UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

Machines against malaria: Artificial intelligence classification models to advance antimalarial drug discovery

by

Ashleigh van Heerden

Supervisor: Prof. Lyn-Marié Birkholtz
Co-supervisor: Prof. Nelishia Pillay

Submitted in the partial fulfilment of the requirements for the degree

Philosophiae Doctor
(Specialisation in Biochemistry)

In the Faculty of Natural and Agricultural Sciences
Department of Biochemistry, Genetics and Microbiology
University of Pretoria
Pretoria
South Africa

December 2023

I, **Ashleigh van Heerden**, declare that the thesis, which I hereby submit for the degree ***Philosophiae Doctor*** in the Department of Biochemistry, at the University of Pretoria, is my own work and has not previously been submitted by me for the degree at this or any other tertiary institution.

SIGNATURE:..... 

DATE:..2023/12/07

PLAGIARISM DECLARATION

University of Pretoria

Faculty of Natural and Agricultural Science

Department of Biochemistry

Full names of student: Ashleigh van Heerden

Student number: 14020590

Title of work: Machines against malaria: Artificial intelligence classification models to advance antimalarial drug discovery

Declaration

1. I understand what plagiarism is and I am aware of the University‟s policy in this regard.
2. I declare that this thesis is my own original work. Where other people‟s work has been used (either from a printed source, internet or any other source), due acknowledgement was given and reference was made according to departmental requirements.
3. I did not make use of another student's previous work and submit it as my own.
4. I did not allow and will not allow anyone to copy my work with the intention of presenting it as his or her own work.

SIGNATURE STUDENT:.....  .....

DATE:..2023/12/07

ACKNOWLEDGEMENTS

I would like to thank the following people for their aid and contribution to this research project:

Jesus Christ, my Lord and Saviour, that has helped me keep my sanity throughout my studies and peace amidst the storms of life as well as answering my prayers when my models were not working!

My family who have provided much comfort and support throughout my studies. My cousin, Claudia Lezzi, for being the best friend, cousin and sister in heart that anyone can ask for. My mother, Verna, and my father, Johan, who has been encouraging me throughout my studies.

Professor Lyn-Marie Birkholtz, for her amazing insight and guidance throughout all these years under her supervision, you truly are an inspiration and a fantastic mentor.

Dr. Gemma Turon, Dr. Miquel Duran-Frigola and Professor Nelisha Pillay, whose expertise and input has been instrumental to this project.

I would also like to thank the National Research foundation and Professor Lyn-Marie Birkholtz, whose funding has made it possible for me to focus and complete on my PhD without distraction.

SUMMARY

Multiple hurdles exist within the antimalarial drug discovery pipeline, including the time, and resources required to identify dual-active transmission-blocking drugs to further malaria elimination. Vast amounts of phenotypic screening data have been generated for antimalarial drug discovery, and yet this information regarding the chemical space for activity is underutilised and can be a valuable resource when explored with machine learning (ML) to improve the efficiency and cost associated with antimalarial drug discovery. Using the efficacy data from diverse chemical libraries, screened against (ABS) parasites and transmissible gametocyte stages of the malaria parasite *Plasmodium falciparum*, we successfully built robust ML models capable of predicting compounds with singular activity against ABS parasites and dual-activity against both life cycle stages. Support Vector Machines (SVM) specifically showed superior abilities with high recall (90% and 66% against the ABS-specific and dual-activity models, respectively) and low false positive predictions (15% and 1%, respectively). The predictive capabilities of the models held in novel diverse chemical spaces, indicating that the ML models are robust and can serve as a prioritisation tool to drive and guide phenotypic screening and medicinal chemistry programs. Feature importance analysis upon these models further allowed the identification of chemical features enriched within active and inactive compounds, an important outcome that could be mined for essential chemical features to streamline hit-to-lead optimisation strategies of antimalarial candidates. Additionally, this information could provide chemical features associated with inactive compounds, eliminating any future unnecessary screening of similar chemical analogues.

This doctoral study developed robust models that can predict compound ABS and/or dual-activity, which can serve as valuable pre-screening tools to prioritize the screening of compounds most likely to show activity against the parasite. Furthermore, we highlighted features that are predictive and enriched within active/inactive compounds against ABS parasites and/or transmissible gametocyte stages and can serve as a guide for derivatisation strategies to increase potency and stage-specific activity of novel antimalarial candidates.

Table of Contents

Table of Contents

List of Tables

Chapter 1

Chapter 2

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

ABBREVIATIONS

ABS	Asexual blood stage(s)
ACT	Artemisinin combination therapy
ANN	Artificial neural network
CNN	Convolutional neural network
DNN	Deep neural network
ECFP	Extended-Connectivity Fingerprints
EF	Enrichment factor
FPR	False positive rate
GBM	Gradient boosting model
G-mean	Geometric mean
IDC	Intraerythrocytic development cycle
KNN	K nearest neighbours
ML	ML
MLR	Multinomial logistic regression
MMD	Maximum mean discrepancy
MMV	Medicines for malaria venture
MoA	Mode of action
NB	Naïve Bayes
PC	Principal component
PCA	Principal component analysis
RBF	Radial basis function
RF	Random forest
RL	Reinforcement learning
RNN	Recurrent neural network
ROC-AUC	Receiving operator curve area under curve
SVM	Support vector machine
TL	Transfer learning
VAE	Variational autoencoder
UMAP	Uniform Manifold Approximation and Projection
WHO	World Health Organization

CHAPTER 1

LITERATURE REVIEW, BACKGROUND AND MOTIVATION

1.1) Impact of technology on biological knowledge

Technological advancements have created an explosion of biological knowledge driven by cost-effective and efficient technical capabilities. In the field of genomics, this is especially true since, in the past, the monitoring of gene expression was restricted to a few genes compared to modern-day technology, where thousands of genes over multiple conditions can be monitored simultaneously [1]. The advancement of technology has had a significant impact in increasing the density and multivariate nature of data we are able to obtain in experimental settings. Not only this, but technological advancements have also also increased the efficiency of experimental assays, especially regarding phenotypic screening. Phenotypic screening is used to discover new small molecules that have a lethal effect on a particular cell type and can be explored as drug candidates. This approach has also benefited from, for example, the application of robotics that has allowed the automation of assays and enabled such screening to become more high-throughput and accelerate hit compound identification [2]. This rapid accumulation of biological information, where numerous variables are monitored concurrently, has, however, resulted in a massive increase in the dimensionality of data that can be rapidly generated. This presents scientists with the big data conundrum - finding hidden structures or patterns within such high-dimensional data. Traditional statistical methods become less insightful as the number of variables increases more than the number of observations within the data [3], limiting these approaches' applicability.

Fortunately, the development of more advanced computational tools spurred the use of machine learning within various industries. Machine learning (ML), a subset of artificial intelligence, employs algorithms to discover hidden, non-obvious patterns within data and, based on the learning of these patterns, produce reliable statistical predictions [4]. Where traditional statistical methods draw population inferences from a sample/population, ML, by contrast, tries to identify generalisable patterns within the data that would enable correct predictions on similar data. Though statistical inference allows a better understanding of how a system behaves based on observations, ML tries to predict new data based on

patterns learned from similar observations rather than understanding the underlying mechanisms [5]. Often, the numerous data features (high dimensionality) and complexity within biological systems make identifying relevant features difficult for statistical inference. As an alternative, ML is a more attractive approach [6]. Due to this, ML has become a powerful tool in drug discovery, expediting decision-making, optimising efficiency, reducing wasteful expenditure, and increasing the success rate of candidates progressing through the pipeline and clinical trials [7]. Various types of ML approaches are used in drug discovery, but most of these can be summarised within four categories, namely, supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning (Figure 1.1) and these categories will be covered in the next section. Examples of standard ML algorithms commonly used for model building are given in grey (Figure 1.1).

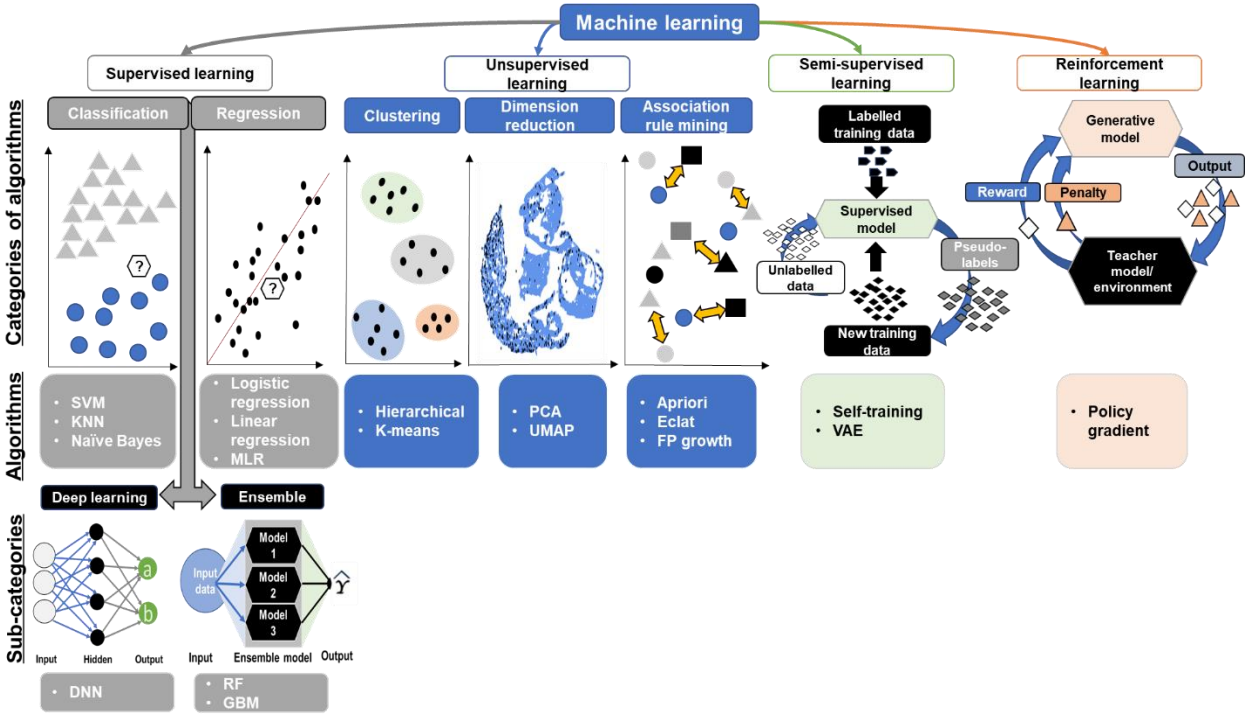


Figure 1.1: Overview of the types of machine learning.
Supervised, unsupervised, semi-supervised and reinforcement learning are the different types of ML.

1.1.1) Supervised learning

Supervised learning is a type of ML where a model is trained on a labelled dataset, i.e., a dataset where inputs and corresponding correct outputs are provided. During training, the model's algorithm aims to learn how to accurately map the inputs to the correct outputs to predict the output for new, unseen inputs. Supervised learning can be further grouped into classification or regression [8], whereas deep learning and ensemble models can be

considered a subcategory within the algorithms as they employ more intuitive approaches for model building than standard machine learning algorithms. Classification models are commonly used to predict the categorical class of new data, whereas regression models try to predict a numerical value. Depending on the type of output to be predicted, supervised ML can build either regression models where a continuous, usually numerical target variable is predicted or classification models where a categorical target variable is predicted. Regression ML models are often used in drug discovery to predict the potency [9], pharmacodynamic [10, 11], pharmacokinetic [12, 13], and physiochemical properties [14] of compounds. Other than predicting compound activity (inactive/active), classification models have been used in drug discovery to predict drug-target interactions [15, 16], drug mode of action [17], as well as the probability of a compound possessing properties that may pose a safety risk such as blood-brain barrier penetration [18] or cardiotoxicity [19].Classification models, rather than a numerical variable, predict a class the user provides and can be more malleable due to the flexibility within user-labelling. Due to this flexibility, classification models are more expansive than regression models and can be applied to a broader range of scientific classification problems, assuming the training data is curated and of good quality and applies to the classification problem. This is known as the "garbage-in, garbage-out" principle and applies to most ML algorithms, where a model's prediction quality entirely depends on the quality of the data the model is trained on. Data that inherently contain noise or irrelevant/redundant features can result in poor model performance [20].

The most frequently used supervised ML techniques used in drug discovery include classification models (e.g. support vector machines [SVM]; K-nearest neighbours [KNN]; and Naïve Bayes models) and regression models (e.g. linear and logistic linear regression and Multiple linear regression [MLR]). However, ensemble methods (e.g. random forest [RF] and gradient boosting machines [GBM]), where several weaker models are combined for a more powerful output or deep learning methods (e.g. deep neural networks [DNN]) are powerful additions to identify patterns within large datasets (Figure 1.1).

1.1.1.1. Classical machine learning algorithms used in drug discovery

During the training process for SVM classification models, multiple hyperplanes are generated and evaluated in their ability to separate classes in a high-dimensional feature space. From these multiple hyperplanes, the optimal hyperplane that produces the maximum margin, i.e. distance from observations of each class, and that can robustly

separate two different classes is selected [21]. Within SVM regression models, the concept remains the same, however, the optimal hyperplane that best maximises the distance between the closest data points in a continuous space while simultaneously reducing the error in predictions of the target variable is selected [22].

SVM was initially created for binary prediction and can be used for either regression or classification. Over recent years, SVM has been adapted to multiclassification problems, e.g., a 'one-against-one' approach can be taken where several binary classifiers are combined [23, 24], and the final class is identified by a majority vote of the combined classifiers. To make SVM also more applicable to non-linear data, data transformation techniques, a.k.a. "kernel trick", are applied to the data to create a non-linear feature space where the observations of the two classes can then be more clearly separated by a hyperplane [24, 25]. Although data transformation techniques are powerful enough to reveal hidden patterns within data, they also make it difficult to calculate a feature's actual weight in predictions as it undergoes such transformation. Nevertheless, SVM has been used for multiple purposes within drug discovery, including prediction of physiochemical properties [26], drug-drug interaction [27] and bioactivity.

The Naïve Bayes algorithm is a probabilistic classifier that employs Bayes' theorem to model input data distribution for a given class. The Bayes theorem is a statistical method for calculating the conditional probabilities, i.e., the likelihood of an outcome occurring based on previous data of when such outcome occurred [28]. Naïve Bayes models assume that all features of the input data are conditionally independent and contribute equally to the class [29, 30].

Within linear regression models, a linear relationship is assumed between the input and output variables, and a generalised linear model is built to describe this linear relation [31]. In multiple linear regression (MLR) models, however, the linear regression algorithm is adapted to allow the prediction of more than two output variables and is also more amiable for the modelling of non-linear data. MLR models make multiple assumptions of the training data, including that the input and output variable is linear, that all variables show multivariate normality, observations are assumed to be independent of one another and independent variables are assumed to show no high correlation with one another. Despite its simplicity, MLR has been used in drug discovery to predict compound activity, toxicity [32], solubility

[33] and anticancer drug efficacy [34], however, the assumptions made on the training data limit the use of MLR algorithm within other data settings.

Within K-nearest neighbours' models, it is assumed that similar data will have similar labels or values; hence, during the training of such models, the training data is used as a reference set. When exposed to new data, the models will calculate the distance between the new data and the training data, using e.g. Euclidian distance as a distance metric. From the distances calculated between the training data and the input, the model will assign the most frequent label/value to the input from among the K neighbours as the predicted outcome. KNN has been used in compound activity and metabolic stability prediction [35]. One disadvantage of KNN, however, is that there is no standard by which to select the number of nearest neighbours (K value) and selecting a too high/low value can result in higher false positive or negative rates [36].

1.1.1.2. Ensemble machine learning algorithms in drug discovery

Ensemble modelling is a ML technique where several weak models are combined to generate an optimal model that would yield better prediction performance than a single model. Examples of such techniques used in drug discovery include random forest (RF) and gradient boosting machines (GBM). RF combines several decision trees as base models to generate an optimal model yielding better prediction performance than a single decision tree [37-39]. These decision trees are built via a ‘bagging’ approach (Figure 1.2), where the trees are built on different randomly selected training samples from the training data. This bagging approach is beneficial to reduce overfitting and improve model performance when training on imbalanced data when the training data has skewed class proportions [40]. Within classification models, the majority vote of the base models is used to predict the correct class, while in regression models, the final prediction is the mean of all the base model outputs. Like RF, GBM is also an ensemble ML method combining several base models; however, the algorithm differs using a ‘boosting’ approach (Figure 1.2). During boosting, numerous base models are built over multiple iterations; during the first iteration, the base model performs weakly in predicting samples, and the algorithm then adds weight to samples that contributed to the error prediction. In the next iteration, using these weighted samples, a new base model is created with better performance than the previous base model, and this process is repeated over multiple iterations to produce an ensemble model that combines all base models to have better predictive performance and lower prediction

error. Such ensemble models have been used in drug discovery to predict bioactivity [41], drug sensitivity [42], cardiotoxicity[19] and druggable protein targets [43] .

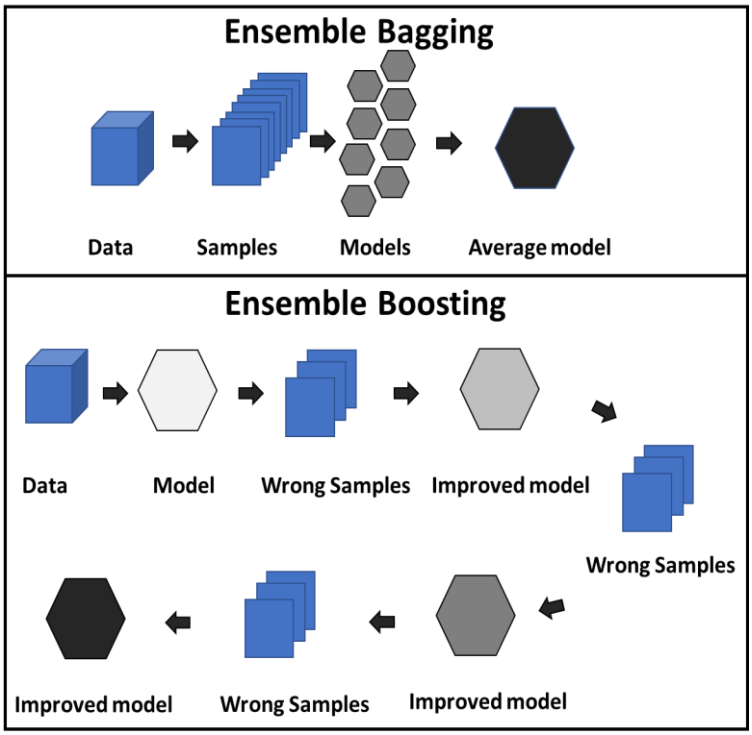


Figure 1.2: Ensemble modelling methods.
Ensemble bagging is where individual models are made from bootstrap aggregating. Each model trains on a subset of the training data (samples) resulting in multiple models being built each different from one another. When making a prediction the output will be the majority vote of all the models' predictions. Ensemble boosting is where a model is trained on the training data, however, the samples incorrectly predicted are used together with the old model to train a new model performs better in the prediction of samples compared to the old model. This is repeated until the number of iterations is reached or no more samples are incorrectly predicted.

1.1.1.3. Deep learning in drug discovery

Deep neural networks (DNN) is a subsection of ML that has gained much popularity due to its remarkable ability to process information from complex systems prone to nonlinearity, noise, and high parallelism and produce good generalisation [44]. Due to these characteristics, DNN have been used extensively in drug discovery, including predicting bioactivity, drug resistance [45], ADMET properties [46] and drug repurposing [47]. The basic layout of a DNN (Figure 1.3) typically consists of an input layer where data is fed into the network, followed by multiple hidden layers that transform the data and extract important features for prediction and finally, the output layer where a prediction is made. Within each of the layers in the DNNs are nodes that each function as processing units and can be wholly interconnected or partially connected between neighbouring layers [48]. The input layer contains nodes that represent the input variables of the model, and the data of these variables can be transformed using an activation function to highlight patterns within the data better as it passes through to the hidden nodes. As information is introduced through

the input layer and fed towards the hidden layers, higher-level features from the data are gradually extracted using the nodes within these layers, which function as multi-layered processing units. Ultimately, the processed data is fed into the output layer, where the number of output nodes corresponds to the number of classes or prediction values that are calculated for classification or regression prediction [49]. Additionally, backpropagation is applied during model training, where the weight between nodes in different layers is adjusted to minimise the error/loss in predictions.

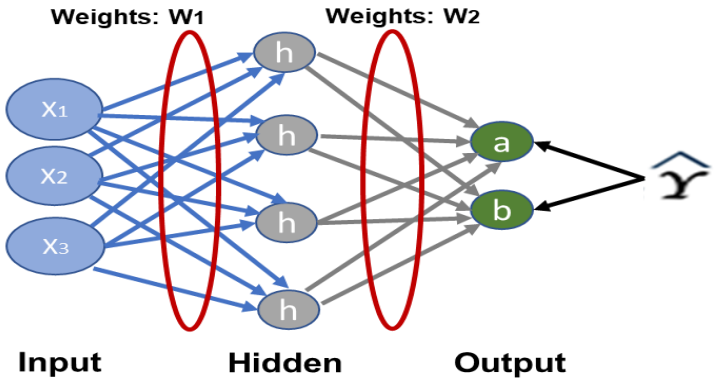


Figure 1.3: Typical artificial neural network layout.
Neural networks have input nodes where data are fed into an input layer where hidden nodes can assess information from the input nodes. This hidden layer can be extended to multiple layers to create a deep neural network and the hidden nodes (processing units) can also be increased. The last hidden layer then connects to output nodes which can be increased to the number of classes or events to be predicted. The hidden nodes give to each output node/class a probability of being true based on the input information fed into the input layer.

1.1.2) Unsupervised learning

In contrast to supervised ML, unsupervised ML does not require labelled data or prior knowledge of the data and is often used to detect and highlight non-obvious patterns within the data. Unsupervised ML can be divided into three groups based on their use: dimensionality reduction, clustering, and association rule mining. Dimensionality reduction is often used to project high-dimensional data into a lower-dimensional feature space while maintaining the meaningful properties and structure inherent within the data [50]. When using supervised learning on high-dimensional data, as the number of features within the data increases, this can impact the model's performance due to the potential signal within the data being diluted as the features increase [51]. In such instances, dimensionality reduction is often performed before model training to lower the training data's dimensionality whilst keeping the data's inherent structure and properties for training. Principal component analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) are examples of dimensionality reduction techniques.

PCA reduces the dimensionality of large datasets by finding orthogonal axes, i.e., principal components (PC), in K-dimensions that best captures the maximum variances within the high-dimensional data. These PCs are linear combinations of the original variables in the data, and when all these PCs are added together, they are equal to the total variance observed in the dataset. However, to reduce the dimensionality of the dataset, the first and second PCs, which explain most of the variances within the dataset, are selected to represent the dataset [50].

UMAP, on the other hand, tries to learn the manifold structure within data and to identify a low-dimensional embedding that closely mimics the topological structure of the manifold of the original data. UMAP learns the manifold structure of data by constructing a high-dimensional weighted graph (Figure 1.4A), with edge weights representing the likelihood that two points are connected. Connectedness between points is then determined by extending the radius of points, and two points are connected when the radii of two points overlap within the graph. The graph is then made "fuzzy" by reducing the likelihood of a connection between points in the graph as the radius grows (Figure 1.4B). Lastly, each point is then connected to its closest neighbour to preserve the local data structure through these connections, whilst the global structure is maintained through the lower dimensional graph embedding.

