Artificial Intelligence In Medicine 147 (2024) 102700

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

Optimisation-based modelling for explainable lead discovery in malaria

Yutong Li [a], Jonathan Cardoso-Silva [b], John M. Kelly [c], Michael J. Delves [c], Nicholas Furnham [c], Lazaros G. Papageorgiou [d], Sophia Tsoka [a,*]

[a] Department of Informatics, King's College London, Bush House, London, WC2B 4BG, UK
[b] Data Science Institute, London School of Economics and Political Science, Houghton St, London, WC2A 2AE, UK
[c] Department of Infection Biology, London School of Hygiene and Tropical Medicine, Keppel St, London, WC1E 7HT, UK
[d] The Sargent Centre for Process Systems Engineering, Department of Chemical Engineering, University College London, Torrington Place, London, WC1E 7JE, UK

ARTICLE INFO

Keywords:
Quantitative Structure–Activity Relationship (QSAR)
Mathematical optimisation
Piecewise linear regression
Drug discovery
Malaria
Machine learning

ABSTRACT

Background: The search for new antimalarial treatments is urgent due to growing resistance to existing therapies. The Open Source Malaria (OSM) project offers a promising starting point, having extensively screened various compounds for their effectiveness. Further analysis of the chemical space surrounding these compounds could provide the means for innovative drugs.

Methods: We report an optimisation-based method for quantitative structure–activity relationship (QSAR) modelling that provides explainable modelling of ligand activity through a mathematical programming formulation. The methodology is based on piecewise regression principles and offers optimal detection of breakpoint features, efficient allocation of samples into distinct sub-groups based on breakpoint feature values, and insightful regression coefficients. Analysis of OSM antimalarial compounds yields interpretable results through rules generated by the model that reflect the contribution of individual fingerprint fragments in ligand activity prediction. Using knowledge of fragment prioritisation and screening of commercially available compound libraries, potential lead compounds for antimalarials are identified and evaluated experimentally via a *Plasmodium falciparum* asexual growth inhibition assay (PfGIA) and a human cell cytotoxicity assay.

Conclusions: Three compounds are identified as potential leads for antimalarials using the methodology described above. This work illustrates how explainable predictive models based on mathematical optimisation can pave the way towards more efficient fragment-based lead discovery as applied in malaria.

1. Introduction

Malarial represents a current unmet medical need. The World Health Organisation estimated 247 million cases across 84 endemic countries, leading to 619,000 deaths [1] in 2021. New therapeutics are urgently required due to the high prevalence of parasite resistance to the existing approved drugs [2,3]. The development of new antimalarial drugs, however, is a complex and expensive process comprising multiple stages with many potential points of failure. Machine learning (ML) strategies can reduce the cost and time requirements associated with early-stage drug discovery by excluding unsuitable compounds and directing the search towards the most promising drug candidates [4,5].

Identification of active antimalarial compounds relies on a blend of experimental and computational strategies. High-content imaging can be used to screen large compound libraries effectively, but has limitations relating to chemical diversity and high resource needs. On the computational side, Quantitative Structure–Activity Relationship (QSAR) modelling [6] is prevalent and popular in antimalarial drug research [7]. QSAR methods predict the biological activity of chemical compounds based on their structural properties and can link different functional groups to the relevant activity [8].

Nowadays, AI methods to train QSAR models have gained popularity, as models to predict the activity of new compounds can accelerate the virtual screening process. In antimalarial drug discovery examples, an SA-SVM-based model was used to predict the activity of fusidic acid derivatives as antimalarial agents [9], we note a GA-SVM-based predictor used to model falcipain inhibitors [10], a standard protocol built with Support Vector Machine (SVM), K-Nearest Neighbours (KNN) and Navier Bayes (NB) to develop molecular descriptor-based predictive models [11], as well as MAIP, a consensus model that combines the prediction of eleven "partner" models trained on a large dataset [12].

Explainable AI (XAI) has become an important research target that aims to provide informative explanations alongside ML models to aid human decision-making and reasoning [13]. XAI is particularly

* Corresponding author.
E-mail address: sophia.tsoka@kcl.ac.uk (S. Tsoka).

https://doi.org/10.1016/j.artmed.2023.102700
Received 21 March 2023; Received in revised form 17 October 2023; Accepted 29 October 2023
Available online 22 November 2023

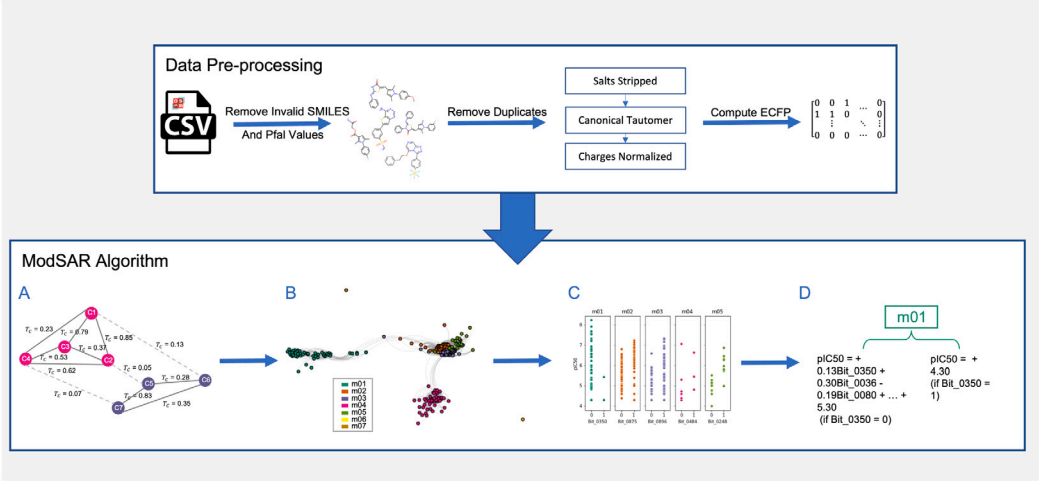


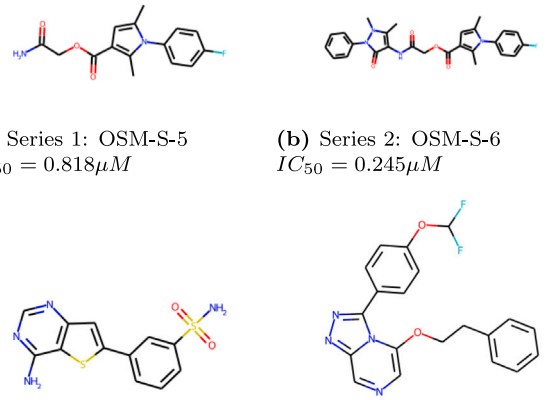
Fig. 1. The pipeline of data processing and analysis in modSAR. (A) The OSM dataset was pre-processed following the steps specified in the OSM Github issue, and the Extended-Connectivity Fingerprints (ECFP) [19] of each OSM compound were obtained. (B) The T_c similarity between compounds was computed using ECFP, and a representative graph was obtained by linking compounds with respect to their similarity. (C) A threshold of similarity was applied to the graph, and network modules were identified. (D) For each module, a bit of ECFP was identified as breakpoint. Each module was then subdivided into two regions according to the values of the breakpoint feature. (E) A regression equation was determined for each region.

important in the pharmaceutical domain due to its potential to support rational molecule design and compound optimisation, as well as to accommodate safety concerns. Therefore, XAI is an integral component of trustworthy AI-based drug discovery and is required to complete the data-information-knowledge-wisdom spectrum [14]. In the application of drug discovery, XAI models allow inference of how a model makes a particular prediction (transparency), elucidation of why the prediction is acceptable (justification), extraction of knowledge from the model to guide human decision-making (informative) and investigation of how reliable the prediction is (uncertainty estimation) [15]. In this work, we attempt to engage the concept of XAI with white-box QSAR modelling based on mathematical optimisation to obtain informative knowledge for wet-lab experiment guidance and compound screening.

