCJ, Yuan JS, Chakrabarti D (2020) DeepMalaria: artificial intelligence-driven discovery of potent antiplasmodial. Front Pharmacol 10:1526. https://doi.org/10.3389/fphar.2019.01526

100. Foko LPK, Meva FEA, Moukoko CEE, Ntoumba AA, Njila MIN, Kedi PBE, Ayong L, Lehman LG (2019) A systematic review on anti-malarial drug discovery and antiplasmodial potential of green synthesis mediated metal nanoparticles: overview, challenges and future perspectives. Malaria J 18(1):1–14. https://doi.org/10.1186/s12936-019-2974-9

101. Latha PP, Sharmila JS (2010) QSAR study for the prediction of IC_{50} and Log P for 5-N-Acetyl-Beta-D-neuraminic acid structurally similar compounds using stepwise (multivariate) linear regression. Int J Chem Res 2(1):32–38

102. Colmenarejo G, Lozano S, González-Cortés C, Calvo D, Sanchez-Garcia J, Matilla JLP (2018) Predicting transmission blocking potential of anti-malarial compounds in the mosquito feeding assay using *Plasmodium falciparum* male gamete inhibition assay. Sci Rep 8(1):1–13. https://doi.org/10.1038/s41598-018-26125-w

103. Comert G, Begashaw N, Turhan-Comert A (2020) Malaria outbreak detection with machine learning methods. BioRxiv 7(21):214213. https://doi.org/10.1101/2020.07.21.214213

104. Adjalley SH, Johnston GL, Li T, Eastman RT, Ekland EH, Eappen AG, Richman A, Sim BK, Lee MC, Hoffman SL, Fidock DA (2011) Quantitative assessment of *Plasmodium falciparum* sexual development reveals potent transmission-blocking activity by methylene blue. Proc Natl Acad Sci 108(47):E1214–E1223. https://doi.org/10.1073/pnas.1112037108

105. Syed AH, Khan T (2019) A supervised classifier-based chemoinformatics model to predict inhibitors essential for sexual reproduction and transmission of the *P. falciparum* parasite into mosquitoes. Int J Adv Appl Sci 6(2):62–72. https://doi.org/10.21833/ijaas.2019.10.011

106. Hakizimana L, Cheruiyot WK, Kimani S, Nyararai M (2017) A hybrid based classification and regression model for predicting diseases outbreak in datasets. Int J Comput 27(1):69–83

107. Zhang L, Fourches D, Sedykh A, Zhu H, Golbraikh A, Ekins S (2013) Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. J Chem Inf Model 53(2):475–492. https://doi.org/10.1021/ci300421n