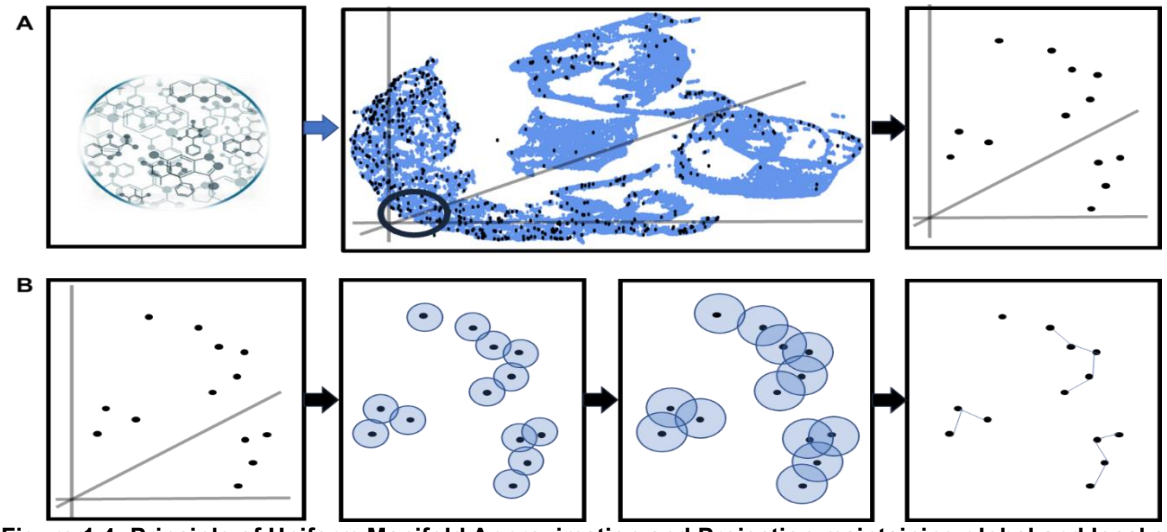


Figure 1.4: Principle of Uniform Manifold Approximation and Projection maintaining global and local structure of chemical structural data.
(A) UMAP constructs a high dimensional topological representation of the chemical structural data and then tries to create a low dimensional space of the data to capture the global structure inherent within the data. (B) To capture the local structure within the data points' radii are increased and connectedness between points is determined when the radii of points overlap, however, the likelihood of a connection between points is reduced as the radius of points grows. From this inference, each point is connected to its closest neighbour to encompass the local structure within the chemical structural data.

Clustering analysis is often used to group data where data points with similar properties fall within the same cluster than other clusters which differ in such properties. Due to the data not being needed to be labelled or prior knowledge required, clustering is a particularly useful tool to highlight inherent structure and groupings within data that may not be as apparent through other analysis methods. Clustering techniques have been used in drug discovery to aid molecular scaffold analysis to identify important core structures pertaining to activity [52] as well as biomarker gene identification to describe a compound's mode of action (MoA) [53].

Association rule learning is an ML technique where a rule-based approach is used to identify meaningful relationships between features within a dataset. Examples of such ML algorithms are the Apriori, Eclat, and FP growth algorithms, which calculate the support, confidence, and lift of variables within the dataset. Support refers to the frequency a feature appears within the data, and confidence refers to how often a rule has been observed to be true, which can be something as simple as feature X co-occurring with feature Z. Lift, however, determines the strength of any rule discovered within the data [54]. Such ML models have been used in cancer to predict drug response based on the various omics databases available [55].

1.1.3) Semi-supervised learning

In cases where an abundance of unlabelled data is available and a small amount of labelled data is available, which is tedious and expensive to acquire, a semi-supervised ML approach is often used, incorporating both supervised and unsupervised ML techniques. Within semi-supervised ML, the labelled data is used to train a primary model for classification/regression. The unlabelled data is used as a test set where the model must predict their pseudo-labels. Predicted pseudo-labels with high confidence levels are then included in the labelled training set, and the model training is reinitiated and tested on the unlabelled data to predict pseudo-labels. This process is repeated to reduce errors and improve the accuracy of the model [56] and has been used to predict drug function [57].

Within a generative modelling setting, another form of semi-supervised ML, the aim is not necessarily to predict a property of the training data or generate pseudo labels, but rather, based on the training examples it observes, to generate similar examples that are novel yet

plausible. An example is when a model trained on facial features has to produce a fake face that is plausible yet not similar to what the model had been trained on [58].

Variational autoencoders (VAE) is an example of such generative models that use an unsupervised way to learn and model the distribution of the training data, which it embeds in a lower dimensional space, referred to as a latent space. VAE aims to learn the probability distribution of the input data rather than just learning a lower representation of the input data [59]. A VAE consists of two neural networks; one network, termed the encoder, receives input data, and through the layers, important patterns and features are extracted and embedded within the latent space. From this latent space, the task of the following neural network, the decoder, is to 'decode' the latent space and thus reconstruct the original input that entered the encoder as closely as possible.

1.1.4) Reinforcement learning

Reinforcement learning aims to train models through a penalty and reward system as the model learns from its' environment through trial-and-error, where desired outcomes are rewarded, and undesirable/incorrect predictions are heavily penalised. This type of ML has gained popularity and has been used to guide drug discovery for *de novo* drug design [60-63]. Within such reinforcement learning models, a policy gradient approach is usually implemented, which ultimately optimises the model output to be directed towards what would achieve the highest rewards. From the ReLeaSE model generated by Mariya Popova *et al.* (2018) [61], for instance, they used two deep neural networks within a reinforcement learning framework, where one DNN is predictive of a particular molecular property (Janus protein kinase 2 inhibition) while the other DNN is a generative recurrent neural network which will be trained via reward/penalty to generate novel chemical libraries targeted towards the desired molecular property (Figure 1.5). The generative model can be considered as a 'student' that generates an output, while the predictive model functions as a 'teacher' that evaluates whether the generated output of the 'student' is correct and updates the policy on whether a reward/penalty should be applied to the generative model. This is repeated until the generative model begins to optimize its' rewards, where outputs with a high probability of having a desired property is continuously generated. Similarly, within ReLeaSE the generated SMILES from the recurrent neural network (RNN) would be assessed by the predictive model; if the generated SMILES are predicted to be inactive, then a penalty would be applied to the RNN, whereas a reward would be given in the

opposite scenario. RNNs have also been used in drug discovery to generate novel chemical structures with good biological inhibition activity [64, 65].

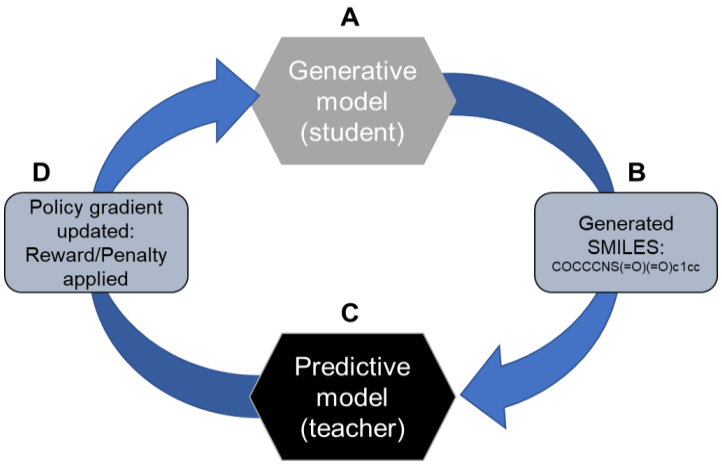


Figure 1.5: Example of reinforcement learning for drug design.
(A) The generative model serves as a student and generates SMILES based on past observations and predictions. (B) The generated SMILES are then fed into the (C) predictive model (which serves as a teacher instructing the student) which will then predict the probability of the generated SMILES having the desired property (for example activity against kinases). (D) If the predicted property is undesirable (inactive) the policy gradient is updated to apply a penalty to the generative model. If the predicted property is the desired outcome, i.e., activity, the policy is updated to apply a reward to the generative model. This is repeated until the generative model begins to optimize its' rewards, where SMILES with a high probability of having a desired property (activity) is continuously generated.

RNNs are a form of DNNs adapted to remember states and inputs from previous instances by changing the weight applied to the current state [66]. With this architecture (Figure 1.6), RNNs perform well with sequence and time-series problems such as sentence construction, weather forecasting, and music generation and can even capture the syntax of molecular representation of chemical SMILES [67, 68]. Despite this, the chemical structures generated from RNNs have low diversity and do not encompass the entire chemical space that hit compounds occupy [69]. Due to this, RNNs have been coupled with RL several studies to allow the RNN algorithm to explore chemical space and broaden the diversity of chemical structures generated. Such models are instrumental in producing targeted libraries towards specific protein targets, however, they may be restricted by the chemical space of the training data for the predictive model.

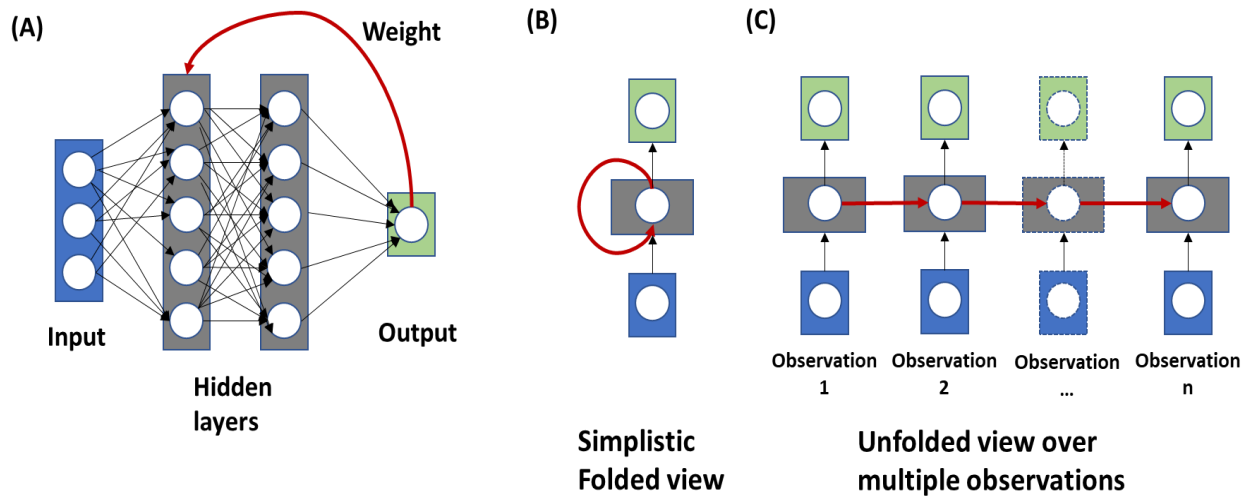


Figure 1.6: Recurrent neural network architecture that allows memory of previous inputs and outputs. (A) RNNs are a modified version of DNNs, where an additional weight is added to the hidden layers which is linked to the output of the RNN. (B) This weight is cyclic and is updated after each observation and prediction of the RNN. (C) Demonstrates how the weight allows the current RNN to remember past observations and predictions.

1.2) ML applications in infectious diseases

The most popular use of ML in infectious diseases was during the COVID-19 pandemic, where ML aided with case forecasting [70], drug repurposing [71, 72], *de novo* drug design [73], case severity prediction [74, 75] as well as optimisation of intensive care resource management [76]. Although COVID-19 provided a platform to display the various uses of ML within drug discovery and case management, ML has been previously used for other neglected diseases. For instance, the diagnosis of Leishmaniasis, a disease caused by a protozoan, relies primarily on microscopy, which can be prone to human error and restrict the efficiency of diagnosis. Through ML, a GBM model trained on microscopic images of macrophages infected with the *Leishmania* parasite achieved good recall and precision in detecting the parasite and was far more cost-effective and faster than individual diagnostic screening [77]. ML has also been used against the Chagas disease, caused by the protozoan *Trypanosoma cruzi*, highlighting electrocardiogram parameters that serve as important markers for acute and chronic case diagnosis through feature importance analysis of their best models [78]. Besides disease diagnosis, ML has also been used in schistosomiasis drug discovery by aiding target validation in predicting the essentiality of proteins within *Schistosoma mansoni* [79].

Due to the limited financial resources available within these neglected infectious diseases, ML has been very valuable in providing models that serve as pre-screening tools to identify

and prioritise compounds with activity against pathogenic organisms (Table 1.1). Such tools are invaluable in reducing wasteful time and cost expenditure during phenotypic screening and accelerating the identification of promising drug candidates. Some of these models try to model the chemical space for inhibition towards specific protein targets, whereas others opt to model whole-cell inhibition activity, however, each model has advantages and limitations as summarised in Table 1.1.

Table 1.1: Overview of current ML models used for pre-screening against parasitic diseases

Model	Model name	Disease agent	Target on which training data is based on	Constraints	Advantages	Ref
Naïve Bayes	MAIP	*P. falciparum* (ABS)	Phenotypic screening data/Whole-cell inhibition activity	• Compound activity restrictive to the biology of the parasite stage on which phenotypic screening was conducted. • Chemical space of specific targets may not be well defined within whole-cell models. • Active compound MoA unknown. • A larger training dataset is required.	• Chemical space of multiple targets can be captured. • Due to whole-cell chemical space captured, predicted actives may still be active towards resistant strains. • Compound activity relating to transport into cell captured. • Multiple compounds with different and novel MoA can be identified.	[80]
Random Forest	NA	*P. falciparum* (ABS and Liver stages)				[81]
Random forest and DNN	DeepMalaria	*P. falciparum* (ABS)				[82]
ANN & KPLS	NA	*Trypanosoma cruzi*				[83]
Bayesian	NA	*Schistosoma mansoni*				[84]
DNN	NA	*P. falciparum*	*P. falciparum* Ion pump (*Pf*ATP4)	• Compound activity restricted to one/few targets. • Compound activity may fail/decrease due to failed transport/permeability of compound into a cell or active site. • The chemical space of the target may change due to gene mutations to target protein in response to drugs and hence resistant strains may be outside the scope of the model. • Compound MoA is restricted	• More defined and finite chemical space for activity. • MoA of active compounds is known • A smaller set of training data is required	[85]
Naïve Bayes and Ensemble models	NA	*P. falciparum*	*P. falciparum* enoyl acyl carrier protein reductase (*Pf*ENR)			[86]
Random Forest	NA	*Leishmania*	*L. mexicana* pyruvate kinase enzyme (*Lm*PK)			[87]

1.3) ML application in antimalarial drug discovery

ML has only recently been applied in antimalarial drug discovery to improve diagnosis, case management and hit identification [86, 88-93], with the most recent ML application in hit compound identification aimed at whole-cell activity prediction against various stages of the main causative agent of the disease, the protozoan parasite, *Plasmodium falciparum* [80, 82] (Table 1.1). This included Naïve Bayes and random forest classification models to predict against asexual and liver stages, respectively (Table 1.1). Additionally, ensemble and deep learning techniques were applied either to allow stage-specific classifications, or to indicate drug target specific probabilities to allow rapid evaluation of compounds with specific activity against frontrunner antimalarial drug targets (Table 1.1). ML has also been used to determine drug partners for synergistic combination therapy [94]. However, ML is still underutilised within antimalarial drug discovery compared to fields like cancer and remains a valuable tool in instances where large datasets are limited, and the parasite's complex biology constrains drug discovery.

Currently, a challenge in antimalarial drug discovery is the lack of compounds that target the sexual forms of the *P. falciparum* parasite due to the inherent restriction of the parasite's biology in these stages and drastically reduced hit rate in phenotypic screening for compounds able to target these stages. Considering the parasite's life cycle (Figure 1.7), the parasite is remarkable in its ability to infect human erythrocytes and undergo massive population expansion during asexual replication within 48 h cycles, resulting in characteristic malaria related fever peaks, severe symptoms and death in an undiagnosed and/or untreated individual.

Malaria infection starts when sporozoites are released into the human bloodstream from an infected female mosquito's salivary glands during feeding. Sporozoites within the human bloodstream migrate to the liver and invade hepatocytes to form hepatic schizonts through asexual exoerythrocytic schizogony [95]. The hepatic schizonts rupture, releasing merozoites into the bloodstream that invade erythrocytes and form the ring stage [95, 96]. The ring stage develops into a trophozoite and then a schizont during asexual multiplication within the infected erythrocyte. Schizonts rupture, releasing merozoites that again infect erythrocytes and this replication cycle repeats every 48 h. Of these asexual bloods stage (ABS) parasites, only a tiny percentage (~10%) commit to sexual development in a process

known as gametocytogenesis [96]. The formation of stage V mature gametocytes, which can take up to 12 days in *P. falciparum*, is the final product of gametocytogenesis and is essential for transmission of the parasite from the human host to a female *Anopheles* mosquito, thereby ensuring survival of the parasites and continuing the spread of the disease. Mature gametocytes within the mosquito midgut develop into micro- and macrogametes. Fusion of the micro- and macrogametes results in a mobile ookinete [97-99] that can penetrate the midgut wall and develop into an oocyst. Within the oocyst, the parasite undergoes asexual replication causing the oocyst to rupture and release sporozoites that migrate to the mosquito's salivary gland.

Most antimalarials are geared towards targeting the ABS parasites to alleviate malaria symptoms and prevent mortality. Only a few candidates in the drug discovery pipeline also target the gametocytes and can therefore be used to prevent the spread of the disease in malaria control and elimination strategies as transmission-blocking antimalarials. Considering observations within the field of increased gametocytaemia after drug treatment [100], commitment to sexual development may serve as an escape strategy of the parasite from drug pressure [101]. Such parasites that escape drug treatment may also contribute to resistance transmission, hence, not only do these transmissible stages function as a reservoir for the spread of the disease, but they may also threaten current antimalarial efficacy. Considering the small percentage of ABS parasites that mature into these transmissible stage V parasites, this biological bottleneck provides a strategically important target for antimalarial drug discovery [102, 103]. To eliminate malaria and make progress towards the disease's eradication, the identification of antimalarials that target these stages is of utmost importance.

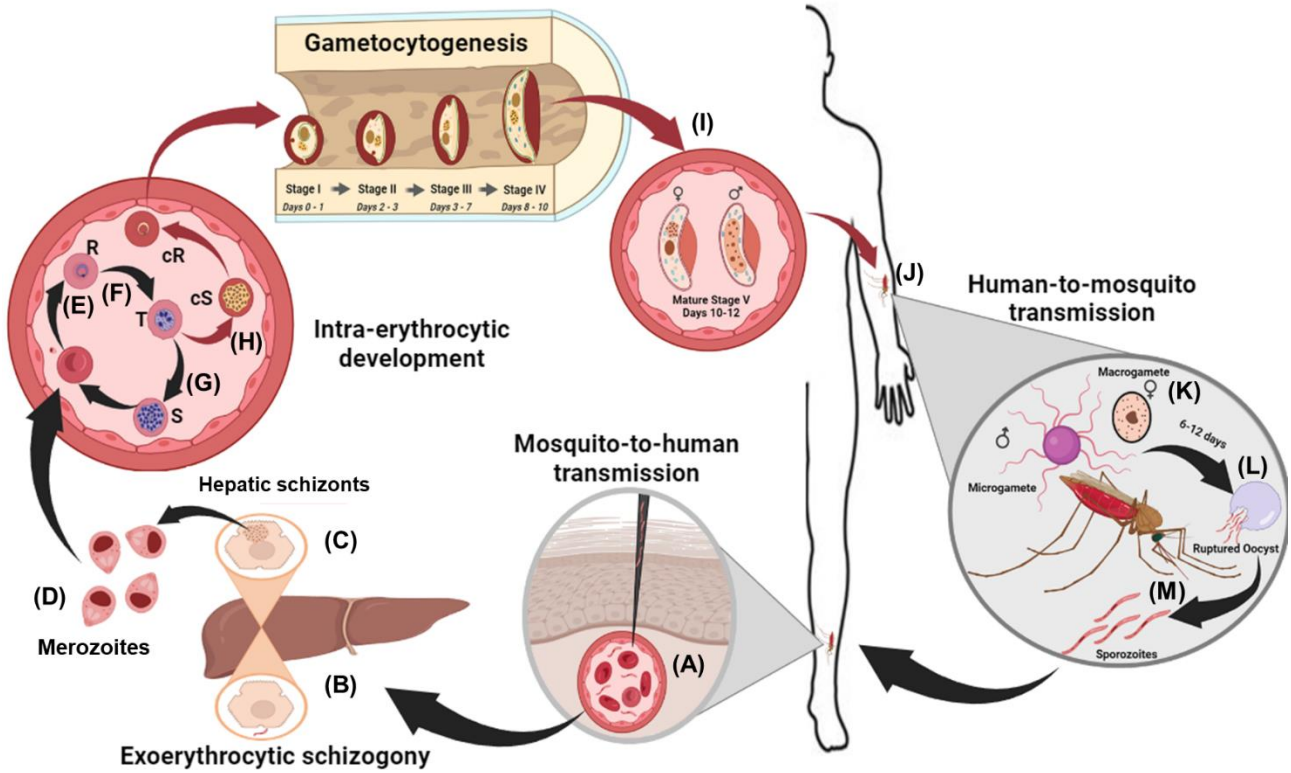


Figure 1.7: Life cycle of *Plasmodium falciparum.*
(A) Sporozoite transmission from mosquitos to humans to initiate liver stage development (B) to form (C) hepatic schizonts and release (D) merozoites. Asexual blood stage (E) ring, (F) trophozoite and (G) schizont stages. (H) Invading merozoites that commit sexually to produce (I) stage V mature male and female gametocytes for (J) transmission for mosquito development as (K) micro- and macrogametes, (L) oocyst, and (M) sporozoites. Created with BioRender.com.

1.4) ML challenges associated with antimalarial drug discovery data

Due to the small portion of parasites that commit to sexual development and the long period to form these transmissible forms of the parasite in *P. falciparum*, phenotypic screening on these transmissible stages is far more costly, laborious and time-consuming than screening against ABS parasites. As a result, the phenotypic screening data for libraries screened against the gametocyte stages of the parasite pales compared to the phenotypic screening data available for libraries screened against ABS. Supervised ML methods are valuable tools to identify patterns and build predictive models for compound activity, however, one pitfall of this type of ML is that it requires a sufficient amount of labelled data to train a model to allow for good generalisation and accuracy. If the training examples are insufficient, it can lead to the model having poor generalisation when exposed to new examples, increasing the prediction error. This is often the case in fields where experimental data is limited and costly to label. In cases where the labelled training data is limited, applying standard ML techniques will produce prediction models with poor generalisation. Another hurdle, common within most phenotypic screening settings, is the class imbalance present within phenotypic

screening data, as only a tiny fraction of compounds show activity against the parasite. Worse still, the identification of compounds which show dual activity (ABS and gametocyte stages) is even rarer within such screens. This imbalance can result in poor prediction performance within models trained on such data, as the models will focus on optimising accuracy at the expense of identifying rare events. The field of ML, however, is a continuously evolving field that improves and develops techniques to mediate areas where conventional ML often fails or obtains poor results. For instance, DeepMalaria has shown that ML can successfully predict asexual hit compounds by employing sampling techniques to address class imbalance within phenotypic screening data and leveraging transfer learning to overcome the constraints associated with limited data [82].

1.5) ML sampling techniques to address class imbalance.

Traditional ML models often struggle with imbalanced training datasets as they strive to obtain the best accuracy, often ignoring the minority class, even though the minority class might be of utmost importance. In such instances, imbalanced data must often undergo class imbalance correction during pre-processing before model building to negate such effects during model training. Class imbalance correction can be done either through oversampling, undersampling, or a hybrid approach. Both sampling methods have their advantages as well as disadvantages. Oversampling, for instance, increases the minority class by replicating samples within the minority class to address the imbalance within the training data. This method may allow better pattern detection within the minority class, however, it may also produce overfitting because certain features are overrepresented within the training data and result in poor generalization when exposed to new instances. Synthetic minority oversampling technique (SMOTE) is another form of oversampling where new minority class instances are generated through interpolation between a minority class instance and its' k-nearest minority class neighbour [104]. Adaptive Synthetic Sampling is an adaptation of SMOTE to generate synthetic samples of the minority class inversely proportional to the density in the minority class to generate new samples in the feature spaces where the concentration of the minority class is low and hard to learn [105]. Border-line SMOTE similarly employs SMOTE in regions where the minority and majority class instances are close to one another to generate border-line minority instances to aid in the decision boundary classification [106]. Although this is a good technique to generate more

samples within the minority class, this may not be representative of the minority class and may introduce noise and generate outliers within the data [107, 108].