This study centres on explainable QSAR modelling for antimalarials, specifically focusing on the chemical space from the Open Source Malaria (OSM) [16] project and modelling the inhibition activity (pIC_{50}) of compounds against *P.falciparum*. Amongst the compounds present in the OSM database, there is particular interest in developing accurate prediction of anti-PfATP4 activity among a set of series of promising compounds [16]. To perform this task, we build upon previous work on development of an interpretable model based on mathematical optimisation, modSAR [17]. We have previously applied modSAR to an earlier version of the OSM data using pre-defined molecular descriptors [18], and here we develop this work further, offering improved understanding of these antimalarial candidates, as well as prioritising and validating potential lead compounds as inhibitors of asexual growth in *P.falciparum* and demonstrating parasite selectivity. This study could pave the way for future SAR explorations, lead optimisation and new *de novo* drug design efforts for malaria, as well as in leveraging high-content screening relating to other disease indications.

2. Methodology

A schematic overview of our methodology is shown in Fig. 1 and comprises data pre-processing, QSAR modelling via modSAR and analysis of the rules generated by the model. ModSAR involves first detecting clusters of chemical compounds, and then applying mathematical optimisation to determine the optimal split of each cluster into appropriate regions and yield piecewise linear regression equations to link molecular descriptors to the biological activity of samples in that region.



(a) Series 1: OSM-S-5 $IC_{50} = 0.818\mu M$

(b) Series 2: OSM-S-6 $IC_{50} = 0.245\mu M$

(c) Series 3: OSM-S-106 $IC_{50} = 0.088\mu M$

(d) Series 4: OSM-S-369 $IC_{50} = 0.251\mu M$

Fig. 2. Initial compounds of each OSM series.

Therefore, modSAR identifies the relationship between compound features and relevant bioactivity in a manner that is both mathematically descriptive, as well as having similar accuracy to popular machine learning methods [17,20]. Below, data processing and methodology are described in more detail.

2.1. Data

Data used in this study were derived from the OSM project, a collaborative consortium aiming to facilitate design of new drugs for malaria guided by open source principles [16]. Data were downloaded from the Google Sheet of all OSM Compounds. Molecules were categorised into four series according to chemotype: an arylpyrrole series (Series 1), the triazolourea singleton (Series 2), aminothienopyrimidine Series (Series 3), and triazolopyrazine series (Series 4). The compounds that characterise each series are shown in Fig. 2.

Although targets of Series 1–3 compounds remain unknown, a promising biological target of *P.falciparum* has been identified as P-type ATPase PfATP4, a parasite cell membrane enzyme which exports Na+ ions and imports H+ ions [21,22]. Based on previous studies, PfATP4

has been implicated as target for Series 4 compounds [23], as well as two experimental antimalarial drugs, Cipargamin (KAE609) [24], (+)-SJ733 [25]. The first three series were derived from the Tres Cantos list of hits against *P. falciparum* released by Glaxo Smith-Kline in 2010 [26] but, although several potent drug candidates were found, structural difficulties have hindered progress.

In **Series 1**, a labile ester created stability concerns, with potency of compounds decreasing whenever changes were made to the central structure. **Series 2**, on the other hand, presented low solubility [27], and in **Series 3**, the mechanisms of action of the initial compound are still under active investigation, as it is believed to inhibit one or more kinases [28,29]. However, the analogues derived and evaluated in the series have not exhibited high potency.

The last set, **Series 4**, is the current series of interest of the OSM consortium [30]. These triazolopyrazine analogues were initially identified in a high throughput screening performed in 2013 by Pfizer and the Medicines for Malaria Venture (MMV) [31,32] and contain many potent compounds, some of which have proven to be potent *in vivo*, and display many desirable physicochemical properties [29]. A correlation has been found between molecular potency and parasite ion-regulated activity [33].

This study focused first on building a predictive model for Series 4 analogues [18], but as our method can inherently distinguish structurally heterogeneous chemical sets, all series and assays were then considered for a more comprehensive analysis. A raw dataset containing all OSM compounds from Series 1 to 4 and their respective assay data was downloaded from the Master List of chemicals provided by OSM. Pre-processing (Fig. 1(A)) was performed as outlined in [34]: (i) compounds with no SMILES or *P.falciparum* IC_{50} values were removed; (ii) molecular structures were normalised using RDkit [35] with salts stripped, canonical tautomer calculated and charges normalised; (iii) data were deduplicated by recalculating each compound's InChiKey; and (iv) pIC_{50} values were computed by taking the negative logarithm of the IC_{50} values to obtain an easy-to-read, intuitive form of data. The final dataset included 386 unique compounds, each with a respective SMILES string and an associated binding activity (pIC_{50}).

For each compound, circular molecular fingerprints were generated by RDKit [35] using the Morgan algorithm [36]. Specifically, each atom in a molecule was viewed as the centre of a radius of perception. The substructures of a molecule were iteratively gathered and recorded by including the immediate neighbours and the neighbours of each centre. In preliminary tests, we observed similar performance of the modSAR algorithm for fingerprints produced with radius = 2 (ECFP4) and radius = 4 (ECFP8) parameters, so we selected the version with a smaller radius. Our final configuration consists of Morgan circular fingerprints of radius = 2 collapsed to 1024 bits, closely resembling the ECFP4 fingerprint algorithm commonly used in cheminformatics studies [19].

2.2. The ModSAR algorithm for network-based piecewise linear regression

ModSAR (Fig. 1(B)-(E)) combines modularity clustering [37] and regularised piecewise linear regression [20] to learn the quantitative structural–activity relationships (QSAR) relationship of compound activity [17]. The algorithm involves two main stages: first modules of molecules that share similar structures are identified, and then each such module is modelled to derive piecewise linear equations through the OPLRAreg optimisation model. An overview of the computational procedure is illustrated in Fig. 1 and described below.

Similarities among compounds are calculated by the pairwise Tanimoto coefficient Tc [38] applied to the circular fingerprints [19]. Pairs of compounds are connected by an edge in the network, if chemical similarity is above a threshold $Tc \geq t_\alpha$, which is identified by modSAR and corresponds to the value that optimises the average clustering coefficient of the network [17,39]. Given the chemical similarity network, compounds are clustered in distinct modules by maximising the modularity metric [40–43] with each module reflecting compounds with a common structural core or scaffold.

The second stage derives the structure–activity mapping, through independent piecewise linear regression equations using the OPLRAreg algorithm [20,44]. Here, one of the features is optimally determined to act as a breakpoint, thereby separating the data into n disjoint sub-groups called "regions", each of which is then modelled by independent linear equations. OPLRAreg identifies all of these properties simultaneously (i.e. optimal feature, number of regions and regression coefficients), maximising the mean absolute error (MAE) of pIC_{50} value through a mixed integer linear programming (MILP) optimisation model.

At each iteration, the breakpoint values are ordered for the selected number of regions R and breakpoint feature f^*:

$$X_{rf^*} \geq X_{r-1,f^*} \qquad \forall r = 2, 3, \ldots, R-1 \tag{1}$$

where X_{rf^*} denotes the breakpoint value of region r on breakpoint feature f^*. Each sample s is assigned to a region r corresponding to the breakpoint values:

$$A_{sf^*} \geq X_{r-1,f^*} - U1(1 - F_{sr}) + \epsilon \qquad \forall s, r = 2, 3, \ldots, R \tag{2}$$

$$A_{sf^*} \leq X_{rf^*} + U1(1 - F_{sr}) - \epsilon \qquad \forall s, r = 1, 2, \ldots, R-1 \tag{3}$$

where A_{sf^*} denotes the numeric value of sample s on breakpoint feature f^*, $U1$ is a large positive number, and binary variable F_{sr} is introduced to decide whether a sample s belongs to a region r. Parameter ϵ is added to the model to ensure that no values of the dataset will equal any of the breaking points.

Each sample is restricted to belonging in only one region:

$$\sum_r F_{sr} = 1 \qquad \forall s \tag{4}$$

For any sample, the predicted value P_{sr} for a sample s in region r is given by Eq. (5):

$$P_{sr} = \sum_f W_{rf} A_{sf} + B_r \qquad \forall s, r \tag{5}$$

where W_{rf} and B_r are the regression and intercept for feature f in region r. For each sample s, the training error is equal to the absolute error E_s between the observed value O_s and predicted value P_{sr} of the assigned region (i.e. $F_{sr} = 1$):

$$E_s \geq O_s - P_{sr} - U2(1 - F_{sr}) \qquad \forall s, r \tag{6}$$

$$E_s \geq P_{sr} - O_s - U2(1 - F_{sr}) \qquad \forall s, r \tag{7}$$

where $U2$ is an arbitrarily large positive number.

The objective function of OPLRAreg is shown in Eq. (8) below,

$$z = MAE + \lambda \cdot REG \tag{8}$$

where λ is a positive user-defined parameter that controls the influence of regularisation. Variables MAE and REG are defined by the set of equations below,

$$MAE = \frac{\sum_s E_s}{|s|} \tag{9}$$

$$REG = \sum_f W^+_{rf} \tag{10}$$

$$W^+_{rf} \geq W_{rf} \qquad \forall r, f \tag{11}$$

$$W^+_{rf} \geq -W_{rf} \qquad \forall r, f \tag{12}$$

The summary of the mixed integer linear programming model, OPLRAreg, is given by:

$$\begin{aligned} &minimise \quad z \\ &subject\ to \quad \text{Eqs. (1)–(12)} \end{aligned} \tag{13}$$

The model is formulated as MILP problem which can be solved to global optimality, and derives an optimal subset of features to be used in each equation, as controlled by a regularisation parameter $\lambda \geq 0$. For $\lambda = 0$, no regularisation is enforced and the linear equation can have as many features as possible, which incurs a risk of overfitting the data. Larger λ values reduce the number of features included in the equation while reducing the risk of overfitting. In order to identify the most common scaffold in relevant groups of compounds the rdScaffoldNetwork algorithm was employed [45].

2.3. Model inference

Piecewise linear regression equations identified by modSAR can form the basis for structure–activity interpretation. It is noted that, as regression equations are fitted independently for each module, data can be further split into as many sub-clusters (i.e. regions) as required to minimise regression error [17,20]. In practice, the algorithm selects a single feature to serve as breakpoint for defining regions in each iteration, and as we are handling binary data, there are typically two disjoint regions for each module.

An advantage stemming from the methodological basis of our work and the associated use of circular fingerprints, is that we can reverse each fingerprint bit to the relevant chemical fragment. As certain bits are selected by the optimisation procedure in modSAR and incorporated to the regression equation, a latent association of the relevant substructure to the binding activity can be generated. We take advantage of the bit–fragment relationship to evaluate the presence and prevalence of certain fragments in network modules and piecewise regions to hypothesise on their contribution to the compound activity.

In addition to fragment prioritisation by modSAR, SHapley Additive exPlanations (SHAP) value [46] analysis is used to further inspect fragment contributions [47]. SHAP values interpret the output of a machine learning model by connecting optimal credit allocation with local explanations using the classic Shapley values from game theory. In this work, we applied SHAP barplot [48,49] for computing feature importance, in addition to results generated by modSAR.

2.4. Tuning λ

To identify a suitable hyper parameter setting, five-fold cross validation was performed for different λ values. For each λ, the dataset was split into five sets, and each set was used separately as test set while the other four portions were used for training in each round of model training. The mean Root Mean Squared Error(RMSE) of the five-fold cross validation for each λ was utilised as evaluation metric to provide an indication of model fitness [50].

2.5. Plasmodium falciparum asexual growth inhibition assay (PfGIA)

To validate the compounds selected by the modSAR algorithm, a *P.falciparum* asexual growth inhibition assay, similar to the one used in the original screens that was performed on the OSM derived data. The assay used *P.falciparum* 3D7 strain parasites maintained in complete culture medium (RPMI supplemented with 25 mM HEPES, 50 μg ml^{-1} hypoxanthine, 2 gl^{-1} $NaHCO_3$, 0.3 gl^{-1} glutamine, 1% Albumax II) at 37 °C under a 5% CO_2 atmosphere at 4% haematocrit in whole human blood (National Blood and Transfusion Service).

Flat-bottomed 96-well assay-ready plates were prepared by automated dispensing of 10 mM DMSO stock solutions of the compounds using a Tecan D300 Digital Dispenser to give appropriate dilutions. DMSO alone was used as a negative control and dihydroartemisinin at a final concentration of 100 nM was used as a positive control. Parasites were synchronised with 5% sorbitol and cultures containing ring stage parasites were diluted to 2% parasitaemia and 1% haematocrit. 100 μl of diluted culture was dispensed into each well of the assay plates ($n = 3$ independent replicates) and incubated in a humidified chamber within a 37 °C incubator. After 72 h, plates were removed and frozen at −20 °C overnight to aid lysis. The next day, plates were thawed and duplicate plates containing 100 μl lysis buffer (20 mM Tris pH 7.5, 5mM EDTA, 0.008% w/v saponin, 0.08% w/v Triton X-100, 1:5,000 SYBR green) were prepared. Resuspended thawed parasite culture from each well (85 μl) was added to the corresponding well of plates containing lysis buffer and incubated in the dark at room temperature for 1 h. SYBR green fluorescence was then read in a fluorescence plate reader with 490 nm excitation and 520 nm emission settings. Emission values for the positive and negative controls were used to normalise the data and convert values to percentage inhibition. IC_{50} values were then calculated using GraphPad Prism 5.

2.6. Cytotoxicity assays

To assess whether compounds that showed activity against *P.falciparum* in the asexual growth inhibition assay had parasite specificity, a cytotoxicity assay was conducted. HeLa cells were seeded into 96-well microtiter plates at 2.5×10^4 m L^{-1} in 200 μL of growth medium, and compounds added at a range of concentrations. Plates were incubated at 37 °C in a 5% CO_2 atmosphere for 5 days, then resazurin (20 μL at 0.125 mg m L^{-1}) was added and the plates incubated for a further 6 h. Fluorescence was determined using a BMG FLUO star Omega plate reader (excitation 488 nm, emission 525 nm), and the data analysed using GraphPad Prism 8 software. Values are expressed as $IC_{50} \pm SD$ and are the average of three replicates.

3. Results

The performance of modSAR is first evaluated using five-fold cross validation and selection of hyper-parameter λ by grid search. After obtaining the best-performing model, we discuss the properties of the network modules derived from modSAR and the relevant piecewise linear equations. We then demonstrate the process of inferring rules via the trained modSAR model and prioritising antimalarial fragments based on their contribution to the compound binding activity.

3.1. Cross validation

The result of five-fold cross validation at 20 different λ values is shown in Figure S1a. The result shows that the RMSE in the test set is almost always slightly higher than the training set, at approximately 0.85, indicating that modSAR fits the OSM dataset well with no indication of overfitting. Overall, the modSAR model performs best around $\lambda = 0.06$. Similar results between training and testing sets were found when performance was assessed via Mean Absolute Error (MAE) (S1b). Average running time for each parameter is shown in Figure S1c.

3.2. Network modules

The chemical similarity network of the dataset as partitioned into clusters, is shown in Fig. 3. Edges represent pairwise Tanimoto similarity that exceed the optimal threshold, calculated by the algorithm to be $t_\alpha \geq 0.20$. Nodes are coloured according to their cluster membership.

To provide a first visual inspection of structure–activity relationships, the node (compound) with the highest within-module degree is selected as the representative compound of each module, thus depicting the structural characteristics of neighbouring compounds. The most common scaffold in each representative compound for each module is highlighted in red (Fig. 3). It is important to note that the highlighted scaffold represents the most dominant structure, and it may be that not all compounds in the module contain that substructure.