In contrast to oversampling, undersampling tries to eliminate instances within the majority class randomly, however, doing so may also result in the loss of essential features for the prediction of the majority class. Cluster-based undersampling is an attractive alternative to circumvent such loss of information, as only instances that are very similar and redundant to one another and that cluster together are removed. In contrast, the remaining instances within the cluster represent the chemical space for training. Such an undersampling technique hence negates the loss of instances important for defining chemical spaces for inactivity. Near-miss is another undersampling technique using KNN where majority class instances are selected if they have the smallest distance to the K nearest neighbours in the minority class [109]. Tomek links a different undersampling approach, which, in contrast, identifies the nearest neighbours between different classes and removes the majority classes, which are prone to misclassification [110]. This, however, may remove significant majority class instances for defining the decision boundary between classes.

1.6) Transfer learning (TL)

To address the problem of limited training data for model building, transfer learning (TL) is an ML technique that transfers predictive ‘knowledge’ from one model or data source to another. TL is based on the phenomenon that when a person learns a task such as chess, specific skills are learned that can indirectly help a person perform better in a related but different task, such as critical decision-making. Within an ML context, the goal is to transfer ‘knowledge’ from a pre-trained model trained on enough training data to perform a defined task to a new model with a different yet related task but where the training data is insufficient [111]. The rationale is that the pre-trained model may contain ‘knowledge’ in identifying essential features that will aid in making accurate predictions for the new ML model towards its’ intended task (Figure 1.8). This has proven successful in numerous studies where the data was unlabelled and/or the training data was insufficient to produce an accurate ML model [112-117]. During training a model, the prediction error is high in the beginning, but as more examples are provided, the algorithm will start identifying patterns within the training examples, and the model will improve in its generalization, thereby lowering the prediction error if enough training examples are available (Figure 1.8). When few training examples

are available, the model's learning to identify patterns and to generalize between examples is halted and the error will remain high or increase. With TL, a pre-trained model is trained on source data (containing many examples) similar to the target data (containing a limited amount of examples) and during the TL, knowledge relating to identifying important features for prediction is transferred to a target model, which is trained on the target data.

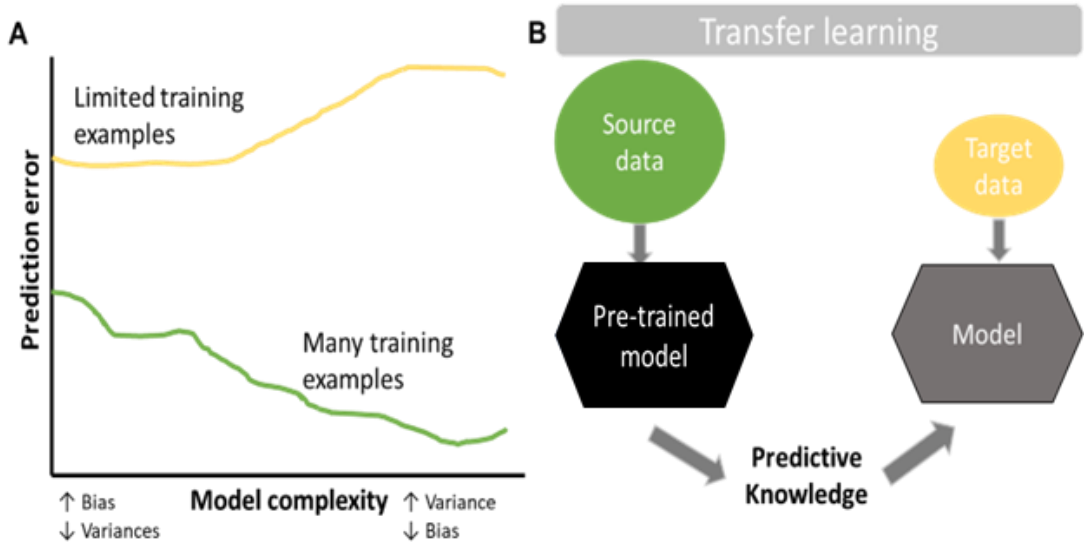


Figure 1.8: Data-limited generalization error and TL.
(A) Graphic representation of the model complexity vs prediction error during model training. (B) Example of TL where a pre-trained model is trained on source data and predictive knowledge is transferred on the target model.

There have been many different types of TL techniques developed, each based on the data available, the task assigned to models and whether the domains of the source data (data used to train the pre-trained model) or the target data (limited data available to train the new model) are similar. TL can ultimately be grouped into three categories. One is inductive TL, where both source and target domains are similar, but the intended tasks are different. Transductive TL is where both source and target domains are not entirely the same, and only labelled data is available from the source data [111]. Unsupervised TL, however, is when both target and source domains are similar, yet no labels are available in either domain [118]. Within a drug discovery context, the feature space of chemical matter screened against organisms is essentially the same, the only difference being that the marginal distributions, i.e. the number and diversity of the compounds screened and hence the chemical space, may differ between different targets/organisms or stages of development. Ultimately, in a phenotypic screening context, the chemical space is the same if all chemical matter is considered, meaning a homogenous TL approach can be taken as both the source

and target domains are similar. Considering this, three types of homogenous TL can be applied within drug discovery and are discussed below.

1.6.1) Instance-based TL

Instance-based TL, as the name suggests, uses instances in the source domain and applies various weighting tactics to such instances to use them with the target data to train the target model. Examples of such algorithms include TrAdaBoost and TransferBoost, which use boosting strategies within TL to optimise model performance. TrAdaBoost is based on the AdaBoost algorithm, where misclassified training instances are more highly weighted to force the correct prediction of such instances within the models. With TrAdaBoost, however, if source instances are misclassified, the algorithm will reduce the weight of such instances and their influence on model training. In contrast, source instances that aid in correct classification have more weight than those different from the target data [119]. TransferBoost similarly adds weight to instances within the source domain that allow positive TL (correct predictions within the target domain), but it also utilises AdaBoost on both target and source instances to weight them accordingly to achieve higher accuracy [120]. In such a way, source instances that are predicted correctly but do not aid in generalisation in the target domain are weighed less than source instances that improve generalisation.

1.6.2) Parameter-based TL

Parameter TL is a more straightforward method of TL in which the parameters defining the source model architecture are shared with the target model. This approach has been most extensively applied within ensemble and neural network models such DNN and convolutional neural networks (CNN) [121, 122]. Due to the architecture of neural networks, they are very efficient in extracting high-level features from data that are important for producing correct predictions and more informative than low-level features. Many training instances, however, are needed to find the optimal architecture of the model that can extract these features to make correct predictions. If this neural network architecture is optimised to extract important predictive features, it can be applied to a different but related classification problem and be fine-tuned to that problem (Figure 1.9). The general architecture of an ANN consists of an input layer where data points are fed into input nodes in the ANN, this information is then passed to nodes within the hidden layer. A neural network with multiple hidden layers present is called a deep neural network and can use these layers

to transform the data for better pattern detection. Connections between input and hidden layers have weights that are updated and fine-tuned during model training on the data to add weight and importance to information that leads to the correct prediction (output).

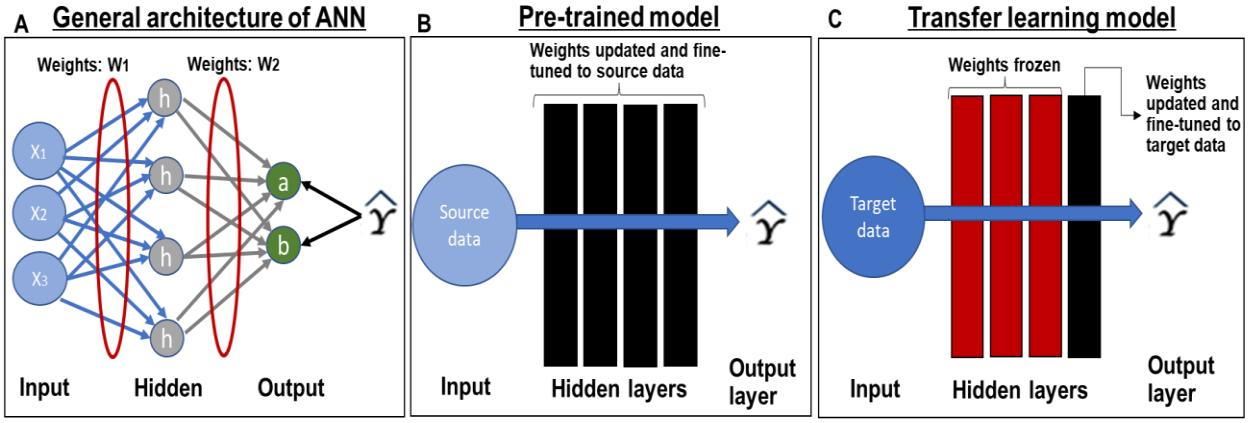


Figure 1.9: TL through transferring DNN architecture.
(A) The general architecture of an ANN.(B) A pre-trained DNN base architecture optimisation and (C) a DNN base architecture transferred to the target model.

A pre-trained model's architecture is optimised through training on the training examples in the source dataset. The DNN base architecture is then transferred to the target model by freezing the weights of the hidden layers. Typically, the last hidden layer within the architecture is unfrozen to allow fine-tuning the last layer towards the target data, whereby the weight withinin the last layer is updated to fit the target data.This method of transferring a neural network's architecture from one classification problem to another has shown to be successful and proven to give better results in classification accuracy than training a neural network on a dataset with limited training instances [82, 123].

1.6.3) Feature-representation TL

Feature-representation TL employs domain adaption to transfer knowledge from the source domain to the target domain by identifying a good feature representation across both domains, which can minimise the difference in distributions between the two domains. An example of this is joint distribution adaption (JDA), which takes the difference between the marginal and conditional distribution into account and aims to learn a feature representation that reduces differences in the marginal and conditional distributions of the source and target domain. Marginal distribution refers to the probability distribution describing the likelihood of only one variable occuring independent of other variables [124]. Conditional distribution,

however, refers to the probability distribution describing the likelihood of only one variable given the condition of the outcome of other variable(s) [125]. When plotting the distributions of the target and source domains in a feature representation, the maximum mean discrepancy is calculated and used to measure the difference between the marginal and conditional distributions of the source and target domains. This maximum mean discrepancy is then used to create a new feature representation of the source and target domain where the differences in the marginal and conditional distributions of the source and target domain are reduced. This new feature representation can then be used by traditional ML methods to build a model [126].

1.7) In summary

Transmission-blocking compounds are urgently needed to combat the spread of malaria parasites and the formation of resistance against current treatments. The parasite's biology, however, makes the identification of gametocytocidal and dual-active compounds through phenotypic screening costly and time-consuming. Multiple areas in drug discovery have utilised ML to accelerate drug discovery by guiding decision-making in compound prioritization. This can similarly be applied to identify and prioritize the phenotypic screening of compounds with a high probability of having gametocytocidal activity. The limited chemical libraries screened against gametocytes to serve as training data and severe class imbalance present within phenotypic screening data may prove challenging for model building. Nonetheless, we theorise that ML techniques such as class sampling and transfer learning may alleviate these limitations within gametocytocidal phenotypic screening data to allow for the building of robust models in gametocytocidal activity prediction.

Hypothesis

ML can build robust models demonstrating good performance to predict compounds with transmission-blocking activity against the *P. falciparum* parasite.

Aim

To build and interrogate different ML algorithms for their ability to robustly predict the chemical space of bioactivity against the asexual and/or gametocyte stages of the *P. falciparum* parasite.

Objectives

1. Training and evaluation of TL models' robustness towards identifying compounds with activity against gametocyte stages (Chapter 2).
2. Training and evaluation of traditional ML models in their robustness towards identifying compounds with activity against both ABS and/or gametocyte stages (Chapter 3).
3. Analysis and interrogation of chemical features predictive of stage-specific activity against *P. falciparum* (Chapter 4).

Outputs generated

Publications

1. **Heerden, A.v.,** Turon, G., Duran-Frigola, M., Pillay, N., and Birkholtz, L.-M. (2023). Machine learning approaches identify chemical features for stage-specific antimalarial compounds. bioRxiv, 2023.2008.2015.553339. doi: 10.1101/2023.08.15.553339.
2. **Heerden, A.v.,** Turon, G., Duran-Frigola, M., Pillay, N., and Birkholtz, L.-M. (2023). Machine learning approaches identify chemical features for stage-specific antimalarial compounds. ACS Omega. 2023 Nov 7;8(46):43813-43826. doi: 10.1021/acsomega.3c05664. eCollection 2023 Nov 21.

Conferences

1. **van Heerden, A**., van Wyk, R., and Birkholtz, L.M. (25 – 28 October 2022). , Applying ML to chemo-transcriptomic profiles to stratify antimalarial compounds with similar modes of action (Ashleigh van Heerden, University of Pretoria). Oral presentation. 4th H3D symposium *A Spectrum of Opportunities in Infectious Disease Drug Discovery to Enhance Global Health Cape Town, South Africa.*

CHAPTER 2

Evaluating TL models on phenotypic screening data against asexual and gametocyte stages.

Some of the work in this chapter was included in a publication as follows:

Heerden, A.v., Turon, G., Duran-Frigola, M., Pillay, N., and Birkholtz, L.-M. (2023). Machine learning approaches identify chemical features for stage-specific antimalarial compounds. ACS Omega. 2023 Nov 7;8(46):43813-43826. doi: 10.1021/acsomega.3c05664. eCollection 2023 Nov 21.

2.1) Introduction

With the innovation of ML, many models capable of identifying compounds with desirable activity and properties have been generated and evaluated for implementation to aid in drug discovery [19, 26, 27, 32-34, 41, 45, 46]. Such models can prioritise compounds with desirable activity without using resources to screen compounds beforehand [127, 128]. These models have the added benefit of making predictions regarding compound activity within seconds compared to phenotypic screening, which can take days to weeks, depending on the biological system.

In the case of antimalarial drug discovery, these advantages become enticing, particularly to identify gametocytocidal compounds or even dual-active compounds (activity towards both gametocytes and ABS parasites), since the *in vitro* production of gametocytes is cumbersome and time-consuming, and assays lengthy and expensive [129]. However, developing such models is challenging due to the limited, curated gametocytocidal phenotypic screening data available [130-132] compared to ABS [130, 132-136]. Models trained on limited data can have poor generalisation and be error-prone in classifying compound activity [137]. Severe class imbalance within phenotypic screening data, where the majority of compounds are inactive with few active compounds against the gametocyte stages of the parasite, further exasperates the building of models predicting gametocytocidal activity. This imbalance complicates pattern recognition within traditional ML models, resulting in models being unable to detect the minority class (active compounds) and classifying most inputs as the majority class (inactive) to achieve ‘good accuracy’ [138], making them ineffective as pre-screening tools. Despite this, there are ML methods to circumvent such limitations within the gametocytocidal phenotypic screening data, such as using TL and class imbalance correction techniques. Due to antimalarial models such as

DeepMalaria showing success in improving model performance in predicting ABS inhibition activity through the use of TL and oversampling [82], we wanted to investigate whether a similar approach could be used for the prediction of dual-active/gametocytocidal compounds.

2.2) Methods

2.2.1) Ethics

All computational work, the saving, modifying and storing of data has ethics approval (University of Pretoria Ethics Application NAS345/2020).

2.2.2) Acquisition of *in vitro* phenotypic screening data for database assembly

Chemical libraries screened against either the asexual and/or gametocyte stages of the parasites were acquired (Table 2.1), and the inhibition data was extracted and pre-processed. During pre-processing, compounds with missing inhibition data and inorganics and organometallic compounds were removed. These chemical libraries contain multiple chemical spaces, each with several chemically similar analogues, with some compounds having additional dimensionality in that they were screened against additional parasite stages.

Table 2.1: Chemical libraries phenotypically screened against *P. falciparum* asexual blood stages and/or gametocyte stages

Chemical library	No. compounds in library	Stage screened	No. compounds obtained	No. compounds (balanced database)	Ref
GSK Library	~ 2 million	ABS	40 510	ABS: 40 398	[133]
TCAMS library	~14 000	Dual	13 533	Dual: 409	[131]
St. Jude's Library	~ 310 000	ABS	5 456	ABS: 3 754	[134]
Novartis-GNF Malaria Box	1.7 million	Dual	11 394	ABS: 9 689	[135]
Global Health Chemical library	~70,000	Dual	68 614	ABS: 14 584 Dual: 7 901	[130]
MMV Box	400	Dual	400	ABS: 304 Dual: 277	[132, 139]
PRB Box	400	Dual	400	NA	[136]
Pathogen Box	400	Dual	367	NA	[140]
Total screened (ABS)	-	-	126 374	68 729	-
Total screened (Dual)	-	-	93 941	8 614	-

ABS = asexual blood stage. Dual = ABS and sexual stages

Two databases were created from these pre-processed chemical libraries. The first is the ABS database containing SMILES and inhibition data from the five chemical libraries screened against *P. falciparum* ABS parasites [130, 132-136]. The second database (referred to as the dual-active database) similarly contained SMILES and inhibition data from

chemical libraries phenotypically screened against ABS of the parasite and any of the gametocyte stages (stage I-V) of *P. falciparum*, however, the majority of screening data obtained focused on stage IV/V gametocytes [130-132] (Table 2.1).

2.2.3) Defining inhibition thresholds for compound activity

The chemical libraries acquired had been screened by different research groups using different assay platforms, with different thresholds set to define parasite inhibition and gametocytocidal activity. This made defining compound activity difficult, as setting a standard threshold to define activity across all phenotypic screens may lower the data quality and unwittingly introduce noise into the data. For example, certain groups phenotypically screened compounds at higher concentrations using the same assay platform as others; hence, they raised the threshold defining compound activity to ensure the identification of potent compounds at much higher screening concentrations. Thus, to maintain data quality, rather than selecting a standard threshold, the thresholds specified within the respective phenotypic screens were used verbatim to define active/inactive compounds for parasite viability inhibition (Table 2.2).

For both databases, inactive compounds were retained as these compounds are informative in defining the chemical space for inactivity. Compound SMILES were extracted for all compounds included in the respective databases and used to calculate physiochemical properties and molecular descriptors using RDKit [141]. From the calculated physiochemical properties and molecular descriptors, a uniform manifold approximation and projection (UMAP) analysis was performed to identify and remove any outliers and to generate a projection of the chemical composition of the databases with the umap-learn python package version 0.5.3 [142].

Table 2.2: Chemical libraries included in databases and the assay platform and threshold used to define active compounds

Chemical library	Assay	Threshold (defining active compounds)	Ref	ChEMBL ID
ABS database				
Global Health	SYBR green	>70% inhibition at final compound concentrations of 5 and 10 µM	[130]	CHEMBL4513216
GSK	LDH assay	>50% inhibition at final compound concentrations of 2 µM	[133]	CHEMBL1157539
Novartis	SYBR green assay	>50% inhibition at a final concentration of 1.25 mM	[135]	CHEMBL1156829
St. Jude	both SYBR Green and YOYO-3 assays	Single point growth inhibition ≥80% in both SYBR Green and YOYO-3 assays were defined as hits At a final concentration of 7 µM of the test compounds.	[134]	CHEMBL1154198
MMV Box	asexual HCI assay utilizes the DNA-intercalating dye DAPI	>50 % inhibition at a final concentration of 5 µM	[132]	See respective article for data
Dual-active database				
Global Health (Stage V)	Real-time luciferase-dependent parasite viability assay	>50% Average gametocyte count reduction at final concentration of 2 µM	[130]	CHEMBL4513216
TCAMS (GSK)	Female gametocyte activation assay (Pf FGAA), which assesses stage V female gametocyte viability and functionality using Pfs25 expression	>50% inhibition at a final compound concentration of 2 µM	[131]	See respective article for data
MMV Box	Multiple **groups** conducted phenotypic screens on gametocytes using different *assay platforms* and *final screening concentrations*	Average Activity >50% inhibition (Note: hits were defined as compounds that obtained >50% average activity for Stage I-III (For **Winzler**[*final concentration of 12.5µM, MitoTracker® Red]* and **Avery** group[*final concentration of 5 µM and 0.5 µM, transgenic line with MitoTracker® Red]*) as well as >50% activity for Stage IV-V (**Avery** [*final concentration of 12.5µM, transgenic line with MitoTracker® Red*] and **D. Taramelli** group screening [*final concentration of 3.7 µM, pLDH assay*]))	[132]	See respective article for data

2.2.4) Pre-processing of acquired datasets into ABS and dual-active database

The generated databases (Table 2.2) included structural information (SMILES) of active and inactive compounds that have been phenotypically screened on gametocyte stages and/or ABS of *P. falciparum* parasites. As is common in phenotypic screens, the data within the databases were the majority of inactive compounds and necessitated class imbalance correction techniques to address the class bias. Of all undersampling techniques to sample inactive compounds, cluster-based undersampling was the most attractive as it did not result in loss of chemical space information and would not remove borderline instances, which could be important for defining decision boundaries or removing instances prone to misclassification. Cluster-based undersampling was thus performed on inactive compounds in both the ABS and dual-active databases, respectively. Inactive compounds were clustered using Tanimoto dissimilarity (0.4 distance threshold) with RDKit version 2022.9.5 [141] to allow chemical substructure searching and clustering of inactive compounds with similar substructures. Emphasis should be placed on the fact that the goal of clustering was to aid in undersampling and not to cluster the inactive compounds entirely and that this was performed independently for both the 'ABS' and 'dual-active' databases.

Inactive compound SMILES of the respective databases were converted into RDKit molecular fingerprints (parameter max path = 5) for cluster-based undersampling. To prevent creating any inherent bias within the training data, different molecular fingerprints were used for cluster-based undersampling and model training. With such large databases to limit the computation power required for clustering, the inactive compounds were divided into seven subsets for the ABS database but only three subsets for the dual-active database, with each subset containing 15 – 20 000 compounds. Parallel clustering (Figure 2.1) was then applied to each of these subsets individually. From this, a representative compound was randomly selected from each inactive cluster within the individual subsets. All representatives for the inactive clusters from the different subsets were merged into a single set before chemical clustering was repeated. This process was repeated until the number of representatives (inactive compounds) was ≥95% of the number of active compounds, after which the database was considered balanced. Unfortunately, complete balancing could not be achieved with cluster-based undersampling for the dual-active database, and the iterative clustering process was halted before reaching the minimum number of clusters.

This was done to prevent severely limiting the chemical space as using the minimum number of clusters would limit the model training to only one compound defining the chemical space of a whole cluster of inactive compounds. This may cause difficulty in pattern detection by only having one compound instead of a select few compounds to be representative of a chemical space, possibly resulting in the inability to capture such chemical spaces and a loss in chemical space information.