The clustering procedure captures the chemical properties of the dataset in terms of molecular properties reflecting the OSM compound series. The five modules as partitioned by algorithm, closely match the analogue series present in the OSM dataset (see Table S1 and the

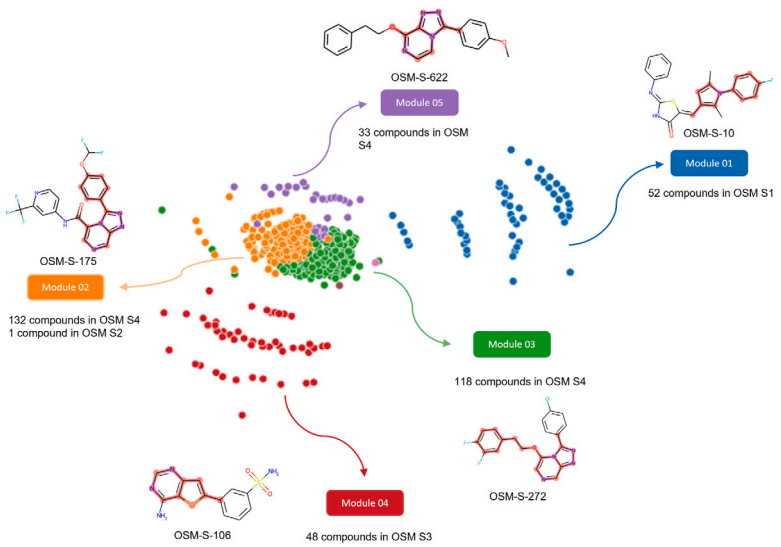


Fig. 3. Compound network modules identified by modularity optimisation. Edges indicate Tanimoto similarity, $t_\alpha \geq 0.20$, and colours signify cluster membership. Representative compounds and dominant scaffolds for each cluster are also shown.

comparative visualisation in Figure S2). OSM Series 1 and Series 3 compounds are members of Modules m01 and m04, respectively, while Series 4 compounds have been allocated to three distinct modules: m02, m03 and m05. The Series 2 structure (OSM-S-66) was assigned to m02, as it shares structural similarities to two compounds in that module, namely OSM-S-359 and OSM-S-570. There were two singletons, OSM-S-89 and OSM-S-69 (not shown in the figure) as these two structures differ significantly from the rest in the dataset.

A closer inspection of modules related to OSM Series 4 (m02, m03 and m05) and their associated highlighted scaffold in Fig. 3 allows for further insight into this dataset. Each module represents core substructures that are more specific than the Series 4 triazolopyrazine core, and it is assumed that each detected module would have their structure–activity relationships modelled individually. Therefore, the subsequent sections below describe how the predicted equations of each of these modules compare, and how the molecular fragments that relate to bioactivity within the core are represented by each module.

3.3. Analysis of modules via regression equations

A summary of rules and equations is shown in Table 1, where the breakpoint features and the equations identified for each cluster are shown. The distribution of pIC_{50} under different subsets and the presence of certain bits can be seen in Figure S3. All bits selected by modSAR are visualised in Figure S4. A detailed description and interpretation of each module is discussed below.

Module related to OSM series 1. The structure–activity relationship of compounds in OSM Series 1 derived by modSAR, is represented by module m01 and the equations shown in Table 1. Most chemical compounds in this module share a common scaffold, with 43 of the 52 compounds containing the fragment highlighted in Fig. 3.

Deriving from regression equations, activity of compounds in this module is predicted by one of two linear equations depending on presence or absence of fragment Bit_0350. When the fragment corresponding to Bit_0350 is not present (i.e. Bit_0350 = 0), the activity of that compound is predicted by the presence of eleven fragments (Bit_0350 included). On the other hand, if a compound includes this fragment (Bit_0350 = 1), the model predicts a bioactivity pIC_{50} = 4.30, thus the compound is inactive against *P.falciparum* (assuming an activity threshold $pIC_{50} \geq 5.80$ [18]. By exploring the **m01** (Region 01) equation, we can also see which of the remaining fragments selected by the algorithm make positive or negative contributions to the bioactivity of these compounds.

Beyond the observation of signal and magnitude of regression coefficients, we rank the importance of fingerprint bits according to their relative SHAP values (Figure S5a). In decreasing order of importance, the presence of fragments Bit_0290, Bit_0036, Bit_0703, Bit_0332, Bit_0031, Bit_0175, and Bit_0961 are predicted to make a positive contribution, while Bit_0745, Bit_0350, Bit_0080, Bit_1017, and Bit_0790 have a negative contribution to activity. Indeed, we can observe the association of the most positive and negative bits (Bit_0290 and Bit_0745 respectively) to the pIC_{50} activity of compounds in Fig. 4 and Figure S3a. Comparing the activity of compounds in this module, the compounds which only contain Bit_0290 is much higher than the ones which only contain Bit_0745. The combination of positive and negative contributing fragments is illustrated in Figures S6a and S6b.

Modules related to OSM series 3. Module **m04** describes the binding activity of compounds in OSM Series 3. The regression equations in Table 1 suggest that the Series 3 compounds are not likely to be active, as the maximum activity values calculated from the two equations are either pIC_{50} = 5.47 (when Bit_0484 = 1) or pIC_{50} = 4.60 (when Bit_0484 = 0 and Bit_0179 = 1), lower than our defined threshold of $pIC_{50} \geq 5.80$ [18]. This is also supported by the distribution of the true pIC_{50} values of Series 3 compounds. As shown in Figure S3c, the median and mode value of Series 3 are both less than 5, and only 2 out of 49 compounds are predicted to be active, namely OSM-S-106 and OSM-S-590. Importance ranking of the two bits can be seen in Figure S5e and combination of positive and negative contributing fragments can be seen in Figure S6g and S6h.

Modules related to OSM series 4. Module **m02**, module **m03** and module **m05** describe the binding activity of Series 2 and Series 4 via the piecewise equations demonstrated in Table 1. Since Series 2 corresponds to a single compound, we focus our analysis on Series 4 properties.

A common feature of these modules is the prominence of Bit_0890 in the equations, with presence of the relevant fragment predicted to make a positive contribution in binding affinity. Interestingly, the structure of Bit_0890 (Fig. 5(b)) is a close, albeit not exact, match to the triazolopyrazine core of the series (Fig. 5(a)).

If all fragment bits that are predicted to make positive contributions to the binding activity are combined, the fragment shown in Fig. 5(d) is obtained. This visualisation suggests that, in addition to the triazolopyrazine core for Series 4 which is approximately represented by

Table 1
Equations and breakpoints identified for modules indicated in Fig. 3.

Module	Region	Decision rule	Selected fragment	Equation
Modules relating to **OSM Series 1**				
m01	01	if Bit_0350 = 0		$pIC_{50} = +0.13\ \text{Bit\_0031} + 0.30\ \text{Bit\_0036} - 0.19\ \text{Bit\_0080} + 0.12\ \text{Bit\_0175} + 0.53\ \text{Bit\_0290} + 0.14\ \text{Bit\_0332} + 0.14\ \text{Bit\_0703} - 0.81\ \text{Bit\_0745} - 0.15\ \text{Bit\_0790} + 0.06\ \text{Bit\_0961} + 0.38\ \text{Bit\_1017} + 5.30$
	02	if Bit_0350 = 1		$pIC_{50} = +4.30$
Modules relating to **OSM Series 3**				
m04	01	if Bit_0484 = 0		$pIC_{50} = +0.20\ \text{Bit\_0179} + 4.40$
	02	if Bit_0484 = 1		$pIC_{50} = +5.47$
Modules relating to **OSM Series 4**				
m02	01	if Bit_0875 = 0		$pIC_{50} = +0.03\ \text{Bit\_0711} + 0.03\ \mathbf{\text{Bit\_0890}} + 5.00$
	02	if Bit_0875 = 1		$pIC_{50} = +6.15$
m03	01	if Bit_0896 = 0		$pIC_{50} = +0.16\ \mathbf{\text{Bit\_0890}} + 4.84$
	02	if Bit_0896 = 1		$pIC_{50} = -0.04\ \text{Bit\_0650} + 6.03$
m05	01	if Bit_0248 = 0		$pIC_{50} = +0.06\ \mathbf{\text{Bit\_0890}} - 0.06\ \text{Bit\_0171} + 0.002\ \text{Bit\_0333} + 0.06\ \text{Bit\_0399} - 0.12\ \text{Bit\_0512} - 0.06\ \text{Bit\_0715} - 0.06\ \text{Bit\_0753} - 0.06\ \text{Bit\_0769} - 0.16\ \text{Bit\_0781} - 0.06\ \text{Bit\_0785} + 0.06\ \text{Bit\_0819} - 0.06\ \text{Bit\_0838} - 0.12\ \text{Bit\_0841} - 0.06\ \text{Bit\_0939} + 5.00$
	02	if Bit_0248 = 1		$pIC_{50} = +0.50\ \text{Bit\_0904} + 5.82$

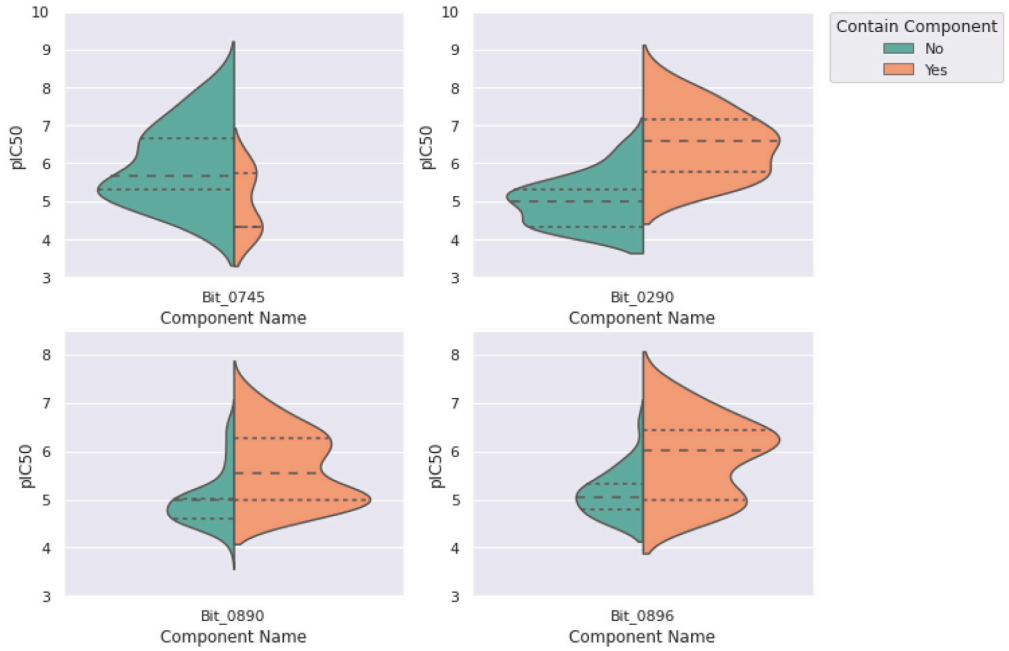


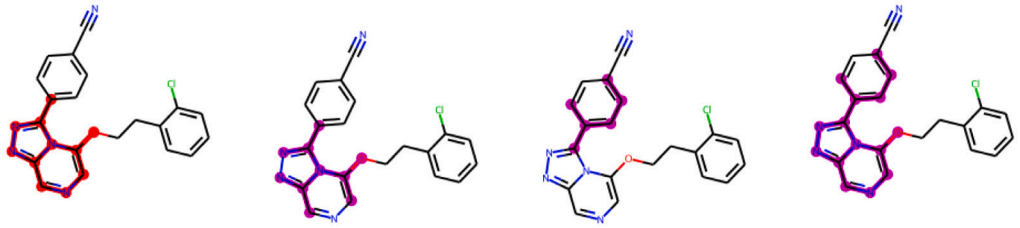
Fig. 4. Comparison of pIC_{50} distributions in different subsets of the data split according to fragment presence.

Bit_0890 = 1, the fragment Bit_0896 (highlighted in Fig. 5(c)) should be retained to maximise the activity of Series 4 compounds, as suggested by the regression equations and SHAP value analysis.

A similar conclusion can be drawn by comparing the two distributions of pIC_{50} values corresponding to compounds with and without Bit_0896, using a one-sided t-test with the following hypothesis:

$$H_0 : E(A) = E(B), H_1 : E(A) < E(B) \tag{14}$$

where A denotes the population of compounds which do not contain Bit_0896, B denotes the population of compounds which contain



(a) The fragment that represents triazolopyrazine core.

(b) The fragment that represents Bit_0890.

(c) The fragment that represents Bit_0896.

(d) Combination of all positive bits in Series 4 modules.

Fig. 5. Visualisation of relevant fragments present in OSM Series 4 modules.

Table 2
Comparison of test set error in y-randomisation. Comparative results of the original model and permutations are shown in terms of average RMSE and standard deviation.

λ	Model 1 y vs. rx	Model 2 py vs. x	Model 3 ry vs. x	Model 4 ry vs. rx	Model 5 py vs. rx	Original model
0.05	0.9569 (± 0.1007)	0.9480 (± 0.1636)	1.2334 (± 0.1146)	1.330 (± 0.1593)	0.9005 (± 0.0717)	**0.8367** (± 0.1504)
0.06	0.9515 (± 0.0959)	0.9685 (± 0.1429)	1.1473 (± 0.1331)	1.2064 (± 0.1340)	1.0363 (± 0.1463)	**0.8065** (± 0.1302)
0.07	0.9092 (± 0.0732)	0.9093 (± 0.0811)	1.1387 (± 0.14284)	1.2959 (± 0.1399)	0.9208 (± 0.0639)	**0.8660** (± 0.1741)
0.08	0.9044 (± 0.1279)	0.9192 (± 0.1035)	1.1820 (± 0.0851)	1.2859 (± 0.0953)	0.8950 (± 0.1135)	**0.8591** (± 0.1442)
0.09	0.8858 (± 0.1011)	0.9036 (± 0.1208)	1.1464 (± 0.0552)	1.2668 (± 0.7561)	0.9164 (± 0.1289)	**0.8790** (± 0.1193)
0.10	0.9056 (± 0.1073)	0.9308 (± 0.0781)	1.2048 (± 0.1262)	1.2224 (± 0.0936)	0.8908 (± 0.0864)	**0.8819** (± 0.1462)

Bit_0896. The obtained p-value, $p = 2.83e-16$, suggests that it is safe to reject the null hypothesis at any widely adopted confidence level (95% or 99%).

A similar analysis can be made for each module. For example, Figures S6i and S6j compare the positive and negative contributing fragments specific to Module m05. Additional plots for OSM Series 4 modules can be seen in Figures S5c, S5d, S5b, and S6.

4. Model evaluation

Having tuned the λ parameter as outlined in section Tuning λ, performance of modSAR is assessed through Sections Y-Randomisation and Applicability Domain [51].

4.1. Y-randomisation

Y-randomisation was employed as validation to compare the performance of a QSAR model with pseudo-random models trained on permuted datasets [52]. To benchmark modSAR against predictions by chance, pseudo-random models were implemented as follows. Three different sets of pseudo-random data were generated via randomised fingerprints (rx), randomised pIC_{50} (ry), and permuted pIC_{50} (py). Randomised fingerprints, rx, were generated by assigning 0 or 1 to the 1024 fingerprints bit for each molecule, ry was generated using random numbers within the range of real pIC_{50} value, and py was generated by shuffling the real pIC_{50} value.

Five pseudo-random models were trained with synthetic datasets, i.e. (1) model 1: trained with rx and y; (2) model 2: trained with x and py; (3) model 3: trained with x and ry; (4) model 4: trained with rx and ry; and (5) model 5: trained with rx and py. We compared the performance of the five pseudo-random models with the original modSAR model using different λ (from 0.05 to 0.1) through 10-fold cross validation. The mean and standard deviation of the model RMSE outperforms the pseudo-random datasets as shown in Table 2.

4.2. Applicability domain

The applicability domain (AD) defines the chemical space covered by the model, indicating its reliability in predicting new compound properties. In this study, the AD of modSAR is determined by the leverage approach [50], which calculates the leverage and standard residual of a compound, visualised via Williams plot. The training data point with higher leverage is considered to have a larger impact during the training process. A critical leverage value h^* is calculated by the equation: $h^* = 3p'/n$, where p' is the number of model variables plus one, and n is the number of the objects used to calculate the model. The data point whose leverage is higher than h^* is considered to be outside the AD of the QSAR model. Due to the sparsity of the original feature space, the leverage of the dataset cannot be computed directly. Therefore, Principal Component Analysis (PCA) was applied to the feature space before calculating the leverage, and the first two principal components (PCs) were selected to represent the original features.