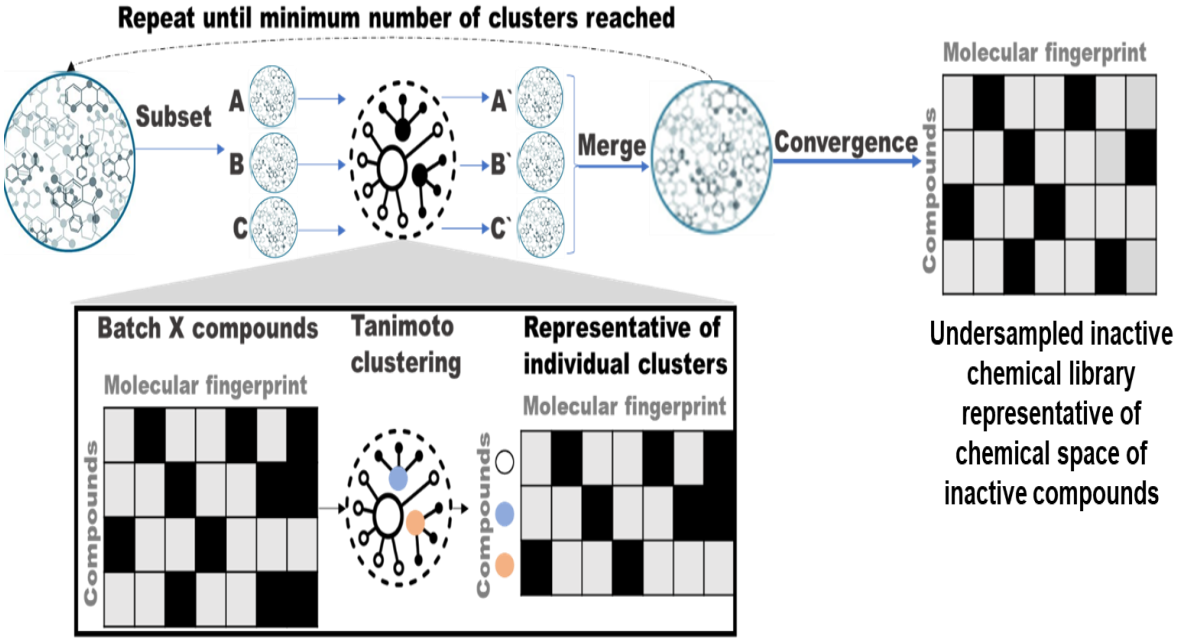


Figure 2.1: Parallel Tanimoto clustering-based undersampling of compounds inactive compounds against the parasite.
Subsets of the inactive chemical compounds screened against either gametocytes and/or ABS of the parasites were chemically clustered, and representatives of the clusters were randomly selected and merged with the other subset representatives (e.g., A'). Clustering was re-initiated with the compounds of the merged subsets until the inactive compounds representing the chemical space of inactive compounds were almost similar in number to the active compounds within the respective databases.

If class imbalance was still present within the target data (dual-active database) after undersampling, a hybrid sampling approach was performed where inactive compounds underwent cluster-based undersampling and active compounds were oversampled. Due to the concerns with SMOTE generating synthetic examples of compounds possibly not representing the actual activity or the feasibility of said compounds, random oversampling was performed via imblearn version 0.11.0 [143]. For comparative purposes, models trained on undersampled data and models trained on under- and oversampled data would be evaluated to determine which sampling technique enabled better performance.

2.2.5) TL model building

The ABS database as source and dual-active database as target domain are similar and the task in identifying active compounds is the same. However, the source and target domains differ in their marginal distributions and conditional distributions. The marginal distribution difference is due to more instances in the source than in the target domain. Hence, there is a more significant representation of the chemical (feature) space in the source domain than within the target domain. The conditional distribution difference between the two domains links to the compound activity differing between different stages, as there are compounds that can be labelled as ‘active’ within the source domain due to the presence of some chemical feature, however, within the target domain they are labelled as ‘inactive’. There will also be many more compounds labelled ‘inactive’ in both the source and target domain, whereas compounds labelled ‘active’ in both the source and target domain are rare.

Based on this homogenous data setting, ‘knowledge’ from the source domain can be transferred to the target domain through instance-based transfer, feature-representation transfer or parameter-based transfer. Instance-based TL may enable better model performance for gametocytocidal prediction by identifying instances within ABS phenotypic screening data that aid in predicting gametocytocidal activity. Alternatively, considering the possible differences in marginal and conditional distributions within the source and target domains, feature representation TL like joint distribution adaptation (JDA) may also be of use by adapting both the source domain and target domain to reduce differences between the two domains and allow better performance when trained on this adapted data. On the other hand, with parameter-based TL, the DNN architecture for predicting ABS inhibition activity may contain hidden knowledge in extracting important chemical features for gametocytocidal activity prediction, such as features that allow better membrane permeability to specific target sites. Utilising this architecture and transferring it to gametocytocidal models through parameter-based TL might drastically limit the number of training examples required to identify the optimum DNN architecture for gametocytocidal prediction.

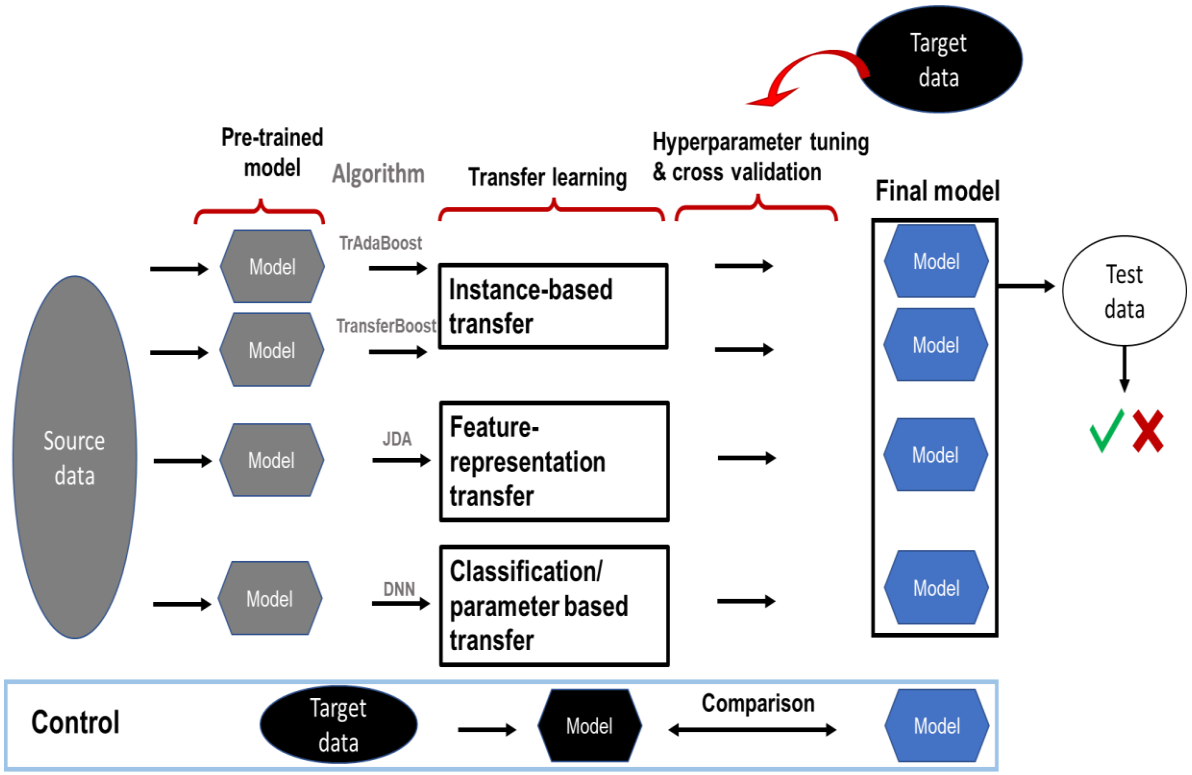


Figure 2.2:TL workflow and analysis.
Pre-trained models were trained from the source dataset (SMILES of compounds screened against ABS parasites). Different TL methods (black boxes) will be investigated in their ability to increase model performance/accuracy on the target dataset (SMILES of compounds screened against sexual stages) compared to the control models, which were trained solely on the target dataset. The models from these TL methods will undergo hyperparameter tuning from which the best 'dual-activity' prediction model will be selected and validated through our test set.

All three TL methods were investigated for their ability to transfer 'active compound' knowledge from the source domain to the target domain using the workflow developed in Figure 2.2. Where possible, the optimal hyperparameters of TL models were identified and used for model building. For instance-based TL, the TrAdaBoost and Transferboost algorithms would be used. TrAdaBoost models were created using the adaptation (version 0.2.0) python package [144]. To determine whether TL positively impacted model performance, the performance of the TrAdaBoost model was compared to that of control Adaboost models built via scikit learn [145] that would serve as a baseline. Similarly, for TransferBoost models, the transferboost (version 0.1.3) python package was used for model building and compared to XGBoost models built using the xgboost (version 2.0.1) python package. For feature representation TL, the JDA algorithm will be used. Joint distribution adaption (JDA) models were created using the online python code provided by Lone *et al*. [126] and would be compared to KNN models built via scikit learn [145] as a baseline for model performance without TL. Finally, to investigate parameter-based TL, a DNN model was trained on the ABS database using keras [146] and tensorflow [147] during which the

optimal DNN architecture and hyperparameters were identified via keras-tuner [148]. The DNN architecture with its weights was fixed, and the last hidden layer within the DNN architecture was fine-tuned and weights adjusted towards the dual-active database. TL DNN models were then compared to a baseline DNN model trained on the dual-active database without any TL. During TL model training, both cluster-based undersampling and a hybrid approach of cluster-based undersampling and oversampling were used to evaluate model performance. Hence, two models were generated for each algorithm investigated, where training was done on either the undersampled training set or an under-and oversampled training set. Models were initially trained on 100-bit ECFP at two atom radius, however, to determine whether increasing the number of chemical features resulted in improved model performance, the ECFP bit-length was then increased to 500 bit-length at five atom radius.

2.2.6) Evaluating performance of different ML and TL models on test set in predicting gametocytocidal compounds

Fine-tuned models were evaluated on test set results to determine model performance on untrained imbalanced chemical data. Performance metrics used for model evaluation were accuracy, recall (equation A), and precision (B). To determine whether these TL methods enabled positive TL (improved model performance), TL models were compared to control models, which would serve as the baseline for model performance without any TL employed.

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \qquad \text{(A)}$$

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \qquad \text{(B)}$$

2.3) Results

2.3.1) Database assembly and pre-processing

Data from chemical libraries screened against the asexual and/or gametocyte stages of *P. falciparum* parasites (Table 2.1) [130-136] were used to compile the two databases for two different modelling environments. The first modelling environment was for the training of models to predict compounds with stage-specific activity against ABS parasites containing chemical libraries screened against the ABS of the parasite and would serve as our source dataset. The second modelling environment containing chemical libraries screened against

gametocyte stages and ABS stages of *P. falciparum* parasites would serve as our target dataset for model training towards predicting dual-activity.Within both ABS and dual-active databases, after removing organometallics and inorganic compounds, the chemical space was visualized using UMAPS and no outliers were detected within any of the databases (Figure 2.3). It was also observed that the chemical space for dual activity was narrower and less defined than that of active compounds for ABS activity.

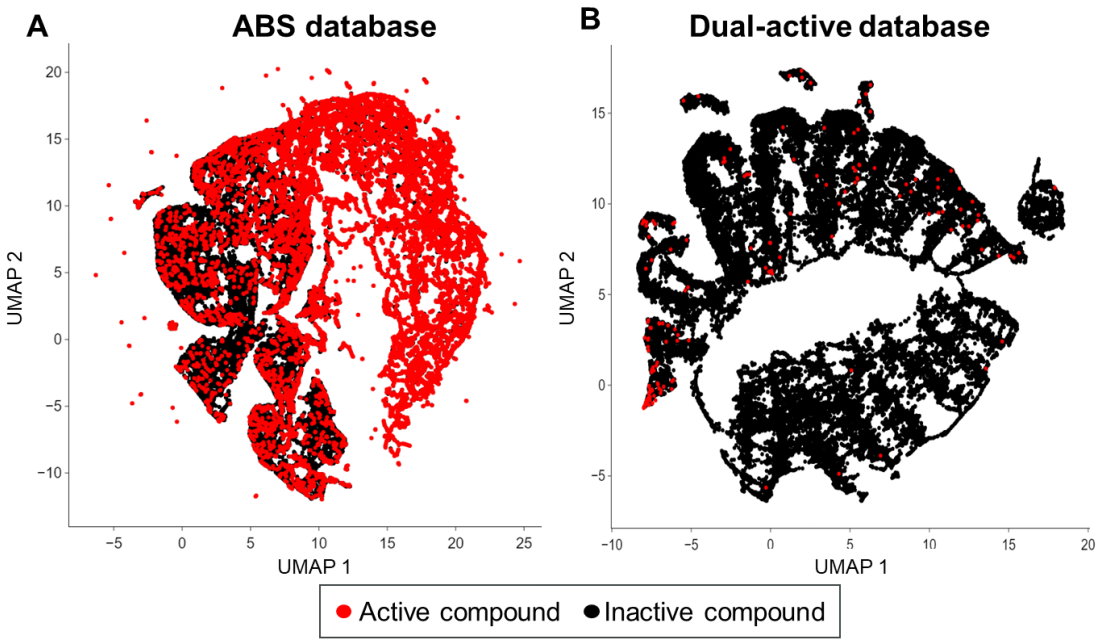


Figure 2.3: Outlier detection in ABS and dual-active database via UMAP.
(A) Chemical space of active (red) and inactive (black) compounds from the ABS database after pre-processing was visualized via UMAPs using Morgan fingerprints (500-bit length) and physiochemical descriptors predicted via RDKit [141]. (B) Similarly, the chemical space of active (red) and inactive (black) compounds from the dual-active database after pre-processing was also visualized using UMAPs using Morgan fingerprints and the same physiochemical descriptors predicted via RDKit. From the chemical space projection of the databases, no outliers were detected. Physiochemical descriptors included molecular weight; log P; number of H donors/acceptors; number of rotatable bonds; TPSA; ring count; number of heteroatoms; aromatic bonds; acidic groups and basic groups. Note: active compounds are layered over inactive compounds for the. visualisation of severe class imbalance within the dual-active database.

2.3.2) Class imbalance correction of databases

The ratio of inactive to active compounds is inherently skewed towards the inactive compounds, which comprises 75% of the ABS database (30 393 active *vs.* 92 178 inactive compounds). This class imbalance is even more severe within the dual-active database, where inactive compounds comprise 99% of the dual-active database (916 active *vs.* 68 801 inactive compounds) (Figure 2.4A and D). Cluster-based undersampling was performed on inactive compounds for each database to lessen the class imbalance within the databases

while preventing the loss of relevant chemical space information in inactive samples [149]. This generated balanced databases that are more useful for ML modelling [150].

Since the database's chemical libraries contained multiple structurally related compounds falling within specific chemical spaces with similar inactivity (Figure 2.4B), such a clustering approach could select representative compounds for such well-defined chemical spaces. UMAP analysis and spatial projection showed retention of the chemical composition and diversity of the ABS database (Figure 2.4B) after two rounds of subset clustering of inactive compounds within the ABS database where the class balance between active and inactive compounds was obtained (30 939 active compounds *vs*. 29 143 inactive compounds, Figure 2.4C). Unfortunately, the class imbalance could not be corrected even after multiple rounds of clustering of the inactive compounds within the dual-active database. Complete clustering of inactive compounds only resulted in a maximal of 3 745 clusters representing the inactive compounds, which still outnumbered the dual-active compounds (916).

To not hinder model performance by selecting only one chemical representative to encapsulate a whole cluster of inactive compounds, additional parallel chemical clustering was performed, and this process halted before reaching the maximal number of chemical clusters. This partial clustering allowed more than one chemical representative of clusters to be present (8 975 clusters) in defining the chemical space for ML of the inactive compounds and resulted in an <8-fold decrease in the observed class imbalance (916 active compounds *vs*. 8 975 inactive compounds, Figure 2.4D and F).

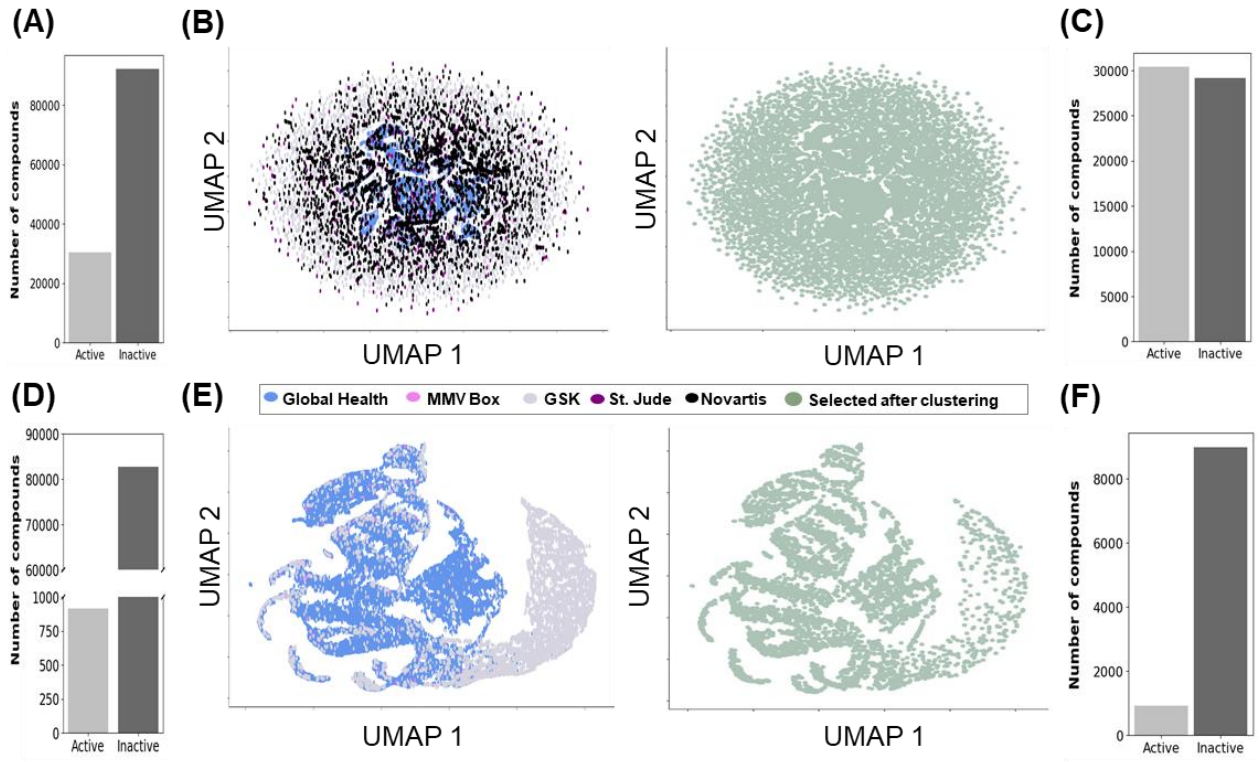


Figure 2.4: Cluster-based undersampling of databases to address class imbalance.
(A and D) Class imbalance in the ABS (A) and dual-active (D) datasets after binary classification of activity vs. inactivity based on criteria as specified in original screens. (B and E) UMAP projection of the chemical space in the databases before (lefthand image) and after (righthand image) cluster-based undersampling on inactive compounds for each database. (C and F) Distribution of active vs. inactive compounds for the ABS (C) and dual-active (F) datasets after cluster-based undersampling.

Such clustering also allowed the retention of inactive compounds' chemical distribution and composition within the dual-active database (Figure 2.4E). Hence, from the cluster-based undersampling, more balanced databases were obtained. The compounds within these databases were then shuffled and randomly split, where 80% of compounds were used for model training, however, the remaining 20% of compounds not included in model training were merged with the inactive compounds excluded during cluster-based undersampling to create an imbalanced test set for model evaluation.

2.3.3) TL model performance in dual-stage activity prediction trained on ECFP with 100 features

Multiple TL models using different TL and class imbalance correction techniques were investigated to build models capable of predicting dual-stage activity. Unfortunately, during the model building of JDA models, the time allocated for model building drastically increased as the sample size from both the ABS and dual-activity databases increased. The computing

power available could no longer support model training after surpassing a sample size of 12 500 compounds from both ABS and dual-activity databases. This high demand for excess computing power could be the result of the JDA trying to calculate the best transformation of the ABS and dual-active database and applying it to each compound therein. Due to this and the coin-toss accuracy achieved, which showed no improvement as the sample size increased, JDA model building was terminated (Figure 2.5).

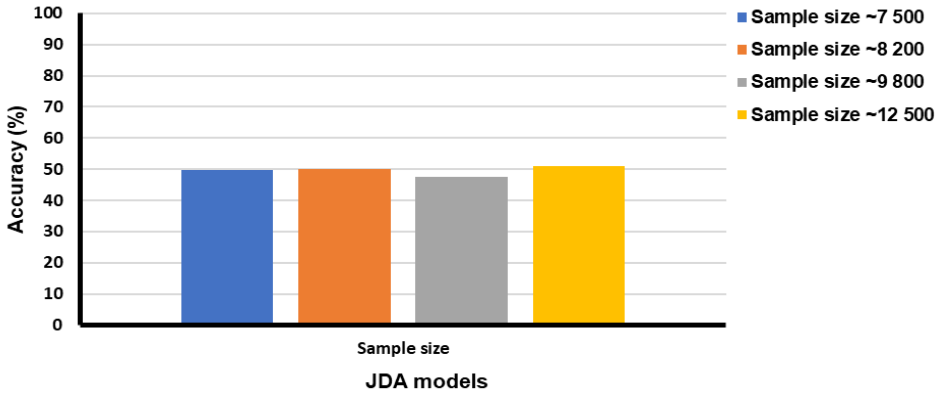


Figure 2.5: JDA accuracy plateau despite sample increase.
Influence of increasing the training sample (consisting of both compounds within the ABS and dual-active database) on JDA model performance in predicting dual-active compounds.

During model training, the optimal hyperparameters for instance-based TL were identified (Table 2.3), and similarly, for parameter-based TL, the optimal DNN architecture for training on either undersampled or under and oversampled data was determined (Figure 2.6). Instance-based TL, TransferBoost and TrAdaboost models trained on under and oversampled data only resulted in models that performed no better than random guessing (~50% accuracy) (Figure 2.7A). Although TrAdaboost displayed good recall (Figure 2.7B), an inspection of model predictions revealed that the model classified almost all compounds as active (Figure 2.8A), which explained the 0.5 precision observed. TransferBoost and TrAdaboost models trained on only undersampled data from the source and target datasets (ABS and dual-active database) showed high accuracy in predicting compound activity with no performance improvement compared to the baseline models (XGBoost and Adaboost) (Figure 2.7A). Upon further investigation, it was discovered that this accuracy was due to models classifying most dual-active compounds as inactive compounds, hence the low precision and recall scores obtained from the models (Figure 2.7B & 2.8B).