As shown in Fig. 6, all compounds were within the warning leverage, showing that there were no highly influential compounds structurally. Seven compounds from the training set, and one compound from the test set had residuals > 3, indicating response outliers or potential activity cliffs.

5. Virtual screening

To validate the results gained from modSAR, virtual screening [53] was performed on compounds commercially available from Molport, followed by experimental validation of predictions using parasite inhibition and cytotoxicity assays. Compounds were selected by the following strategy. Firstly, each positive bit prioritised by modSAR was used to search the Molport catalogue, yielding a total of 934 compounds. Secondly, the activity of these compounds was computed using modSAR and compounds where the predicted pIC_{50} was higher than 5.8 were retained, resulting in 97 compounds. Thirdly, filtering

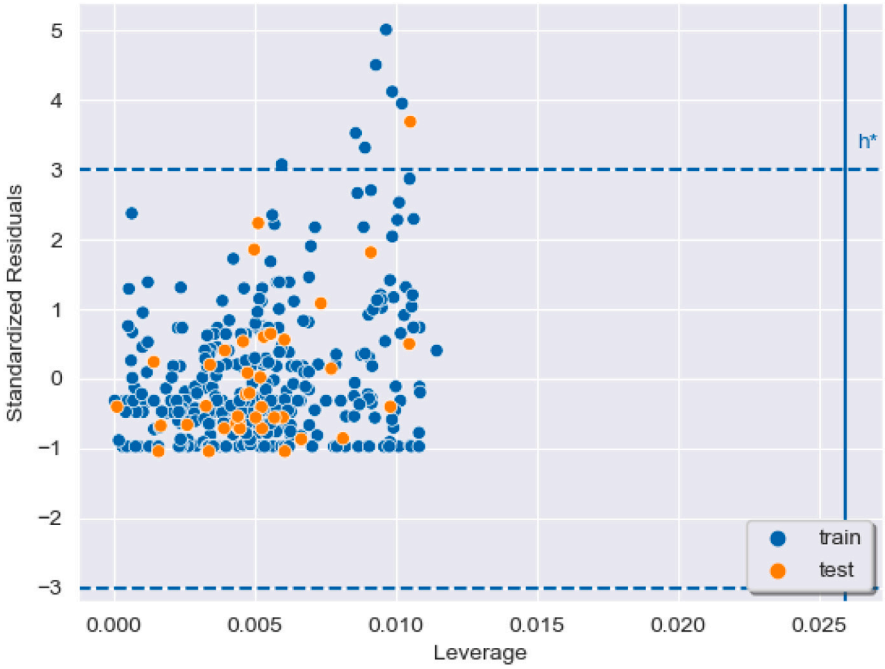


Fig. 6. Williams Plot to evaluate the Applicability Domain of modSAR.

based on violation of the Lipinski rule of 5 and solubility properties returned 22 compounds, which were ranked by activity score and number of positive bits, as identified by modSAR and SHAP analysis. Figure S7 illustrates these 22 compounds together with positive bits and predicted pIC_{50} values. As additional step, the 22 compounds were also assessed through the MAIP resource [12] to evaluate potency, with results provided in Supporting Information. These compounds were tested for their activity against the parasite *in vitro*, as follows.

5.1. Experimental validation

To validate the activity of the compounds predicted from the virtual screen experimentally, the top set of compounds were sourced via Molport from commercial providers. Each compound was evaluated by testing for inhibition of growth in the asexual life stage of *P.falciparum* using a standard assay (described in *Plasmodium falciparum* asexual growth inhibition assay (PfGIA) section) at a fixed concentration of 10 μM. Three compounds (A02, A10, A22 corresponding to MolPort-047-964-374, MolPort-047-964-639, MolPort-046-842-243) (Fig. 7) showed some activity (see Fig. 8). These were further evaluated to obtain IC_{50} values, with the top hit A22 obtaining an IC_{50} of 8.29 μM (see Fig. 9). To test if these three active compounds have parasite specificity, a cytotoxicity assay was performed using human HeLa cells (see methods for details). The IC_{50} values of A02, A10 and A22 are 109 ± 30 μM, 41 ± 1 μM and 115 ± 4 μM respectively, with the most potent compound A22 having a near 14-fold difference in IC_{50} between parasite and human cell activity, demonstrating good parasite selectivity.

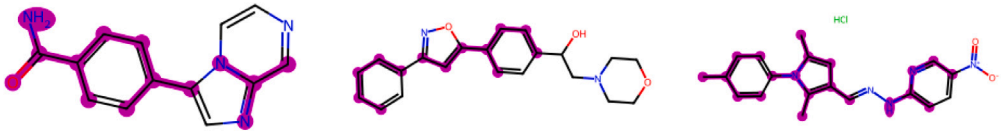
6. Discussion and conclusion

Machine learning methodologies that incorporate explainability principles [54] are particularly pertinent in drug discovery due to their ability to provide a trail of reasoning for model prediction [55,56], also noted in models for antimalarial drug discovery [57–59]. However, in some cases such interpretation may only relate to understanding the decision-making mechanism of the model at computational level, and may not extrapolate to knowledge that can feed back to the discovery process and uncover new compounds, for example via fragment-based drug discovery (FBDD). As reported widely in literature, drug discovery is a notoriously difficult problem, with a small success rate despite the rise in available methodologies from computational and measurement sciences. Methodologies that bridge the gap between computation and wet-lab experimental validation, such as the one reported in our manuscript, have the potential to add significant impact to relevant literature.

In this work, we report the use of optimisation-based regression modelling coupled with network clustering to mine and analyse publicly available data derived from phenotypic screens of potential antimalarial molecules. In contrast to inherently black-box models that cannot respond to the need of pharmaceutical scientists for continuous improvement of molecular screening and compound optimisation, modSAR employs a mathematically descriptive optimisation model that can learn from available datasets, determine optimal breakpoints for piecewise regression and prioritise descriptors by optimal regression coefficients, thereby resulting in fragment-based insights and generating rules to guide virtual screening.

We illustrated the use of the modSAR piecewise linear regression modelling method as applied to the OSM dataset using ECFP fingerprints as features to describe each compound. Analysis of results showed that the method represented the heterogeneity among different chemical series in the OSM dataset well, and that it was capable of modulating separate piecewise linear equations for each molecule group. Model performance was assessed by cross validation, randomisation and applicability domain tests, with results indicating promising performance and interpretable outputs. Implementation of a deep learning QSAR method (Transformer-CNN [56]) showed comparative performance in terms of prediction error, but modSAR offered superior explainability by identifying generalised substructures that can guide FBDD (see Figure S8).

Importantly, modelling results were used to develop a screening strategy to identify suitable compounds that could be experimentally tested for antimalarial activity. The result indicated that the positive fragments prioritised from modSAR analysis can constitute a loose structure of a potential candidate to be further selected from virtual screening. The three most suitable compounds which passed the wet-lab tests (A02, A10, and A22) were assigned to modSAR modules m02, m03, and m01 respectively, and corresponded to Series 4 and Series 1 compounds in OSM. When analysed via modSAR, these three



(a) MolPort-047-964-374 (b) MolPort-047-964-639 (c) MolPort-046-842-243

Fig. 7. Visualisation of the top three hits compounds.

% Inhibition at 10μM					
Screening ID	Plate 1	Plate 2	Plate 3	Mean	SD
A01	18.2	9.2	6.2	**11.2**	6.3
A02	48	37.6	38.5	**41.4**	5.7
A03	0.6	-4.2	-1.5	**-1.7**	2.4
A04	12.4	3.5	1.9	**5.9**	5.6
A05	-1.2	-3.2	-6	**-3.5**	2.4
A06	8.4	-4.4	-2.5	**0.5**	6.9
A07	2.2	-3.1	-6.6	**-2.5**	4.4
A08	3	-3.4	-2.6	**-1**	3.5
A09	5.5	3.8	-2.5	**2.3**	4.2
A10	37.2	35.1	29.8	**34**	3.8
A11	0.2	-3.2	-2.8	**-2**	1.9
A12	2.8	-6.9	-0.8	**-1.6**	4.9
A13	-4.1	-4.6	-7.1	**-5.3**	1.6
A14	17.4	11.4	12.4	**13.7**	3.2
A15	6.1	-1.5	13.7	**6.1**	7.6
A16	0.2	-4.7	-3.9	**-2.8**	2.6
A17	4.2	0.3	-0.3	**1.4**	2.4
A18	9.6	-0.3	-0.2	**3**	5.7
A19	0.2	0.3	-0.4	**0**	0.4
A20	7	2.5	-0.4	**3**	3.7
A21	6.8	3.5	-5.5	**1.6**	6.3
A22	21.5	15.6	18.7	**18.6**	2.9

Fig. 8. The percentage inhibition of *P.falciparum* asexual growth for the top-ranking compounds identified from the *in silico* screen using a single concentration of 10 μM performed in triplicate. The amount of inhibition is indicated by a colour scale from green (no inhibition) to white (complete inhibition).

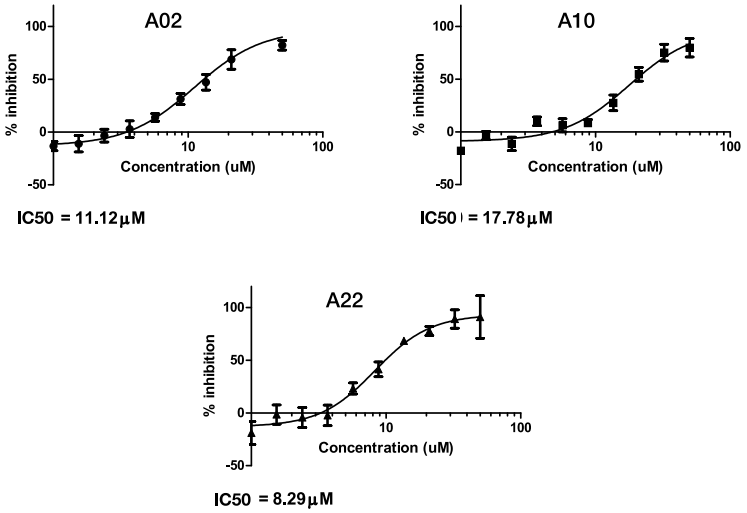


Fig. 9. The percentage inhibition of *P.falciparum* asexual growth based on serial dilutions of the compounds of each of the top three compounds identified by the initial inhibition screen to obtain IC_{50} values.

compounds contained the largest number of positive bits and showed the best (A22) and second best (A02, A10) similarity to the original OSM compounds of the corresponding module. Moreover, A22 also retained the common scaffold that was shared across most of Series 1 compounds, which was considered to possess antimalarial activity [33]. These compounds can be considered as potential lead compounds in the development of antimalarial drug candidates.

In this work our attention focused on demonstrating the ability of modSAR to provide modelling insights and prioritise useful chemical fragments. As our methodological basis comprises modularity clustering and mathematical modelling, modSAR inherits the limitations of the two techniques. Modularity-based community detection suffers from resolution limit [60], which implies that communities smaller than a certain scale cannot be resolved. Moreover, we note that mathematical modelling can potentially suffer in handling very large datasets and may be sensitive to noisy data. Finally, as our aim was to target commercially available compounds for wet-lab experimental validation, future work can be envisaged aiming further towards extending screening libraries, designing and synthesising compounds with the prioritised fragments.

An important aspect of this work lies in the mathematical nature of the modSAR model that offers explainable output, as molecular fingerprint bits selected by each equation can be reverse-engineered to match molecular fragments. This provides valuable and powerful insights into the components that drive activity and can be leveraged to identify potentially active compounds in a different chemical space, driving new lines of drug discovery.

Data and software availability

Code for data preprocessing, feature extraction, modSAR, model evaluation and tuning, as well as result analysis is available through the Github page.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge the open and collaborative research effort of Open Source Malaria without which this study would not have been possible. Yutong Li is supported by the China Scholarship Council. Lazaros G. Papageorgiou acknowledges funding from EPSRC, United Kingdom (EP/V01479X/1, EP/V051008/1). Nicholas Furnham is funded by the Medical Research Council UK (Grant no. MR/T000171/1). Michael J. Delves is supported by a Medical Research Council Career Development Award (MR/V010034/1).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.artmed.2023.102700. The Supporting Information is available free of charge in supporting_information.pdf.

Mathematical details of OPLRAreg, equivalency between modSAR modules and OSM series, average model performance of cross validation, pIC_{50} distribution of compounds related to the important bits, the visualisation and SHAP value of each important bit, Molport ID and molecular structure of prioritised compounds from virtual screening, and the visualisation of feature importance comparing Transformer-CNN and modSAR (supporting_information.pdf). MAIP prediction of the modSAR prioritised compounds (MAIP_results.csv)

References

[1] World Health Organization. World malaria report 2022. World Health Organization; 2022.
[2] Ashley EA, Dhorda M, Fairhurst RM, Amaratunga C, Lim. Spread of artemisinin resistance in plasmodium falciparum malaria. N Engl J Med 2014;371(5):411–23.
[3] Dondorp AM, Smithuis FM, Woodrow C, von Seidlein L. How to contain artemisinin- and multidrug-resistant falciparum malaria. Trends Parasitol 2017;33(5):353–63.
[4] Melville JL, Burke EK, Hirst JD. Machine learning in virtual screening.. Comb Chem High Throughput Screen 2009;12(4):332–43.
[5] Jamal S, Periwal V, Consortium OSDD, Scaria V. Predictive modeling of anti-malarial molecules inhibiting apicoplast formation. BMC Bioinformatics 2013;14(1):55.
[6] Roy K, Kar S, Das RN. A primer on QSAR/QSPR modeling. SpringerBriefs in molecular science, Springer International Publishing; 2015.
[7] Vyas VK, Bhati S, Patel S, Ghate M. Structure-and ligand-based drug design methods for the modeling of antimalarial agents: a review of updates from 2012 onwards. J Biomol Struct Dyn 2021;40(20):10481–506.
[8] Yadav BS, Chaturvedi N, Marina N. Recent advances in system based study for anti-malarial drug development process. Curr Pharm Des 2019;25(31):3367–77.
[9] Rahman F, Lhaksmana KM, Kurniawan I. Implementation of simulated annealing-support vector machine on QSAR study of fusidic acid derivatives as anti-malarial agent. In: 2020 6th international conference on interactive digital media (ICIDM). IEEE; 2020, p. 1–4.
[10] Ambiar MF, Aditsania A, Kurniawan I. QSAR study on falcipain inhibitors as anti-malaria using genetic algorithm-support vector machine. In: 2022 5th international conference of computer and informatics engineering (IC2IE). IEEE; 2022, p. 287–93.
[11] Bharti DR, Lynn AM. QSAR based predictive modeling for anti-malarial molecules. Bioinformation 2017;13(5):154.
[12] Bosc N, Felix E, Arcila R, Mendez D, Saunders MR, Green DV, Ochoada J, Shelat AA, Martin EJ, Iyer P, et al. MAIP: A web service for predicting blood-stage malaria inhibitors. J Cheminform 2021;13(1).
[13] Askr H, Elgeldawi E, Aboul Ella H, Elshaier YA, Gomaa MM, Hassanien AE. Deep learning in drug discovery: an integrative review and future challenges. Artif Intell Rev 2023;56(7):5975–6037.
[14] Combi C, Amico B, Bellazzi R, Holzinger A, Moore JH, Zitnik M, Holmes JH. A manifesto on explainability for artificial intelligence in medicine. Artif Intell Med 2022;133:102423.
[15] Jiménez-Luna J, Grisoni F, Schneider G. Drug discovery with explainable artificial intelligence. Nat Mach Intell 2020;2(10):573–84.
[16] Open Source Malaria. OSM - Main website. 2020, URL: http://opensourcemalaria.org/.
[17] Cardoso-Silva J, Papageorgiou LG, Tsoka S. Network-based piecewise linear regression for QSAR modelling. J Comput Aided Mol Des 2019;33(9):831–44.
[18] Tse EG, Aithani L, Anderson M, Cardoso-Silva J, Cincilla G. An open drug discovery competition: Experimental validation of predictive models in a series of novel antimalarials. J Med Chem 2021;64(22):16450–63.
[19] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model 2010;50(5):742–54.
[20] Cardoso-Silva J, Papadatos G, Papageorgiou LG, Tsoka S. Optimal piecewise linear regression algorithm for QSAR modelling. Mol Inform 2019;38(3):1800028.
[21] Spillman NJ, Allen RJ, McNamara CW, Yeung BK, Winzeler EA, Diagana TT, Kirk K. Na+ regulation in the malaria parasite Plasmodium falciparum involves the cation ATPase PfATP4 and is a target of the spiroindolone antimalarials. Cell Host Microbe 2013;13(2):227–37.
[22] Kirk K. Ion regulation in the malaria parasite. Annu Rev Microbiol 2015;69:341–59.
[23] Lehane AM, Ridgway MC, Baker E, Kirk K. Diverse chemotypes disrupt ion homeostasis in the malaria parasite. Mol Microbiol 2014;94(2):327–39.
[24] Spillman NJ, Allen RJ, McNamara CW, Yeung BK, Winzeler EA, Diagana TT, Kirk K. Na+ regulation in the malaria parasite Plasmodium falciparum involves the cation ATPase PfATP4 and is a target of the spiroindolone antimalarials. Cell Host Microbe 2013;13(2):227–37.
[25] Jiménez-Díaz MB, Ebert D, Salinas Y, Pradhan A, Lehane AM, Myrand-Lapierre M-E, O'Loughlin KG, Shackleford DM, Justino de Almeida M, Carrillo AK, et al. (+)-SJ733, a clinical candidate for malaria that acts through ATP4 to induce rapid host-mediated clearance of Plasmodium. Proc Natl Acad Sci 2014;111(50):E5455–62.
[26] Gamo F-J, Sanz LM, Vidal J, de Cozar C, Alvarez E, Lavandera J-L, Vanderwall DE, Green DVS, Kumar V, Hasan S, Brown JR, Peishoff CE, Cardon LR, Garcia-Bustos JF. Thousands of chemical starting points for antimalarial lead identification. Nature 2010;465(7296):305–10.
[27] Williamson AE, Ylioja PM, Robertson MN, Antonova-Koch Y, Avery V. Open source drug discovery: Highly potent antimalarial compounds derived from the tres cantos arylpyrroles. ACS Cent Sci 2016;2(10):687–701.
[28] OpenSourceMalaria. b. OpenSourceMalaria/series3: Everything to do with Open source malaria series 3. URL: https://github.com/OpenSourceMalaria/Series3.