Table 2.3: Hyperparameter tuning and optimal parameters identified for TL models on 100-bit ECFP

Parameter tuned	Parameters (range) used	Optimal parameter
Dual-activity prediction Transferboost model on undersampled data		
Learning rate	[0.00001,0.0001,0.0011, 0.01]	0.1
N estimators	[5,10,15,20,30,40,50,75,100,120,200]	50
Max depth	[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,20]	6
Dual-activity prediction Transferboost model on under and oversampled data		
Learning rate	[0.00001,0.0001,0.0011, 0.01]	0.0001
N estimators	[5,10,15,20,30,40,50,75,100,120,200]	50
Max depth	[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,20]	3
Dual-activity prediction TrAdaboost model on undersampled data		
Learning rate	[0.9,0.8,0.5,0.3,0.1]	0.9
N estimators	[30,50,100,120]	30
Dual-activity prediction TrAdaboost model on under and oversampled data		
Learning rate	[0.9,0.8,0.5,0.3,0.1]	0.8
N estimators	[30,50,100,120]	50

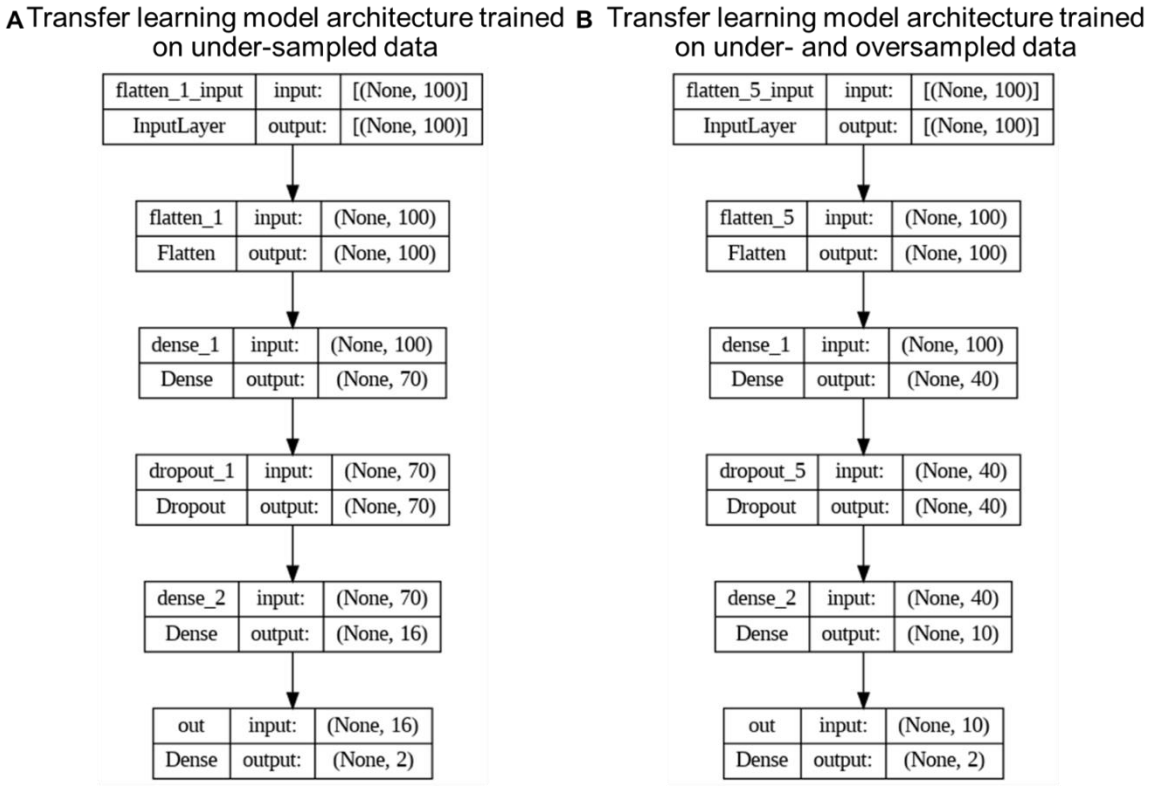


Figure 2.6: TL DNN models' overall architecture trained on either undersampled or under and oversampled data of ECFP with 100-bit-length.
(A) Deep neural network architecture acquired after training on the balanced ABS database, before fine-tuning the last dense layer with 16 nodes in the DNN to the training set of the undersampled (inactive compounds) dual-active database. (B) Deep neural network architecture was acquired after training on the balanced ABS database, before fine-tuning the last dense layer with 10 nodes in the DNN to the training set of the undersampled (inactive compounds) and oversampled (active compounds) dual-active database.

Although most TL models trained on under and oversampled data had better recall and precision than TL models trained on undersampled data, it was clear that such models struggled to differentiate between active and inactive compounds (Figure 2.8). Hence, it was more than likely that when such models are exposed to 'real world' imbalanced chemical

libraries would fail in prioritizing gametocytocidal compounds. Despite using a hybrid approach to address the class imbalance within the data, this did not seem to improve model performance for instance-based TL.

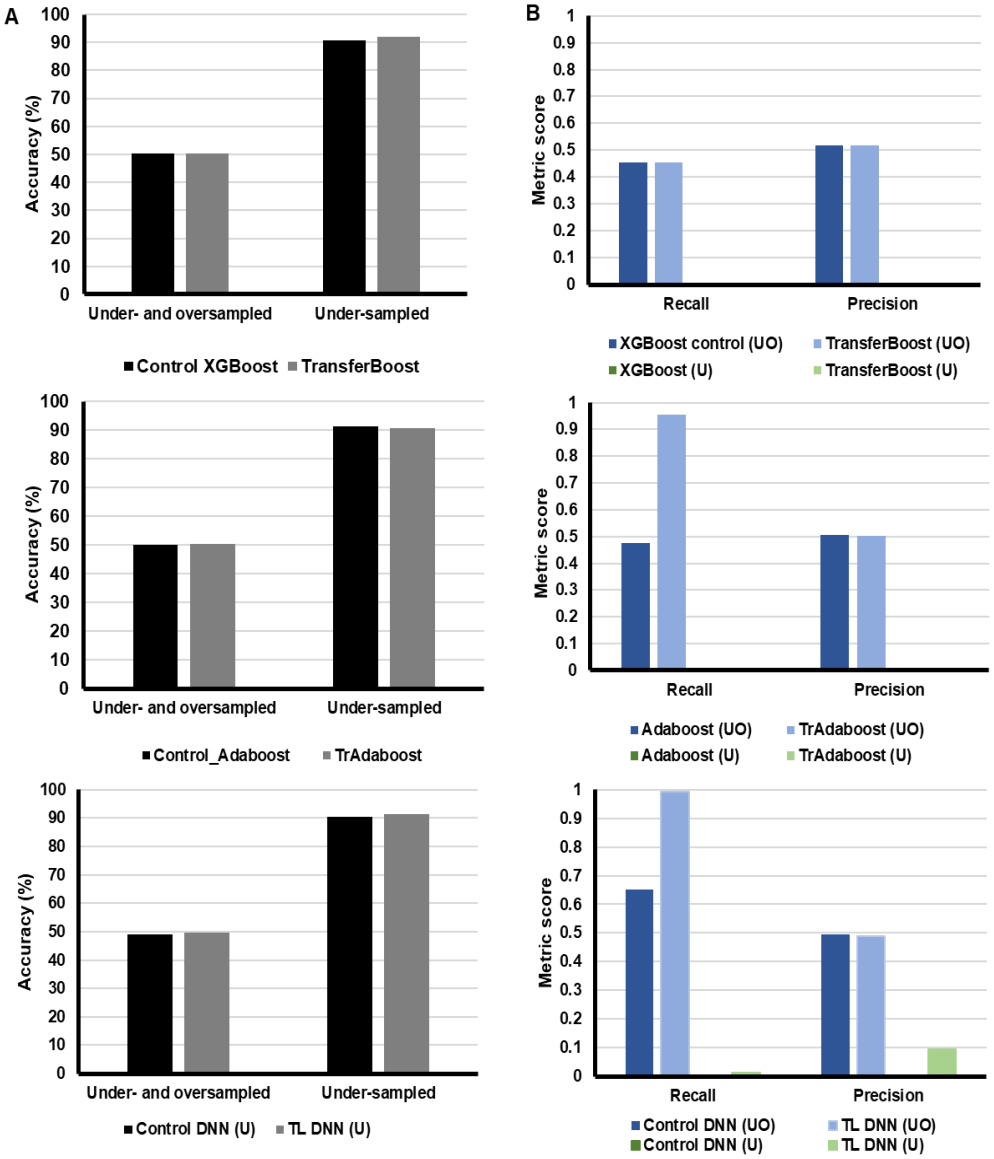


Figure 2.7: TL and baseline models accuracy, recall and precision in predicting dual-active compounds. (A) Accuracy of baseline (black) and TL models (grey) either trained on undersampled data (right) or under-and oversampled data were compared in predicting dual-active compounds. (B) The precision and recall of identifying dual-active were compared between baseline and TL models as well as the different class imbalance correction techniques employed. U= undersampling, UO= under-and oversampling.

With parameter-based TL models, similar trends were observed in that undersampling did not aid TL models in identifying patterns for the detection of dual-active compounds, whereas a hybrid approach utilising both oversampling and undersampling resulted in TL models that classified all compounds as active (Figure 2.8). One theory for the low model performance despite the class imbalance correction techniques implemented, was that the

number of features (100) used for training was too low to allow for pattern detection relating to compound activity. Hence, models were trained on ECFP with 500-bit length at a five-atom radius to determine whether increasing the number of features improved performance.

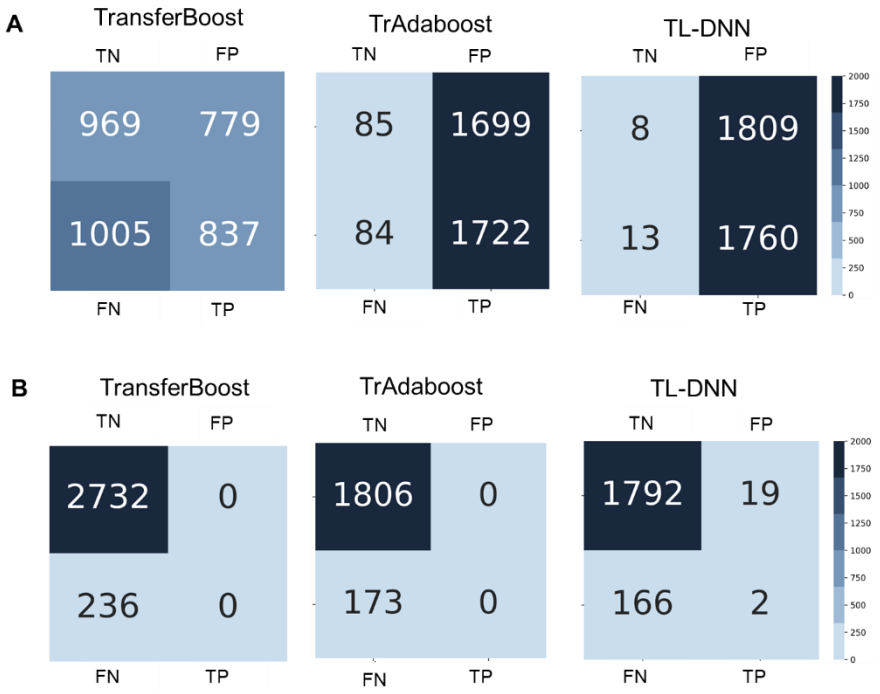


Figure 2.8: TL models predictive performance in identifying dual-active and inactive compounds.
The number of true negatives (TN), false negatives (FN), true positives (TP) and false positives (FP) obtained from predictions of TL models trained on (A) under and oversampled data or (B) only trained on undersampled data.

2.3.4) TL model performance trained on 500-bit ECFP of compounds

Models were additionally trained on 500-bit ECFP instead of 100-bit to determine whether adding more descriptive features would improve model performance in differentiating between inactive and dual-active compounds for both control and TL models. With preliminary experimentation, among the instance-based transfer learning techniques, TrAdaboost performed better than Transferboost using 100-bit length, thus for scalability we decided to focus on TrAdaboost, Additionally, during model training of Transferboost the model didn't scale well when increasing the bit-length to 500. The optimal hyperparameters for instance-based TL on undersampled or over-and-undersampled data were identified (Table 2.4), and similarly, for parameter-based TL, the optimal hyperparameters and DNN architecture for training on either undersampled or under and oversampled data was determined (Table 2.4 and Figure 2.9)

Table 2.4: Hyperparameter tuning and optimal parameters identified for TL models on 500-bit ECFP

Parameter tuned	Parameters (range) used	Optimal parameter
Dual-activity prediction Transferboost model on undersampled data		
Learning rate	[0.00001,0.0001,0.0011, 0.01]	Could not train model
N estimators	[5,10,15,20,30,40,50,75,100,120,200]	
Max depth	[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,20]	
Dual-activity prediction Transferboost model on under- and oversampled data		
Learning rate	[0.00001,0.0001,0.0011, 0.01]	Could not train model
N estimators	[5,10,15,20,30,40,50,75,100,120,200]	
Max depth	[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,20]	
Dual-activity prediction TrAdaboost model on undersampled data		
Learning rate	[0.9,0.8,0.5,0.3,0.1]	0.1
N estimators	[30,50,100,120]	15
Dual-activity prediction TrAdaboost model on under- and oversampled data		
Learning rate	[0.9,0.8,0.5,0.3,0.1]	0.00001
N estimators	[30,50,100,120]	30
Dual-activity prediction TL-DNN model on undersampled data		
Weight decay		None
Epsilon		1e-07
beta_1		0.9
beta_2		0.999
learning_rate		0.001
ema_momentum		0.99
Dual-activity prediction baseline DNN model on undersampled data		
Weight decay		None
Epsilon		1e-07
beta_1		0.9
beta_2		0.999
learning_rate		0.01
ema_momentum		0.99
Dual-activity prediction TL-DNN model on under and oversampled data		
Weight decay		None
Epsilon		1e-07
beta_1		0.9
beta_2		0.999
learning_rate		0.001
ema_momentum		0.99
Dual-activity prediction baseline DNN model on under- and oversampled data		
Weight decay		None
Epsilon		1e-07
beta_1		0.9
beta_2		0.999
learning_rate		0.0001
ema_momentum		0.99

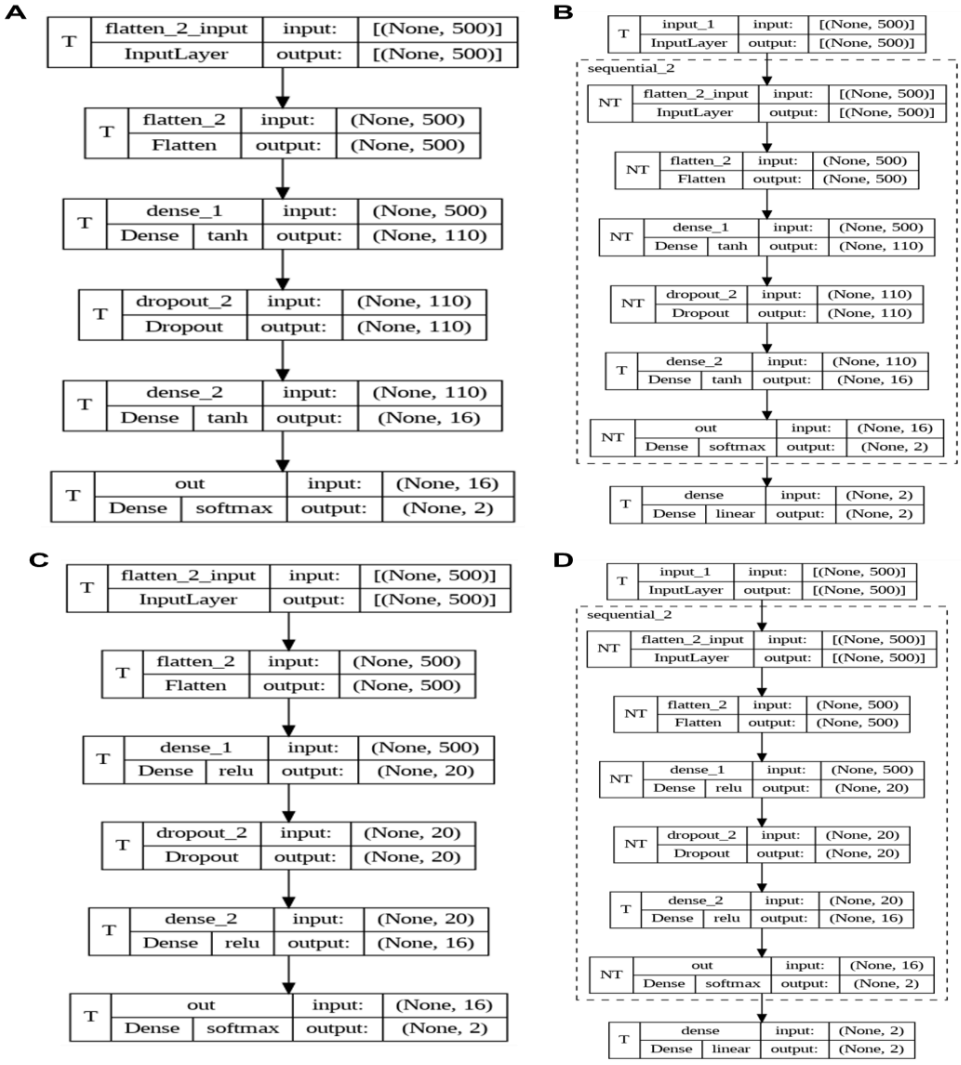


Figure 2.9: Pre-trained model and TL DNN models' architecture trained on either undersampled or under-and oversampled data using ECFP with 500 bit-length.
(A) Pre-trained model deep neural network architecture acquired after training on the balanced ABS database (undersampled) using 500-bit ECFP as training features, before (B) fine-tuning the last dense layer (dense_2) with 16 nodes in the DNN to the training set of the undersampled (inactive compounds) dual-active database. (C) Pre-trained model deep neural network architecture acquired after training on the ABS database (undersampled and oversampled) using 500-bit ECFP as training features, before (D) fine-tuning the dense layer (dense_2) with 16 nodes in the DNN to the training set of the undersampled (inactive compounds) and oversampled (active compounds) dual-active database.

When increasing the number of features within ECFP (from 100 to 500-bit length), a slight increase in recall could be observed within both TrAdaboost and the control DNN model, which was trained on undersampled data (Figure 2.10), however, there was no precision within the compound activity predictions between these models. Although TrAdaBoost did perform better than the baseline AdaBoost model in identifying active compounds, the overall model performance was very poor, especially considering the F1 and G-mean scores, which indicated the model suffered in optimising its' sensitivity and specifically its' specificity within imbalanced data. This indicated that TrAdaBoost cannot clearly distinguish dual-active compounds from inactive compounds, which can also be observed from the false positive predictions of TrAdaboost (Figure 2.10B). The parameter-based TL models trained

on undersampled data showed no positive TL between our TL and baseline DNN models. The baseline model obtained higher recall (even though slight), precision and G-mean scores than the TL-DNN model (Figure 10A). Based on the model predictions, it becomes apparent that the TL-DNN model classified all compounds as inactive, hence, the low recall and G-mean scores observed as the TL-models display low sensitivity in identifying dual-active compounds compared to the baseline DNN (Figure 2.10A & B).

To determine whether oversampling together with cluster-based undersampling would aid models in identifying dual-active/gametocytocidal compounds, the same models were trained on under- and oversampled data using ECFP at 500-bit length. In contrast to TL-DNN models trained on undersampled data, TL-DNN models trained on under-and oversampled data achieved high recall (Figure 2.10C), however, inspection of model predictions revealed the TL-DNN model classified all compounds as active (Figure 2.10D) hence the extremely high FPR observed. Compared to this, the baseline DNN model performed better in distinguishing inactive compounds from active, however, both TL and baseline DNN models displayed poor precision. TrAdaboost models trained on under-and oversampled data, showed slight improvement in precision at the expense of recall compared to models trained on undersampled data. Comparing the model to the baseline AdaBoost model, TL enabled better model performance concerning the G-mean score. This high G-mean score is most likely due to TrAdaboost's higher sensitivity in identifying active compounds. Nevertheless, the specificity of the model appears to be low as the FPR is also high (Figure 2.10C & D).

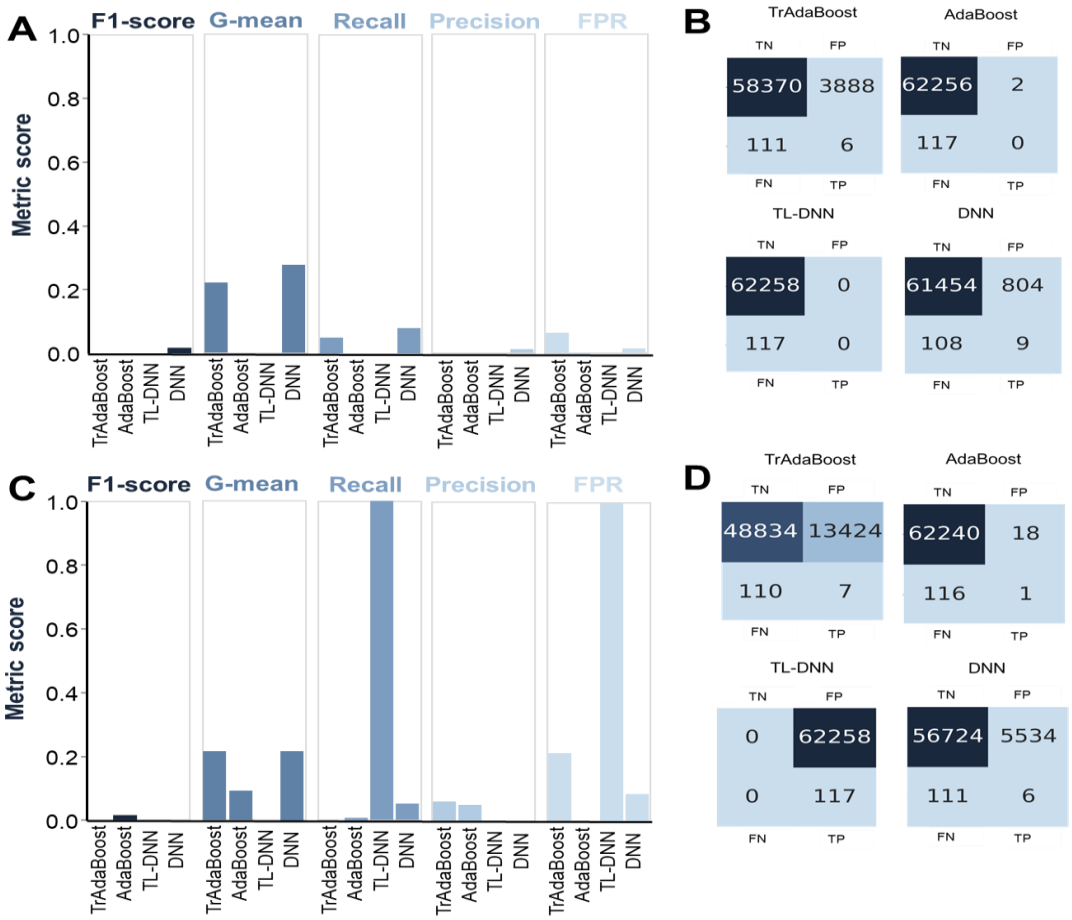


Figure 2.10: TL and baseline model performance on imbalanced test set when trained on 500-bit ECFP.
TL and baseline models trained on undersampled (A and B) or over- and undersampled data (C and D) were evaluated on their performance in predicting an imbalanced test data. (A and C) Model performance metrics are associated with the performance of the different models in predicting the imbalanced test set data. The F1-score evaluated model performance on imbalanced data, whereas G-mean scores determined how well models were able to optimize sensitivity and specificity. Recall and precision indicated the accuracy of activity predictions whereas false positive rate (FPR) indicated error within predictions. (B and D) Model activity predictions, i.e., true positive (TP), true negative (TN), false positives (FP) and false negatives (FN), were visualized as a confusion matrix for both TL and baseline models.

Overall, no discernible improvement could be observed in using TL models compared to baseline ML models for model building, however, it was undoubtedly clear that the performance of TL and baseline models was heavily impacted by the class imbalance present within the dual-active database and techniques such as undersampling and oversampling was not able to improve model performance. Rather than the limited dual-active phenotypic screening data available, the class imbalance was a more glaring issue when building models for dual-activity prediction.

2.4) Discussion and Conclusions

With the cost and time associated with gametocytocidal phenotypic screening, ML models predicting gametocytocidal activity would be a valuable pre-screening tool in reducing the cost and time directed towards screening inactive compounds. However, the training of such models is complicated due to the limited training data available and the class imbalance present within phenotypic screening data. To overcome these challenges in model building, we investigated techniques such as TL and class sampling to build robust gametocytocidal activity prediction models.

Cluster-based undersampling significantly reduced the class imbalance within our target (dual-active) database and completely balanced out the source (ABS) database. Unfortunately, these class imbalance correction techniques did not improve model performance for both TL and baseline models. Models trained on only undersampled data had poor to almost no recall ability and zero precision in identifying dual-active compounds, as most models classified all instances in the test set as inactive. This was true for instance and parameter-based TL methods, however, this was expected as class imbalance was still present within the undersampled data. A hybrid approach of cluster-based undersampling of inactive compounds and oversampling of gametocytocidal compounds allowed slightly better recall and precision within both instance and parameter-based TL models. Closer inspection of both model accuracy and activity predictions revealed that although the models had fair recall and precision, this was primarily due to models classifying all instances in the test set as active compounds or failing to distinguish active and inactive compounds.

Considering the undersampling and hybrid approach results, we theorised that the molecular descriptor used (ECFP) was not descriptive enough to allow effective pattern detection between inactive and active compounds. Unfortunately, increasing the number of chemical training features within our ECFPs, only a few slight improvements could be seen within the recall and precision ability of TL and baseline models. Despite our expectations, TL did not display better performance than traditional ML methods; in the case of parameter-based TL, the baseline DNN model mostly outperformed the TL-DNN model, indicating negative TL concerning recall and precision. Though TL improved model performance in other antimalarial compound activity prediction models, such as DeepMalaria [82], it should be noted that such models were built on phenotypic screening data of large chemical libraries

screened against ABS. This contrasts with our data setting, where there is limited phenotypic screening data conducted on gametocytes, with the class imbalance within these gametocytocidal phenotypic screens being more severe than observed in ABS screens.

Additionally, TL is mainly directed at improving model performance where training examples are limited and expects balanced target and source data for training. Though we could balance our ABS (source) database, we could not do this for our dual-active (target) database, which has limited training examples, and class imbalance is still present. TL techniques aimed at limited and imbalanced data remain an understudied niche within the ML community, and only recently have some studies tried to address such limitations within TL [151-153]. Nonetheless, at the time of this study, we could not further investigate such newly developed TL techniques. Interestingly, a recent study comparing traditional ML and TL algorithms within data domains where class imbalance is present reported that overall TL had poor performance compared to traditional ML models [154]. Due to our limited computing power, this study could not thoroughly investigate whether domain adaptation techniques such as JDA provided better performance than alternative TL techniques and remains an area that could still be further investigated.

Nonetheless, JDA would likely have similarly struggled with the class imbalance present in the dual-active database. Within other TL studies with similar data settings to ours, JDA was shown to perform poorly without additional design techniques to accommodate the severe class imbalance [154, 155]. Since class imbalance proved to be a more daunting challenge for TL and baseline ML models in predicting gametocytocidal compounds and for fear of losing chemical space information by further undersampling inactive compounds, we opted to interrogate ML algorithms more suited for training on imbalanced data.

CHAPTER 3

BUILDING ML MODELS CAPABLE OF PREDICTING COMPOUNDS WITH ABS AND DUAL-ACTIVITY

The work in this chapter has been published as follows:

Heerden, A.v., Turon, G., Duran-Frigola, M., Pillay, N., and Birkholtz, L.-M. (2023). Machine learning approaches identify chemical features for stage-specific antimalarial compounds. ACS Omega. 2023 Nov 7;8(46):43813-43826. doi: 10.1021/acsomega.3c05664. eCollection 2023 Nov 21.

3.1) Introduction

The previous chapter showed that class imbalance was a more challenging hurdle than the limited gametocytocidal phenotypic screening data available for model building. Unfortunately, further undersampling of inactive compounds may result in poor performance of models defining inactivity and increase false positives. In contrast, a hybrid approach of both undersampling and oversampling did not provide any benefit. Additionally, to class imbalance correction techniques, one can employ ML algorithms that are less sensitive to such class imbalances. Such ML algorithms can be those of random forest (RF) and gradient boosting machines (GBM), which have been shown to perform quite well when trained on imbalanced datasets by using bagging and boosting to improve model performance. Alternatively, to such algorithms, one can also use ML algorithms that set the weight of the minority and majority class to penalise the model more heavily for misclassifying the minority class to force the model to learn pattern detection for correctly predicting the minority class.

3.2) Methods

Note: the databases generated from Chapter 2 were re-used in Chapter 3 for the workflow below (Figure 3.1)

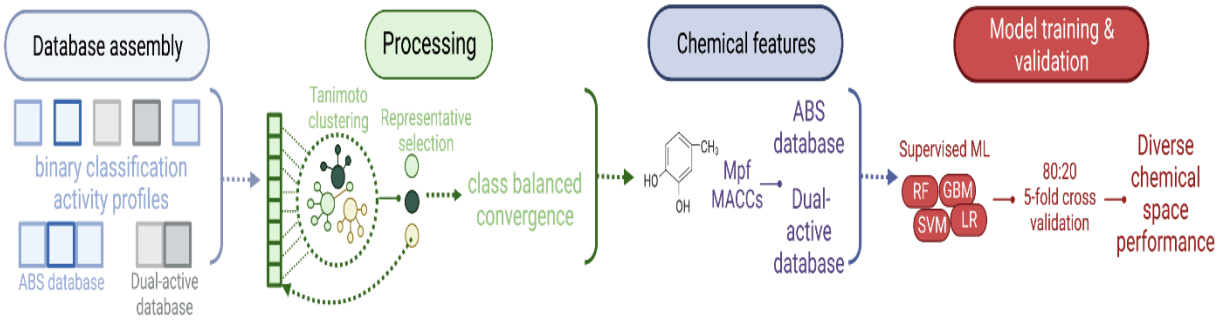


Figure 3.1: Workflow for building and validation of models trained on ABS balanced data and/or dual-active imbalanced data.
Database assembly (light blue), pre-processing and class sampling (green) was conducted as described in Chapter 2. In addition to ECFP as molecular descriptors for compounds, MACCS keys was also investigated (blue). After model training on 80% respective databases models were evaluated on cross-validation results and imbalanced test set predictions and then subsequently validated on external and diverse chemical libraries (red).

3.2.1) Selection of traditional ML algorithms suited for training on class imbalance datasets

For both ABS and dual-active chemical databases, two models were built for each of the algorithms used, one using ECFP and the other using MACCS as molecular descriptors. This would allow comparative evaluation between molecular descriptors to determine which was best suited to predict bioactivity in either ABS and/or gametocyte stages within representative and novel chemical spaces.

Before model building, the respective ABS and dual-active databases were randomly split into training and testing sets at an 80:20 ratio. Model training and hyperparameter tuning were conducted on 80% of the compounds. Each model underwent a grid search cross-validation hyperparameter tuning to identify the optimal hyperparameters. After that, hyperparameter-tuned models underwent 5-fold cross-validation to assess average accuracy and variability within model predictions. The resultant leftover 20% of compounds from the database were merged with the compounds that were excluded during the undersampling process to generate an imbalanced test set for model evaluation on untrained and imbalanced data.

To prevent class bias within models as a result of class imbalance being present, especially within the dual-active database even after undersampling, ML algorithms that performed well on imbalanced training data or applied weight-based penalties on the misclassification of minority classes (dual-active compounds) were nominated for model building from the

scikit-learn python package (version 0.20) [145]. Such models included ensemble methods such as random forest (RF) and gradient boosting machines (GBM), which have been shown to perform well on imbalanced data [156, 157]. Additionally, ML algorithms such as support vector machines (SVM) and logistic regression (LR) were also selected, as these scikit-learn algorithms could attribute weights to a minority class (active compounds) to heavily penalise the model for misclassifying active compounds.

ABS activity prediction models were trained on 80% (47 530) of the compounds from the balanced ABS database. During training, models needed to identify patterns within molecular fingerprints of compounds (ECFP or MACCS keys) that allowed for the correct prediction of compound bioactivity against ABS. For the models built using SVM, RF, GBM, and LR algorithms, the scikit-learn python package was used to train the model to the training set. The optimal hyperparameters were identified for the respective algorithms during model training, and the resultant fine-tuned models underwent subsequent 5-fold cross-validation. After cross-validation, models were evaluated on the imbalanced test set (61 029) excluded from training to assess models' bioactivity prediction accuracy on imbalanced untrained data and for any overfitting to the training data.

Similarly, models were trained on 80% (7 913 compounds) within the dual-active database for dual-activity prediction models. Likewise, scikit-learn was used for training RF and GBM models, however, for LR models, the class weight was additionally set as 1 for inactive compounds and 10 for active compounds to compensate for the class imbalance present within the dual-active database when training dual-activity prediction models on the training set. Similarly, the class weight was set to "balanced" for SVM models to adjust class weights inversely proportionally to the frequency of the class. Dual-activity prediction models were built using the optimal hyperparameters identified during training and underwent 5-fold cross-validation before model evaluation on the imbalanced test set (62 375 compounds) to judge model bioactivity prediction accuracy and overfitting.

3.2.2) Evaluating performance of different ML models on test set in predicting ABS and dual-active compounds

Fine-tuned models were evaluated on cross-validation and test set results to determine model performance on untrained imbalanced chemical data and to highlight overfitting within

models. Performance metrics used for model evaluation were recall (equation A), precision (B), false positive rate (C), receiver operator characteristic curve (ROC-AUC) and the F1-score (D).

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \qquad \text{(A)}$$

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \qquad \text{(B)}$$

$$\boldsymbol{FPR} = \frac{\boldsymbol{FP}}{\boldsymbol{FP+TN}} \qquad \text{(C)}$$

$$\boldsymbol{F1-score} = \boldsymbol{2} \times \frac{\boldsymbol{Precision \times Recall}}{\boldsymbol{Precision+Recall}} \qquad \text{(D)}$$

Recall and precision were calculated to determine if models can correctly predict active and inactive compounds. FPR was used to assess false positive predictions, and the ROC-AUC score determined how well the model could distinguish between the two classes. A score of 0.7 indicates the model has a 70% chance of ranking a randomly chosen active compound higher than a randomly chosen inactive compound The F1 score was also used as a performance measure, combining both recall and precision [158]. Models were also evaluated on their ability to optimise sensitivity (equation E below) and specificity (F), which were used to define the geometric mean (G) when predicting active and inactive compounds within novel chemical spaces.

$$\boldsymbol{Sensitivity} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \qquad \text{(E)}$$

$$\boldsymbol{Specificity} = \frac{\boldsymbol{TN}}{\boldsymbol{FP+TN}} \qquad \text{(F)}$$

$$\boldsymbol{G-mean} = \sqrt{\boldsymbol{Sensitivity} \times \boldsymbol{Specificity}} \qquad \text{(G)}$$

Due to the class imbalance present, especially within the dual-active database, the GHOST python package version 0.6.1 [159] was used to adjust probability thresholds for model decisions. This was to evaluate if adjusting the probability discrimination threshold resulted in better model performance on the test set to the abovementioned metrics. The impact of adjusting the probability discrimination threshold was visualised using the yellowbrick python package version 1.5 [160]. To validate that simplistic models were on par or even better than complex models such as neural networks, the top-performing model based on the above metrics was compared to more complex models generated via the autogluon python package version 0.8.2 [161] utilizing the same training and test set data.

3.2.3) Performance comparison of models trained on either imbalanced, undersampled or oversampled data

To validate that cluster-based undersampling resulted in better model performance than training on imbalanced data, model performance was compared to that of models on imbalanced data. Likewise, to justify that cluster-based undersampling was the ideal class imbalance correction technique compared to oversampling, the performance of models trained on cluster-based undersampled data was compared to models trained on oversampled data pre-processed via imblearn version 0.11.0 [143].

3.2.4) External validation of models on PRB and Pathogen Box

To examine the limits of the activity prediction models in novel chemical spaces and to externally validate the models, models were tested against new chemical libraries that included chemically diverse compounds. These chemical libraries were obtained from the open-source Medicines for Malaria Venture (MMV) Pandemic Response Box (PRB box) and the Pathogen Box, with recorded potent activity against various stages of the parasite validated by multiple research groups (www.mmv.org) [136, 140]. The compounds within the PRB and Pathogen box that were absent within the training or test sets were extracted and used for external model validation. The hit rate of the best-performing model was calculated and then compared to that of the chemical library and random selection. Here, the hit rate of the top 100 compounds, which was ordered according to model probability, was then compared to the hit rate of randomly selecting 100 compounds. To further assess whether the models helped in limiting the number of active compounds in the bottom 100, i.e., least likely to have activity or last to be screened via random selection, the hit rate for the bottom 100 was also calculated with the goal of having a lower hit rate than that of randomly screening. The rationale was that the top 100 would be the first compounds randomly selected to start with during phenotypic screening, whereas the bottom would be the last compounds chosen to screen. Additionally, to better validate compound prioritization within such chemically diverse spaces, the enrichment factor (EF) was calculated for the top 10 and top 50 compounds based on the predicted probability of compound activity.

3.3) Results

3.3.1) ECFP and hyperparameter analysis for best model performance

For ABS activity prediction models, the optimal bit-length that enabled better model performance was determined to be 500 (Figure 3.2A). No apparent change was observed within ROC-AUC values as the bit-length increased, whereas a drastic drop in FPR was observed at 500 bit-length, which then again increased as the bit-length increased. In contrast to FPR, precision peaked at 500 bit-length before dropping as the bit-length increased.

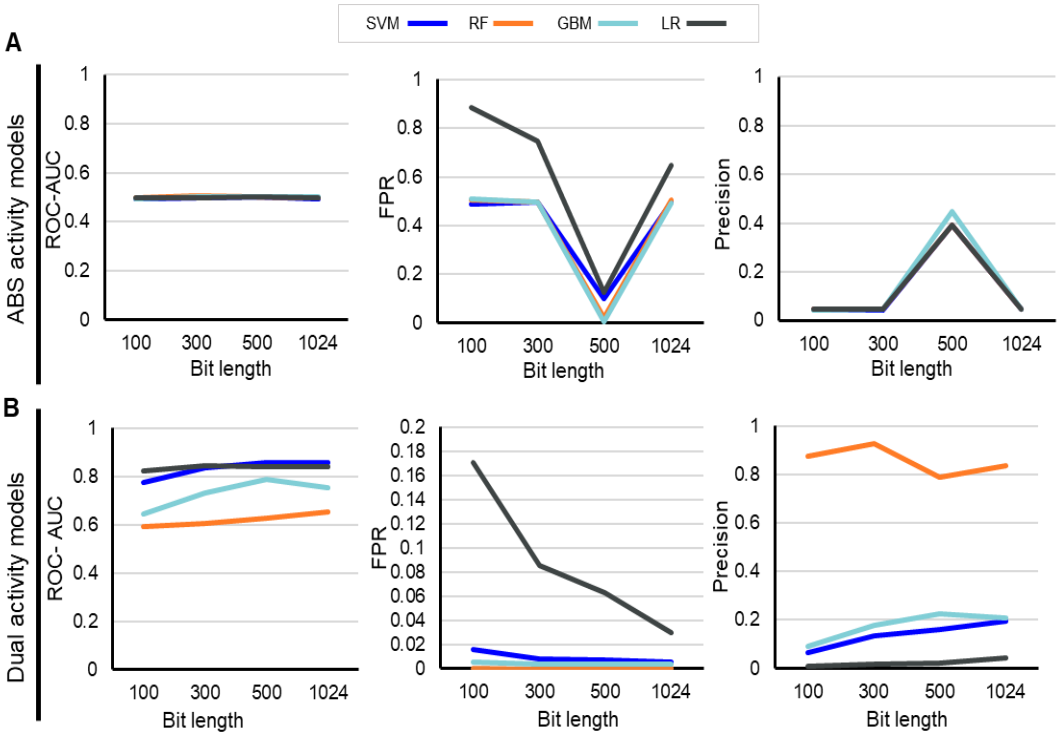


Figure 3.2: Determination of Morgan fingerprints (ECFP) bit-length that enabled better model performance. (A) Performance of ABS activity models using different ML algorithms (blue = Support vector machine, orange = Random Forest, light blue= gradient boosting machine, dark grey= logistic regression) in identifying compounds with ABS inhibition activity within test set using differing bit lengths of Morgan fingerprints (ECFP) during training at a set atom radius of 5. ROC AUC scores indicate the classifier's ability to distinguish active and inactive compounds against ABS. FPR scores indicate the false positive rate of test set predictions. (B) Performance of dual-active models (trained on data with class imbalance) using different ML algorithm's ability in identifying compounds with dual-activity within test set using differing bit lengths of Morgan fingerprints (ECFP) during training.

Within dual-activity prediction models, ROC-AUC values slightly increased as bit length increased, with GBM decreasing in ROC-AUC once a bit-length larger than 500 was used (Figure 3.2B). Like ABS models, FPR values decreased with increasing bit-length, however, models plateaued at 300 bit-length (excluding LR). Precision also tended to increase with increasing bit-length. Based on ROC-AUC, FPR, precision metrics, and the training time required to build models, a bit-length of 500 was determined to be the best bit length to use for ECFP whilst training most models.

The optimal atom radius for ABS and dual-activity prediction models using a bit-length of 500 was determined to be a radius of five, as this increased the precision and G-mean scores of the models whilst not impacting ROC-AUC values (Figure 3.3).

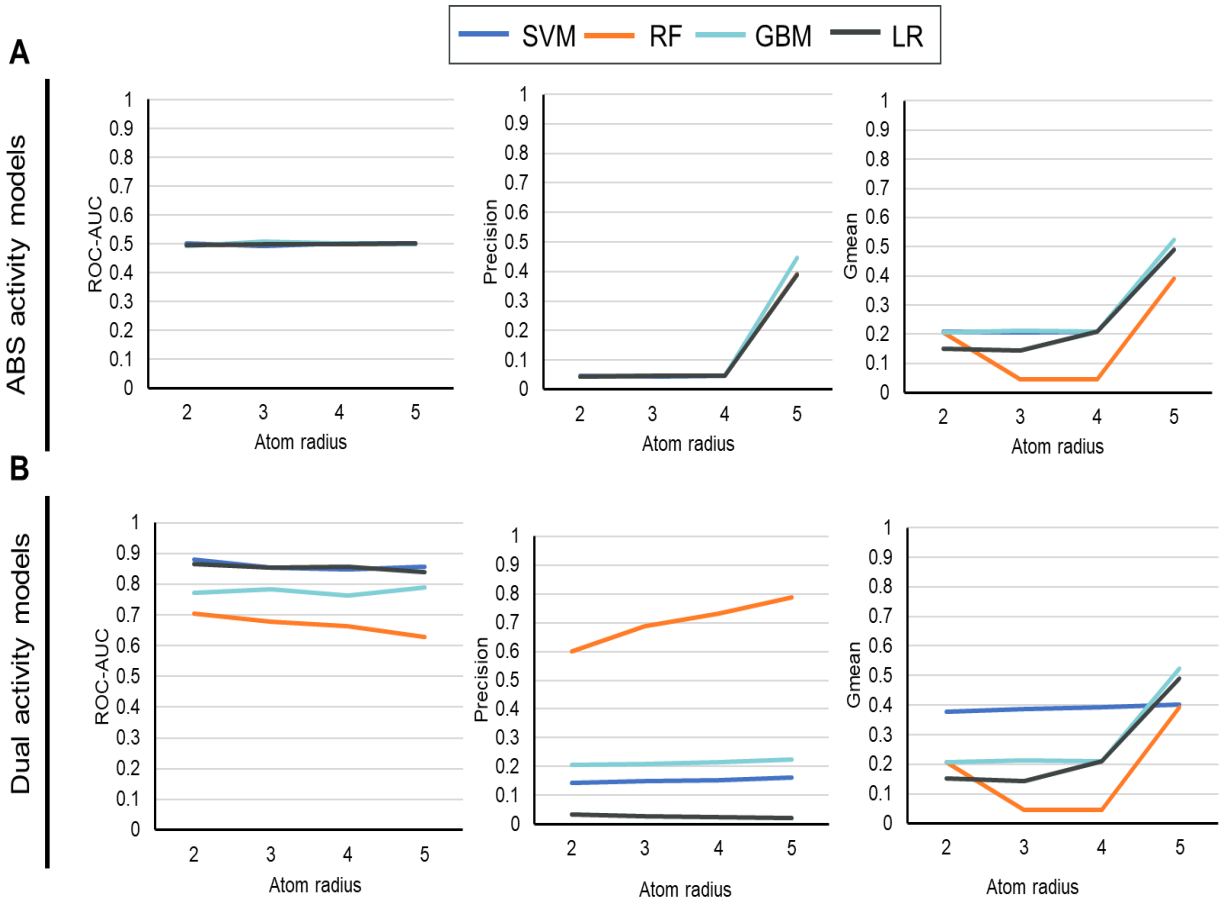


Figure 3.3: Determination of Morgan fingerprints (ECFP) atom radius that enabled better model performance. (A) Performance of ABS activity models using different ML algorithms (blue = Support vector machine, orange = Random Forest, light blue= gradient boosting machine, dark grey= logistic regression) in identifying compounds with ABS inhibition activity within test set using differing atom radius of Morgan fingerprints (ECFP) at 500-bit length during training. ROC AUC scores indicate the classifier's ability to distinguish active and inactive compounds against ABS. FPR scores indicate the false positive rate of test set predictions. (B) Performance of dual-active models (trained on data with class imbalance) using different ML algorithms' ability in identifying compounds with dual-activity within test set using differing atom radius of Morgan fingerprints (ECFP) at 500-bit length during training.

Additionally, an atom radius of five would also aid in identifying substructures during feature analysis; hence, a bit length of 500 with an atom radius of five was selected when generating the ECFP of compounds. Using these parameters for ECFP generation, the optimal hyperparameters were identified for each of the respective algorithms during model training (Table 3.1).

Table 3.1: Hyperparameter tuning and optimal parameters identified for ML models suited for training on imbalanced data

Parameter tuned	Parameters (range) used	Optimal parameter (ECFP)	Optimal parameter (MACCS)
Dual-activity prediction SVM model			
Kernels:	Polynomial, RBF, Sigmoid, Linear	RFB	RFB
Regularization parameter (C)	[0.1, 1, 10, 100, 1000] or default	10	1000
Kernel coefficient (gamma)	[1, 0.1, 0.01, 0.001, 0.0001] or default	0.01	0.01
ABS activity prediction SVM model			
Kernels:	Polynomial, RBF, Sigmoid, Linear	RFB	RFB
Regularization parameter (C)	[0.1, 1, 10, 100, 1000] or default	default	10
Kernel coefficient (gamma)	[1, 0.1, 0.01, 0.001, 0.0001] or default	default	0.1
GBM dual-activity prediction model			
Number of trees	[10, 50, 100, 500]	500	500
Subsample	[0.1, 0.01, 0.001]	0.1	0.1
Max features	1-10	5	-
Learning rate	[1, 0.1, 0.01, 0.001, 0.0001]	0.1	0.01
Max tree depth	1-10	2	9
ABS activity prediction GBM model			
Number of trees	[10, 50, 100, 500]	100	500
Subsample	[0.1, 0.01, 0.001]	0.1	0.1
Max features	1-7	29	-
Learning rate	[1, 0.1, 0.01, 0.001, 0.0001]	0.1	0.01
Max tree depth	1-10	2	9
RF dual-activity prediction model			
Max depth	[10- 15]	14	14
Max features	['auto', 'log2']	auto	auto
Number of estimators	[5, 6, 7, 8, 9, 10, 11, 12, 13, 15]	15	10
ABS RF activity prediction model			
Max depth	10-15	14	14
Max features	['auto', 'log2']	auto	auto
Number of estimators	[5, 6, 7, 8, 9, 10, 11, 12, 13, 15]	15	15
LR dual-activity prediction model			
C-value	[100, 10, 1.0, 0.1, 0.01]	100	100
Solvers	['newton-cg', 'lbfgs', 'liblinear']	lbfgs	lbfgs
Penalty	L2	L2	L2
ABS LR activity prediction model			
C-value	[100, 10, 1.0, 0.1, 0.01]	10	100
Solvers	['newton-cg', 'lbfgs', 'liblinear']	lbfgs	newton-cg
Penalty	L2	L2	L2

3.3.2) Asexual blood stage activity prediction models performance on the test set

A two-pronged approach was used to identify chemical features of compounds associated with activity against ABS (or lack thereof) using ML: SMILES of compounds from the ABS database were transformed into either ECFP or MACCS molecular fingerprints. Afterwards, four different models were trained on 80% of the data from each featurization method. To determine if training on cluster-based undersampled data enabled better model performance on the imbalanced test set than the default database or oversampling, we compared model performance using the same algorithms trained on oversampling or on imbalanced data without pre-processing. The models trained on undersampled data from the ABS database displayed higher sensitivity but similar specificity compared to models

trained on oversampled or severely imbalanced data, with higher G-mean and ROC-AUC scores for the undersampled data (Table 3.2). This highlighted that the models trained on oversampled data failed to identify novel compounds due to the model possibly fixating on patterns associated with the oversampled active compounds in the training data [162]. It was therefore concluded that cluster-based undersampling provided a better class imbalance correction technique and improved model sensitivity within ABS prediction models compared to training models on oversampled data or imbalanced data.