[29] OpenSourceMalaria. c. OpenSourceMalaria/series4: Repository for series 4 of the open source malaria consortium. URL: https://github.com/OpenSourceMalaria/Series4.
[30] Motion A, Tse EG, Korsik M, Macdonald T, Baum J, Michael J. A potent, in vivo active antimalarial series based on a triazolopyrazine core: Communal lead optimization in an open source malaria project series. 2021.
[31] OSM. Announcing OSM series 4 - the triazolopyrazines. 2013, URL: http://malaria.ourexperiment.org/the_osm_blog/7954/Announcing_OSM_Series_4_the_Triazolopyrazines.html.
[32] MMV. Potential new class of antimalarials now open source. 2013, URL: https://www.mmv.org/newsroom/news/potential-new-class-antimalarials-now-open-source.
[33] Tse EG. Open source malaria : Potent triazolopyrazine-based antiplasmodium agents that probe an important mechanism of action (Ph.D. thesis), University College London; 2019, p. 400.
[34] OpenSourceMalaria. a. Competition round 2: A predictive model for series 4-issue #1. URL: https://github.com/OpenSourceMalaria/Series4_PredictiveModel/issues/1.
[35] RDKit. RDKit: Open-source cheminformatics. 2020, URL: http://www.rdkit.org. accessed: 2020-05-12.
[36] Morgan HL. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. J Chem Doc 1965;5(2):107–13.
[37] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008;2008(10):P10008.
[38] Maggiora G, Vogt M, Stumpfe D, Bajorath J. Molecular similarity in medicinal chemistry. J Med Chem 2014;57(8):3186–204.
[39] Zahoránszky-Kohalmi G, Bologa CG, Oprea TI. Impact of similarity threshold on the topology of molecular similarity networks and clustering outcomes. J Cheminform 2016;8(1):1–17.
[40] Newman ME. Modularity and community structure in networks. Proc Natl Acad Sci USA 2006;103(23):8577–82.
[41] Xu G, Bennett L, Papageorgiou LG, Tsoka S. Module detection in complex networks using integer optimisation. Algorithms Mol Biol 2010;12:5–36.
[42] Yang L, Silva JC, Papageorgiou LG, Tsoka S. Community structure detection for directed networks through modularity optimisation. Algorithms 2016;9(4):73.
[43] Silva JC, Bennett L, Papageorgiou LG, Tsoka S. A mathematical programming approach for sequential clustering of dynamic networks. Eur Phys J B 2016;89(2):39.
[44] Yang L, Liu S, Tsoka S, Papageorgiou LG. Mathematical programming for piecewise linear regression analysis. Expert Syst Appl 2016;44:156–67.
[45] Kruger F, Stiefl N, Landrum GA. rdScaffoldNetwork: The scaffold network implementation in RDKit. J Chem Inf Model 2020;60(7):3331–5.
[46] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems. 2017, p. 4768–77.
[47] Rodríguez-Pérez R, Bajorath J. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. J Comput Aided Mol Des 2020;34(10):1013–26.
[48] Rodríguez-Pérez R, Bajorath J. Chemistry-centric explanation of machine learning models. Artif Intell Life Sci 2021;1:100009.
[49] Nick P. shap_barplot: Visualize shap values of top features by magnitude and direction. URL: https://github.com/nick-phillips/shap_barplot.
[50] Gramatica P. Principles of QSAR models validation: internal and external. QSAR Comb Sci 2007;26(5):694–701.
[51] Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inform 2010;29(6–7):476–88.
[52] Rucker C, Rucker G, Meringer M. y-Randomization and its variants in QSPR/QSAR. J Chem Inf Model 2007;47(6):2345–57.
[53] Muegge I, Mukherjee P. An overview of molecular fingerprint similarity search in virtual screening. Expert Opin Drug Discov 2016;11(2):137–48.
[54] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 2020;58:82–115.
[55] Jaganathan K, Tayara H, Chong KT. An explainable supervised machine learning model for predicting respiratory toxicity of chemicals using optimal molecular descriptors. Pharmaceutics 2022;14(4):832.
[56] Karpov P, Godin G, Tetko IV. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. J Cheminform 2020;12(1):1–12.
[57] Neves BJ, Braga RC, Alves VM, Lima MN, Cassiano GC, Muratov EN, Costa FT, Andrade CH. Deep Learning-driven research for drug discovery: Tackling Malaria. PLoS Comput Biol 2020;16(2):e1007025.
[58] Zhang L, Fourches D, Sedykh A, Zhu H, Golbraikh A, Ekins S, Clark J, Connelly MC, Sigal M, Hodges D, et al. Discovery of novel antimalarial compounds enabled by QSAR-based virtual screening. J Chem Inf Model 2013;53(2):475–92.
[59] Caballero-Alfonso AY, Cruz-Monteagudo M, Tejera E, Benfenati E, Borges F, Cordeiro M, Armijos-Jaramillo V, Perez-Castillo Y. Ensemble-based modeling of chemical compounds with antimalarial activity. Curr Top Med Chem 2019;19(11):957–69.
[60] Fortunato S, Barthelemy M. Resolution limit in community detection. Proc Natl Acad Sci 2007;104(1):36–41.