Table 3.2: Comparison of ABS activity prediction models' performance on test set when trained on either undersampled, oversampled or imbalanced training data

Model	G-Mean	FPR (FP/FP+TN)	ROC-AUC	Sensitivity (TP/TP+FN)	Specificity (TN/TN+FP)	F1-Score
ABS activity prediction model trained on imbalanced data						
SVM (ECFP)	0.599	0.098	0.65	0.397	0.902	0.47
RF (ECFP)	0.586	0.14	0.629	0.399	0.86	0.438
GBM (ECFP)	0.461	0.057	0.584	0.226	0.943	0.323
LR (ECFP)	0.7	0.489	0.735	0.96	0.511	0.56
SVM (MACCS)	0.574	0.076	0.64	0.356	0.924	0.45
RF (MACCS)	0.629	0.151	0.657	0.466	0.849	0.484
GBM (MACCS)	0.549	0.088	0.621	0.33	0.912	0.413
LR (MACCS)	0.711	0.469	0.742	0.952	0.531	0.563
ABS activity prediction model trained on oversampled data						
SVM (ECFP)	0.685	0.206	0.693	0.591	0.794	0.53
RF (ECFP)	0.633	0.188	0.652	0.493	0.812	0.474
GBM (ECFP)	0.695	0.185	0.704	0.592	0.815	0.546
LR (ECFP)	0.575	0.665	0.661	0.988	0.335	0.486
SVM (MACCS)	0.788	0.292	0.792	0.875	0.708	0.629
RF (MACCS)	0.685	0.22	0.69	0.601	0.78	0.525
GBM (MACCS)	0.763	0.306	0.766	0.839	0.694	0.6
LR (MACCS)	0.572	0.668	0.659	0.987	0.332	0.485
ABS activity prediction model trained on undersampled balanced data						
SVM (ECFP)	0.875	0.149	0.875	0.899	0.851	0.359
RF (ECFP)	0.825	0.178	0.825	0.828	0.822	0.298
GBM (ECFP)	0.802	0.219	0.802	0.824	0.781	0.257
LR (ECFP)	0.708	0.483	0.743	0.969	0.517	0.161
SVM (MACCS)	0.876	0.168	0.877	0.922	0.832	0.339
RF (MACCS)	0.853	0.162	0.853	0.868	0.838	0.331
GBM (MACCS)	0.725	0.459	0.757	0.972	0.541	0.168
LR (MACCS)	0.865	0.165	0.865	0.896	0.835	0.335

Using the undersampled ABS database for ABS activity prediction models trained on ECFPs of compounds, the SVM model achieved the highest ROC-AUC score with the lowest variability (0.99 ± 0.02) during 5-fold cross-validation (Figure 3.4A). Additionally, the SVM model maintained such accuracy when predicting compound ABS inhibition activity on

untrained imbalanced test data (ROC-AUC score 0.92, Figure 3.4B), showing no overfitting of the model to training data.

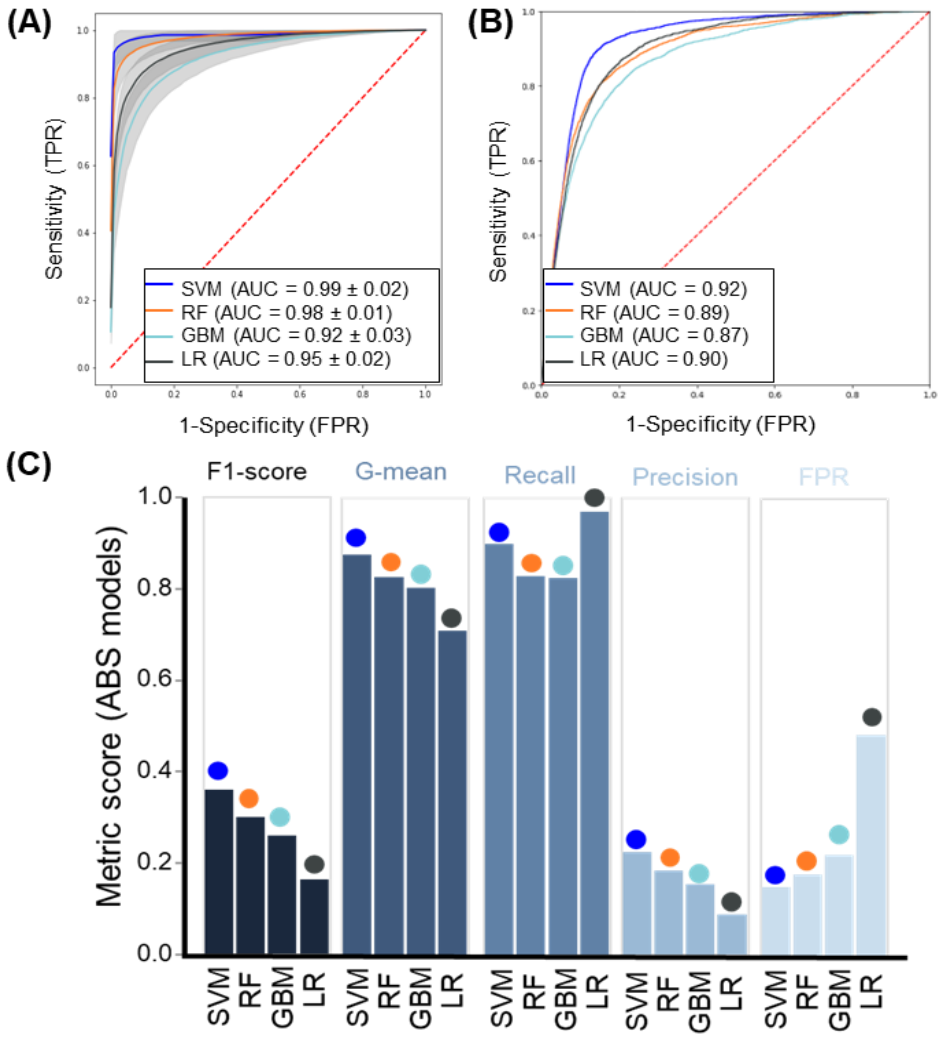


Figure 3.4: Performance of different conventional ML algorithms in identifying compounds with ABS activity. (A) ROC-AUC curves showing the performance of different ML algorithms in predicting compounds with ABS activity when trained on the ECFP of compounds after 5-fold cross-validation. Insert indicates AUC mean values ± standard deviation. (B) ROC-AUC curves showing the performance of different ML algorithms on the imbalanced test set. (C) Model performance metrics are associated with the performance of the different models in predicting the imbalanced test set data. The F1-score evaluated model performance on imbalanced data, whereas G-mean scores determined how well models were able to optimize sensitivity and specificity. Recall and precision indicated the accuracy of activity predictions whereas false positive rate (FPR) indicated error within predictions.

Further evaluation of model performance metrics showed that the SVM models' recall ability (at 0.90) and precision (at 0.22) were comparable to that of the other ensemble models (RF, GMB; Figure 3.4B), with only LR models obtaining a higher recall (0.97) than SVM. Interrogation of the LR models, however, indicated that this high recall within LR models came at the expense of the model's precision (0.08), explaining the low specificity with a higher false positive rate within LR models (FPR: 0.43 vs. <0.22 for the other models) (Figure

3.4B & C). This suggested that the model derived from SVM for ABS activity prediction is more precise and capable of identifying compounds with ABS activity, all while limiting its' false positive rate (FPR: 0.15) (Figure 3.4C).

Like models trained using ECFPs, SVM models trained with MACCS keys of compounds also achieved the highest ROC-AUC scores on both 5-fold cross-validation (0.99 ± 0.02, Figure 3.5A) and untrained test data (0.92, Figure 3.5B). No significant differences could be detected between the performance of models trained on ECFPs or MACCS keys beyond a slight improvement in performance metrics for LR and RF models (Table 3.2). Due to the more descriptive nature of ECFPs compared to MACCS, such molecular descriptors may better highlight chemical features associated with ABS inhibition activity; thus, ABS prediction models trained on ECFPs were further evaluated.

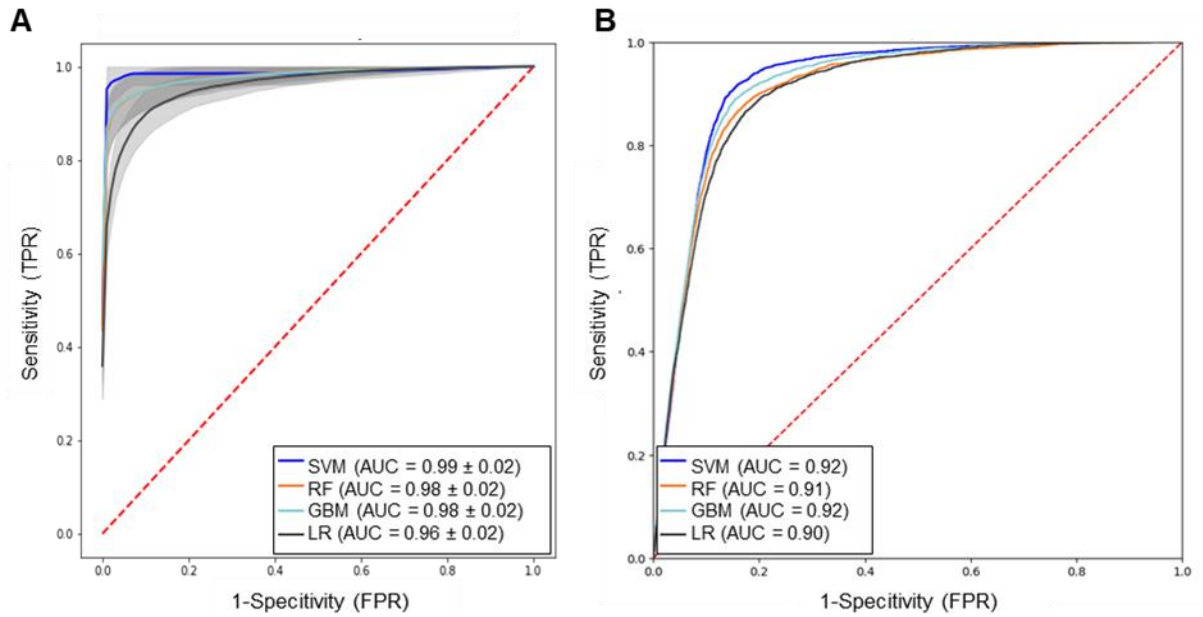


Figure 3.5: ROC-AUC curves from cross-validation and imbalanced test set performance of ABS activity prediction models trained on MACCS.
ROC-AUC curves showing the performance of different ML algorithms in predicting compounds with ABS inhibition activity or inactivity (A) when trained on MACCS keys of compounds after 5-fold cross-validation. Insert indicates AUC mean values ± standard deviation. (B) The ROC-AUC curves indicating model performance of the different models trained on MACCS descriptors upon untrained test set in predicting ABS inhibition activity or inactivity

Within models trained on ECFPs, further improvement could be obtained in lowering the FPR by shifting the discrimination probability, which was associated with gains in precision to as high as 0.675 at higher probability thresholds (Table 3.3).

Table 3.3: Optimised probability threshold for ABS activity prediction models trained on ECFP of compounds

Model	G-Mean	FPR	ROC-AUC	Recall	Precision	Probability threshold
ABS activity prediction model trained on undersampled balanced data						
SVM (ECFP)	0.875	0.149	0.875	0.899	0.224	0.5
RF (ECFP)	0.825	0.178	0.825	0.828	0.182	0.5
GBM (ECFP)	0.802	0.219	0.802	0.824	0.152	0.5
LR (ECFP)	0.708	0.483	0.743	0.969	0.088	0.5
SVM (ECFP)	0.714	0.468	0.745	0.958	0.675	0.8
RF (ECFP)	0.702	0.415	0.714	0.842	0.673	0.64
GBM (ECFP)	0.706	0.071	0.733	0.537	0.267	0.74
LR (ECFP)	0.822	0.140	0.823	0.785	0.212	0.96

Although increasing the discrimination probability threshold improved the models' precision (Figure 3.6A), when increasing this threshold past 0.90, a drastic loss in recall from 0.9 to <0.5 can be observed within the SVM model. Ensemble models such as RF and GBM showed a similar trend when adjusting the discrimination probability threshold, such as shifting the threshold past 0.5 resulted in a gradual loss of recall and increase in precision, however, this recall fell sharply once moving past 0.8.

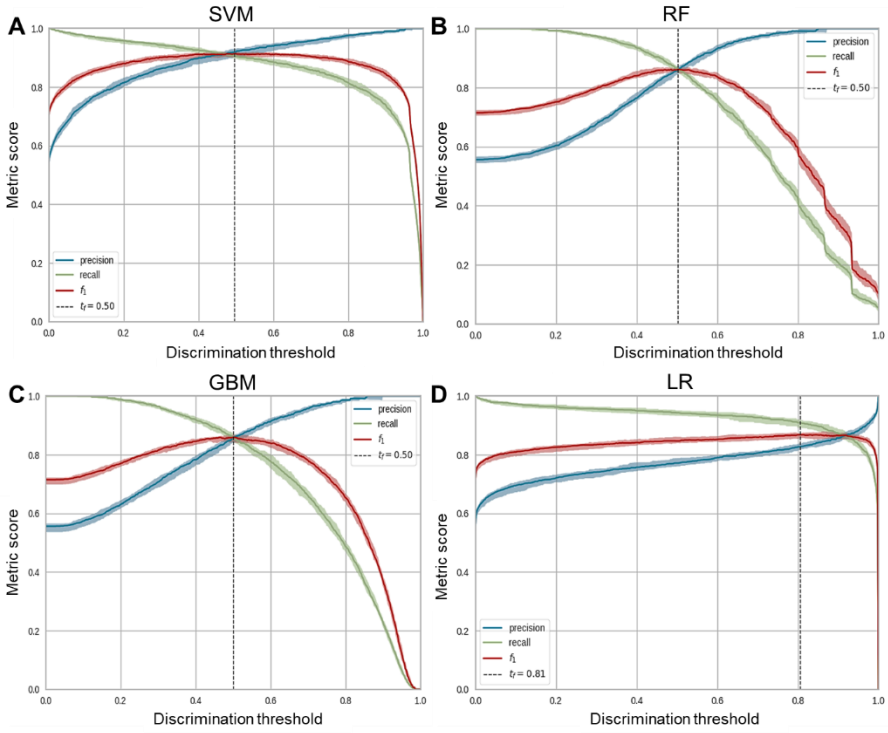


Figure 3.6: Influence of discrimination threshold adjustment on ABS model performance within the untrained test set.

(A) SVM, (B) RF, (C) GBM and (D) LR model performance regarding precision, recall and f1-score was calculated and plotted for each threshold defining active and inactive compounds from the predicted probability, i.e., discrimination threshold. Discrimination threshold adjustment was conducted on test set data with ABS activity models. Tf indicated the threshold at which both recall, and precision was the highest.

To determine whether such traditional models were on par with that of more complex models such as DNN, which has been used in other fields for bioactivity prediction, we compared the performance of our top two ABS activity prediction models to that of more complex models trained on the same training data using the same ECFP (500 bit-length, 5 atom radius). Interestingly, SVM showed similar, if not better, model performance than models such as NeuralNetFastAI (Table 3.4) trained on the same training set and was better at reducing its FPR than such complex models.

Table 3.4: Model performance comparison on test data of best-performing ABS activity prediction models and more complex models

Model	G-Mean	FPR (FP/FP+TN)	ROC-AUC	Recall	Precision	F1-Score
SVM (ECFP)	0.875	0.149	0.875	0.899	0.224	0.359
RF (ECFP)	0.825	0.178	0.825	0.828	0.182	0.298
NeuralNetFastAI	0.863	0.188	0.864	0.917	0.189	0.313
WeightedEnsemble_L2	0.877	0.163	0.877	0.918	0.213	0.345

3.3.3) Asexual blood stage activity prediction models performance on the external validation dataset

SVM and RF models were exposed to previously unseen chemical matter and data from curated datasets to validate further and interrogate the robustness of the best ABS activity prediction models. The inherent chemical diversity within these external datasets served as an additional level of interrogation that allowed us to evaluate how well the models would perform when exposed to very diverse and novel chemical spaces and can indicate what to expect of these models in real-world application. Data from the MMV PRB and Pathogen box were used, which comprised unique compounds not included in the data from the chemical libraries used to train models and activity data available against both *P. falciparum* ABS parasites and gametocytes [136, 140]. The compounds within these boxes were individually distinct and chemically diverse, providing additional datasets to validate the robust nature of the models compared to the larger databases used to train the models, where structurally related compounds were present within a chemical space (Figure 3.7A and D). The best ABS activity prediction model must thus maintain fair accuracy and recall under these conditions while also optimizing sensitivity and specificity in predictions to limit the models' FPR for these novel and diverse chemical datasets. Additionally, to the chemical diversity of the dataset, the inherent class imbalance was also not corrected to evaluate the

performance of the models, with observed hit rates against ABS being 18% for the PRB box [136] and 31% for the Pathogen Box [140] (Figure 3.7A and D). The top-performing ABS activity prediction models (SVM and RF) obtained similar scores for most metrics on the PRB box (Figure 3.7B), with the RF model leading in F1 and G-mean scores while limiting its' FPR. When considering the enrichment factor of the top 10 (3.35 vs. 1.52) and top 50 (2.24 vs. 2.15) compounds, RF outperforms the SVM (Figure 3.7C) in prioritizing ABS active compounds within the PRB box. Both models showed significant enrichment of hits (hit rates >30%) in the top 100 compounds, compared to random selection (17%) and the hit rate of the PRB box itself (18%) (Figure 3.7C). Model specificity was also confirmed as both models shifted the selection of compounds that would be inactive towards the bottom 100.

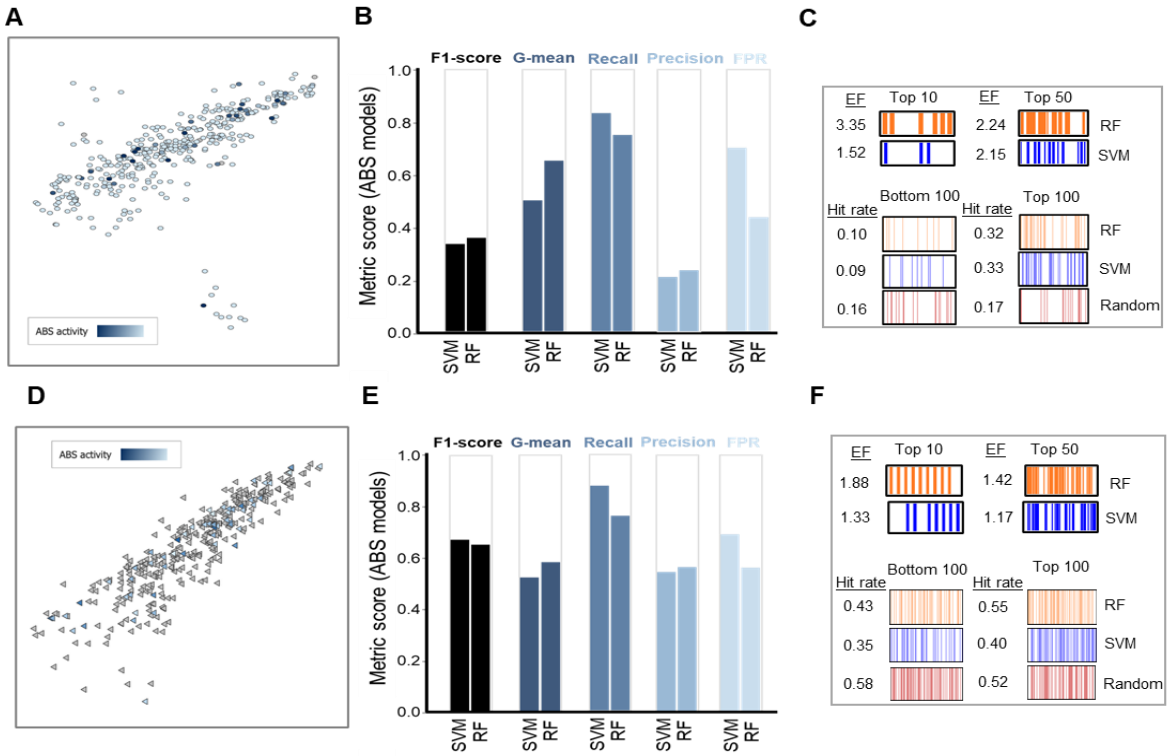


Figure 3.7: Model performance of different ML algorithms in identifying compounds with ABS inhibition activity within novel diverse chemical spaces.
Model robustness was evaluated by exposing the top two ABS activity prediction models to extreme datasets from the PRB box (A) and Pathogen box (D) displayed in the context of the launched drug chemical space (available on StarDrop v 7.3.0), with heat bars indicating potency. Performance of ABS activity prediction models trained on ECFP descriptors within the PRB box (B) and the Pathogen box (E) for F1-scores (model performance exposed to imbalanced data) G-mean scores (ability to optimize sensitivity and specificity), recall, precision, and false positive rate (FPR). The hit rate of the best-performing model within these chemical spaces (C and F) compared to random selection. The enrichment factor (EF) of calculated for the top 10 and top 50 compounds to determine model efficacy.

Within a different novel and diverse chemical space (Pathogen Box, Figure 3.7D), both SVM and RF ABS activity models obtained higher F1-scores (Figure 3.7E) compared to the PRB box (Figure 3.7B), and this could be attributed to the higher hit rate within the Pathogen Box

resulting in a less severe class imbalance compared to the PRB box. Regardless, both models maintained their recall performance within these different chemical spaces (Figure 3.7E), with RF again obtaining a slightly higher G-mean score and lower FPR than SVM. Both models also enabled the enrichment of hits for ABS inhibition activity within the top 10 and 50 compounds, however, RF seemed better suited to doing this compared to SVM (Figure 3.7F).

To ensure we were not biased in our model selection by only considering ABS activity prediction models trained on ECFP, we also evaluated the performance of models trained on MACCS keys on these extreme external datasets. Most strikingly, models using MACCS tended to have higher FPR and lower specificity in predictions compared to models using ECFP (Table 3.5). Additionally, MACCS models tended to have slightly lower G-mean and F1 scores. This could result from MACCS keys not being as descriptive as ECFP to allow better evaluation of compounds within such novel and diverse chemical spaces. From this result, we could conclude that ECFP allowed better evaluation and prediction of compounds within novel chemical spaces compared to MACCS keys.

Table 3.5: Comparison of ABS activity prediction models' performance within novel diverse chemical spaces

Model	G-Mean	FPR (FP/FP+TN)	Sensitivity (TP/TP+FN)	Specificity (TN/TN+FP)	F1-Score	Accuracy
ABS inhibition activity prediction model on PRB box						
SVM (ECFP)	0.499	0.701	0.833	0.299	0.331	39.5
RF (ECFP)	0.650	0.437	0.750	0.563	0.356	51.3
GBM (ECFP)	0.547	0.585	0.722	0.415	0.329	47.0
LR (ECFP)	0.327	0.890	0.972	0.110	0.323	26.5
SVM (MACCS)	0.432	0.784	0.861	0.216	0.317	33.0
RF (MACCS)	0.446	0.748	0.792	0.252	0.302	34.0
GBM (MACCS)	0.007	0.999	0.059	0.001	0.313	35.3
LR (MACCS)	0.255	0.933	0.972	0.067	0.313	23.0
ABS inhibition activity prediction model on Pathogen box						
SVM (ECFP)	0.521	0.689	0.876	0.311	0.670	58.3
RF (ECFP)	0.581	0.558	0.763	0.442	0.646	59.7
GBM (ECFP)	0.536	0.600	0.718	0.400	0.608	55.3
LR (ECFP)	0.215	0.953	0.977	0.047	0.652	49.6
SVM (MACCS)	0.493	0.726	0.887	0.274	0.665	56.9
RF (MACCS)	0.532	0.626	0.757	0.374	0.623	55.9
GBM (MACCS)	0.511	0.663	0.774	0.337	0.623	54.8
LR (MACCS)	0.296	0.911	0.977	0.089	0.662	51.8

Considering the extreme datasets these models were subjected to, it was expected that the precision of these models in such chemical spaces would be low. Regardless, to confirm that the molecules within these extreme datasets were indeed very dissimilar compared to the training data used for model building, five compounds were randomly selected from the test set as well as the PRB box for comparison on chemical similarity to the training data and their respective prediction accuracy within models. Comparing the five compounds selected from the PRB box to the training set (Figure 3.8) revealed that the chemical similarity distribution and average similarity of compounds over the training data is lower than that of the test set compounds selected.

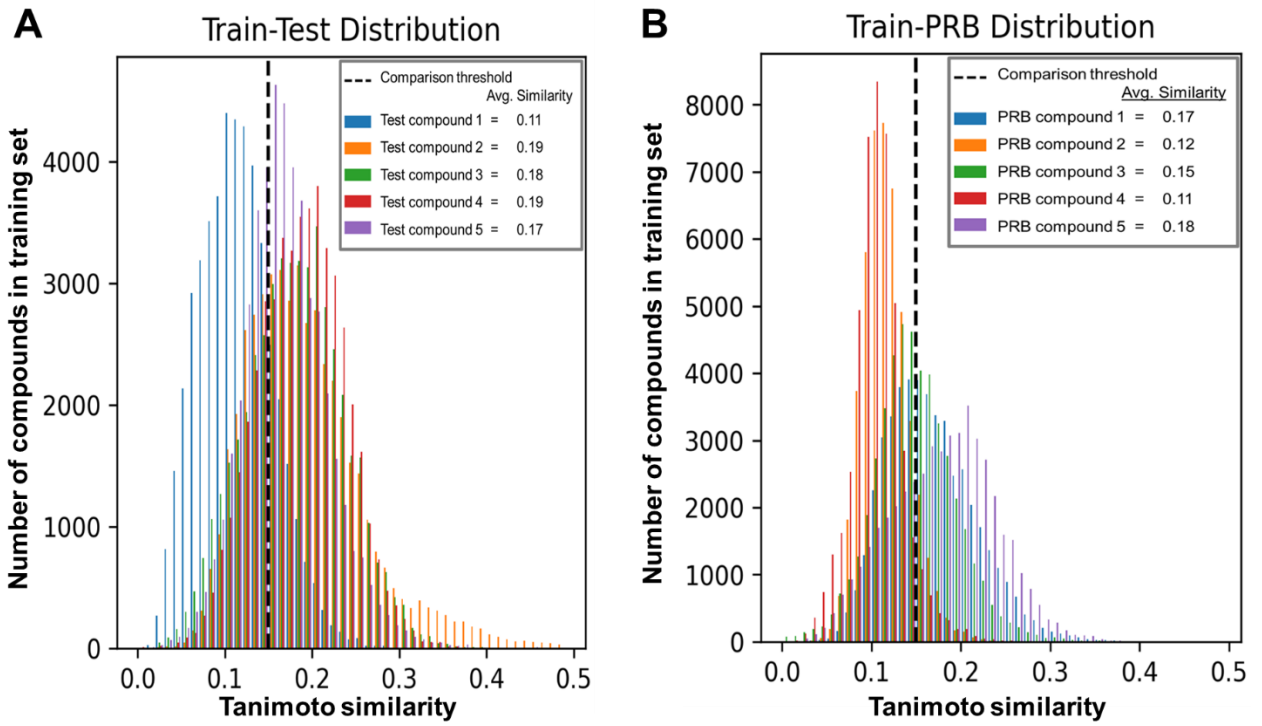


Figure 3.8: Tanimoto similarity distribution of PRB box or test set compounds on ABS model training set. Tanimoto similarity distribution plots were generated for each of the five compounds randomly selected from the (A) ABS test set and PRB Box (B) based on their structural similarity to the training set used for the ABS activity prediction models. Average Tanimoto similarity of compounds over the training set are indicated within the label The predicted activity of the randomly selected compounds is shown in Table 3.6.

Error in predicting activity of compounds tended to occur when compounds showed low similarity to the training data. For example, PRB compound 2 and test set compound 1, for both SVM and RF models, were misclassified as inactive, and for both these compounds, the chemical similarity distribution towards the training data falls below 0.15. This may suggest that some of these compounds fall outside of the model's applicability domain, lowering the precision in correctly identifying ABS active compounds (Figure 3.7, Table 3.6).

Interestingly, these models performed surprisingly well when exposed to external datasets despite training data (ABS database, Table 2.2) containing chemical libraries that varied in their activity cut-offs which was a point of concern in inadvertently introducing noise into the data.

Table 3.6: Activity predictions of compounds with low chemical similarity to ABS model training set

Compound	Activity	Predicted activity	Predicted probability
ABS activity SVM model			
Test set compound 1	ABS activity/Hit	Inactive	0.004
Test set compound 2	ABS activity/Hit	Active	1.000
Test set compound 3	ABS activity/Hit	Active	0.985
Test set compound 4	ABS activity/Hit	Active	0.991
Test set compound 5	ABS activity/Hit	Active	0.991
PRB compound 1	ABS activity/Hit	Active	1.000
PRB compound 2	ABS activity/Hit	Inactive	0.640
PRB compound 3	ABS activity/Hit	Active	0.920
PRB compound 4	ABS activity/Hit	Active	0.987
PRB compound 5	ABS activity/Hit	Active	0.996
ABS activity RF model			
Test set compound 1	ABS activity/Hit	Inactive	0.121
Test set compound 2	ABS activity/Hit	Active	0.780
Test set compound 3	ABS activity/Hit	Active	0.795
Test set compound 4	ABS activity/Hit	Inactive	0.297
Test set compound 5	ABS activity/Hit	Inactive	0.233
PRB compound 1	ABS activity/Hit	Active	0.615
PRB compound 2	ABS activity/Hit	Inactive	0.420
PRB compound 3	ABS activity/Hit	Inactive	0.455
PRB compound 4	ABS activity/Hit	Inactive	0.423
PRB compound 5	ABS activity/Hit	Active	0.883

3.3.4) Dual-activity prediction models performance on test set

Considering the success of generating ABS activity prediction models with fair performance within novel chemical spaces, we further wanted to investigate whether these same algorithms could be applied and provide useful dual-activity prediction models with good performance when trained on our imbalanced dual-active database. Like the previous section, the same model building pipeline was followed, where SMILES of compounds from the dual-active database were transformed into ECFP or MACCS molecular fingerprints. Subsequently, four different models were each trained on 80% of the data from each featurization method, respectively.

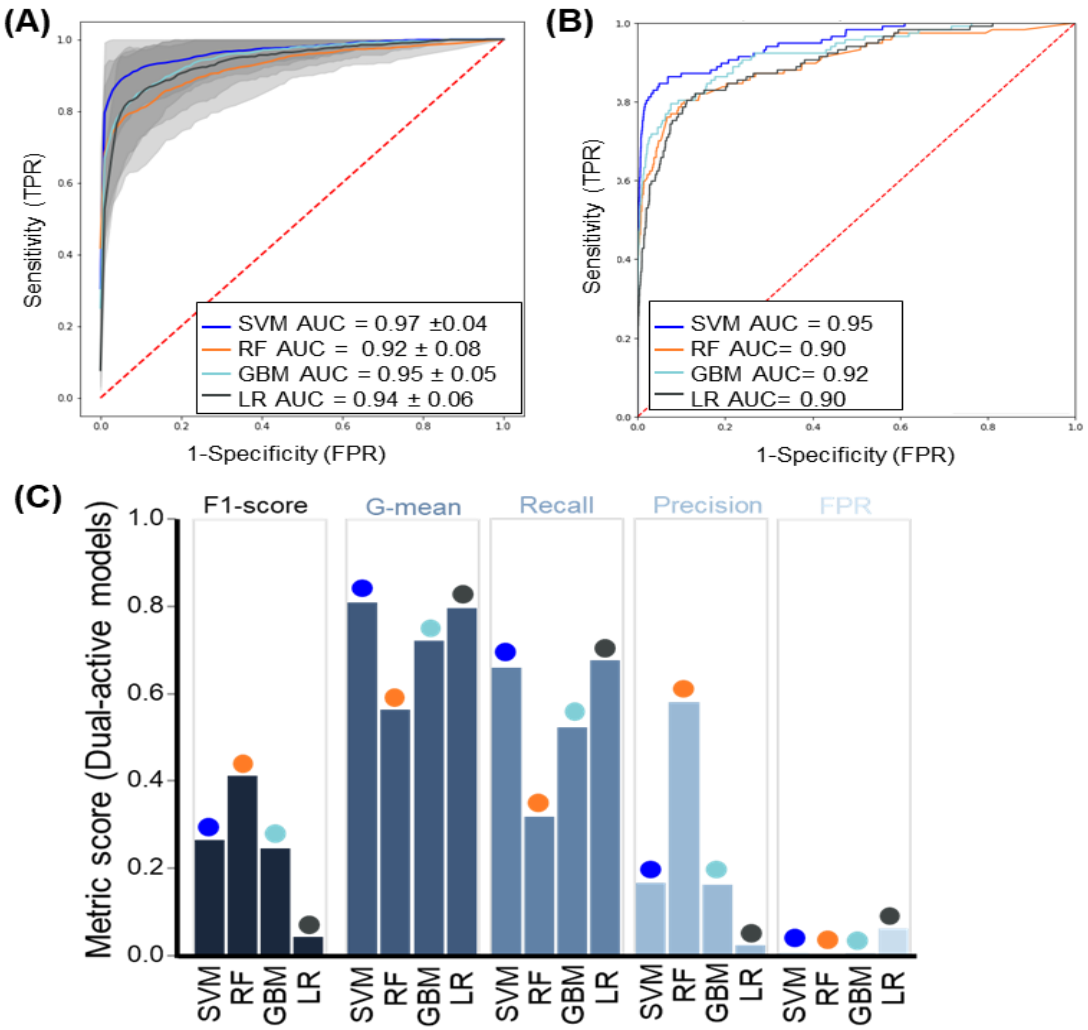


Figure 3.9: Performance of different conventional ML algorithms in identifying compounds with dual-activity. (A) ROC-AUC curves showing 5-fold cross-validation performance of different ML algorithms in predicting compounds with dual-activity when trained on the ECFP of compounds. Insert indicates AUC mean values ± standard deviation. (B) ROC-AUC curves showing the performance of different ML algorithms on the imbalanced test set. (C) Model performance metrics associated with the performance of the different models in predicting imbalanced test set data. The F1-score evaluated model performance on imbalanced data, whereas G-mean scores determined how well models were able to optimize sensitivity and specificity. Recall and precision indicated the accuracy of activity predictions whereas false positive rate (FPR) indicated error within predictions.

Due to the class imbalance still present within the dual-active database, even after undersampling, different metrics were used to accurately evaluate model performance in predicting compounds with dual-activity. Hence, during model evaluation, emphasis was placed on recall and precision in identifying dual-active compounds. SVM, trained on ECFP, outperformed other models within 5-fold cross-validation by obtaining ROC-AUC means >0.96 (Figure 3.9A). This extended to the performance of the models against imbalanced test data, where SVM reached an ROC-AUC score of 0.95 for models trained on ECFP (Figure 3.9B), indicating no overfitting of models. To ensure undersampling was the best class imbalance correction technique and improved model performance, models trained on

undersampled data were compared to models trained on oversampled data or in the case where imbalance was unaddressed. Dual-activity prediction models trained on undersampled data tended to better distinguish between dual-active and inactive compounds, as can be observed by the G-mean scores when comparing this to models trained on oversampled or imbalanced data (Table 3.7). Surprisingly, the FPR within most models tended to be low, however, models trained on oversampled or imbalanced data had lower sensitivity in identifying dual-active compounds. For oversampling, this could result from overfitting and fixating on overrepresented chemical features within the training data. This indicated that undersampling improved model performance for dual-activity prediction and was a better alternative to oversampling overall.

Table 3.7: Comparison of dual-activity prediction models' performance on the test set when trained on either undersampled, oversampled or imbalanced training data

Model	G-Mean	FPR (FP/FP+TN)	Sensitivity (TP/TP+FN)	Specificity (TN/TN+FP)	F1-Score
Dual-activity prediction model trained on imbalanced data					
SVM (ECFP)	0.755	0	0.571	1	0.72
RF (ECFP)	0.703	0.001	0.495	0.999	0.647
GBM (ECFP)	0.647	0	0.419	1	0.58
LR (ECFP)	0.853	0.014	0.737	0.986	0.543
SVM (MACCS)	0.671	0	0.45	1	0.618
RF (MACCS)	0.816	0.001	0.667	0.999	0.768
GBM (MACCS)	0.72	0.001	0.519	0.999	0.662
LR (MACCS)	0.884	0.014	0.792	0.986	0.543
Dual-activity prediction model trained on oversampled data					
SVM (ECFP)	0.452	0	0.204	1	0.333
RF (ECFP)	0.247	0	0.061	1	0.113
GBM (ECFP)	0.684	0.005	0.469	0.995	0.343
LR (ECFP)	0.636	0.01	0.408	0.99	0.191
SVM (MACCS)	0.571	0.002	0.327	0.998	0.372
RF (MACCS)	0.429	0	0.184	1	0.29
GBM (MACCS)	0.706	0.024	0.51	0.976	0.125
LR (MACCS)	0.771	0.029	0.612	0.971	0.126
Dual-activity prediction model trained on undersampled data					
SVM (ECFP)	0.809	0.006	0.658	0.994	0.262
RF (ECFP)	0.562	0	0.316	1	0.409
GBM (ECFP)	0.72	0.005	0.521	0.995	0.244
LR (ECFP)	0.796	0.061	0.675	0.939	0.04
SVM (MACCS)	0.836	0.015	0.709	0.985	0.143
RF (MACCS)	0.766	0.005	0.59	0.995	0.295
GBM (MACCS)	0.793	0.007	0.632	0.993	0.248
LR (MACCS)	0.851	0.069	0.778	0.931	0.04

Considering the significant disparity in the number of active vs. inactive compounds for dual-activity, even with undersampling, the F1, G-mean and recall scores were also considered. From these performance metrics, the SVM model was identified as the best-performing model trained on ECFP for dual-activity prediction with optimal F1 (0.26), G-mean (0.81), and recall (0.66) scores and low false positive predictions (0.006) (Figure 3.9C). For models trained on MACCS, SVM and GBM obtained the best ROC-AUC scores during cross-

validation (>0.95) and upon the test set (>0.91, Figure 3.10). Interestingly, for models trained on MACCS, tended to have a higher FPR was observed when compared to models trained on ECFP, possibly indicating that MACCS is not as good at generating molecular descriptors of the chemical space for dual-activity, however this difference in FPR was very slight. In contrast, ECFP is more extensive and descriptive, allowing better capability to distinguish active and inactive compounds.

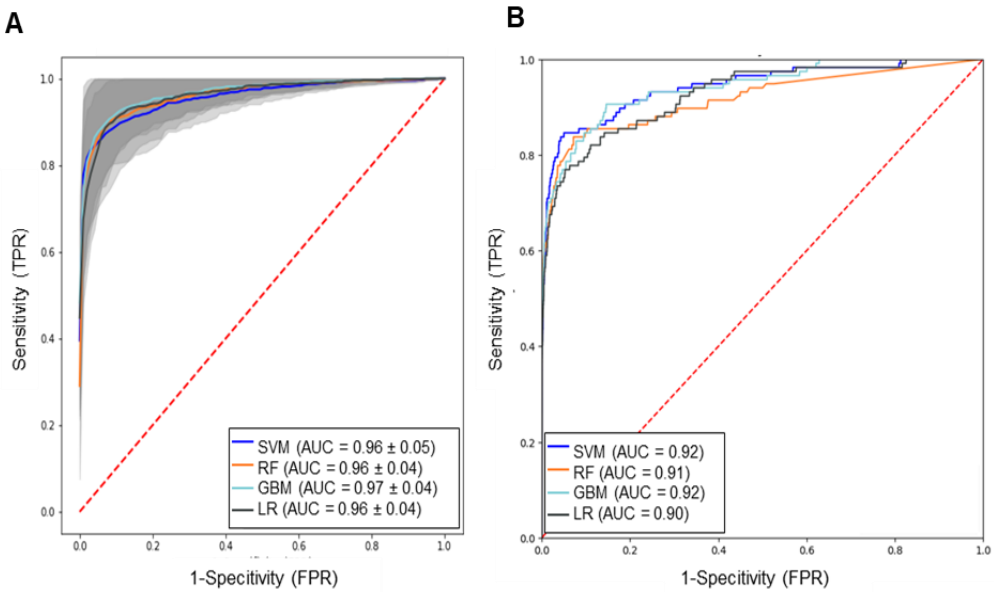


Figure 3.10: ROC-AUC curves from cross-validation and imbalanced test set performance of dual-activity prediction models trained on MACCS.
ROC-AUC curves (A) when trained on MACCS keys of compounds after 5-fold cross-validation. Insert indicates AUC mean values ± standard deviation. (B) Models trained on MACCS descriptors on untrained test sets in predicting dual-activity or inactivity of compounds.

Due to the models' low precision within the test set, we set out to identify the ideal threshold that would enable better precision whilst not drastically compromising the recall of our dual-activity prediction models. For the SVM model, precision could be increased to 0.562 using higher probability thresholds (>0.9) (Table 3.8, Figure 3.11A) at the cost of drastically decreasing recall (from >0.6 to <0.3) in identifying dual-active compounds.

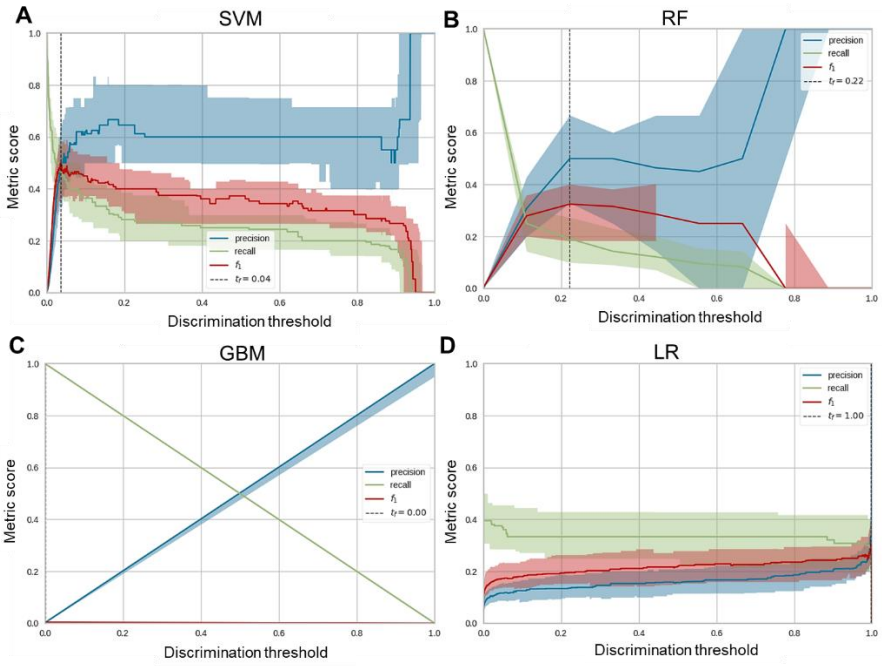


Figure 3.11: Influence of discrimination threshold adjustment on dual-activity model performance within the untrained test set.
A) SVM, (B) RF, (C) GBM and (D) LR model performance regarding precision, recall and f1-score was calculated and plotted for each threshold defining dual-active and inactive compounds from the predicted probability, i.e., discrimination threshold. Discrimination threshold adjustment was conducted on test set data with dual-activity models. Tf indicated the threshold at which both recall, and precision was the highest.

Table 3.8: Optimised probability threshold for dual-activity prediction models trained on ECFP of compounds

Model	G-Mean	FPR	ROC-AUC	Recall	Precision	Probability threshold
Dual-activity prediction model trained on undersampled data						
SVM class-weighted (ECFP)	0.809	0.006	0.826	0.658	0.164	0.5
RF (ECFP)	0.562	0.000	0.658	0.316	0.578	0.5
GBM (ECFP)	0.720	0.005	0.758	0.521	0.159	0.5
LR class-weighted (ECFP)	0.796	0.061	0.807	0.675	0.020	0.5
SVM class-weighted (ECFP)	0.654	0.001	0.713	0.427	0.562	0.92
RF (ECFP)	0.562	0.000	0.658	0.316	0.617	0.52
GBM (ECFP)	0.547	0.000	0.649	0.299	0.565	0.82
LR class-weighted (ECFP)	0.768	0.041	0.787	0.615	0.027	0.96

To assess whether more complex models would be more suited for dual-activity prediction, we compared our top two models to those of more complex models generated via autogluon. Although our conventional models obtained lower G-mean scores, the FPR within model predictions were considerably lower than that of the more complex models, indicating that such complex models are more prone to classify inactive compounds as dual-active, which

is undesirable (Table 3.9). This could be because complex models tend to overfit the data and, as a result, have poor generalisation on these relatively small datasets. In conclusion, SVM was the best model for predicting compounds with dual-activity with low false positive predictions on test data within representative chemical spaces.

Table 3.9: Model performance comparison on test data of best-performing dual-activity prediction models and more complex models

Model	G-Mean	FPR (FP/FP+TN)	ROC-AUC	Recall	Precision	F1-Score
Dual-activity prediction model						
SVM (ECFP)	0.809	0.006	0.826	0.658	0.164	0.485
RF (ECFP)	0.562	0.000	0.658	0.316	0.578	0.409
NeuralNetFastAI	0.992	0.632	0.813	0.632	0.147	0.238
WeightedEnsemble_L2	0.995	0.641	0.818	0.641	0.201	0.305

3.3.5) Dual-activity prediction models performance on the external validation dataset

To determine whether the top-performing dual-activity prediction models would show the same robustness as ABS activity prediction models when exposed to novel and chemically diverse compounds, these dual-activity models were also tested against the PRB and Pathogen Box. Like the selection of the ABS activity prediction model, the best dual-activity prediction model must maintain fair accuracy and recall under these conditions while simultaneously optimizing sensitivity and specificity in predictions to limit the models' FPR. Class imbalance was not corrected for these datasets to evaluate the performance of the models, with a gametocytocidal activity hit rate of ~13% for the PRB box and ~24% for the Pathogen box. It should be noted that the hit rate of gametocytocidal activity was much lower than that of the ABS inhibition activity, which makes the external validation dataset a prime example of what to expect in the real-world application of the models.

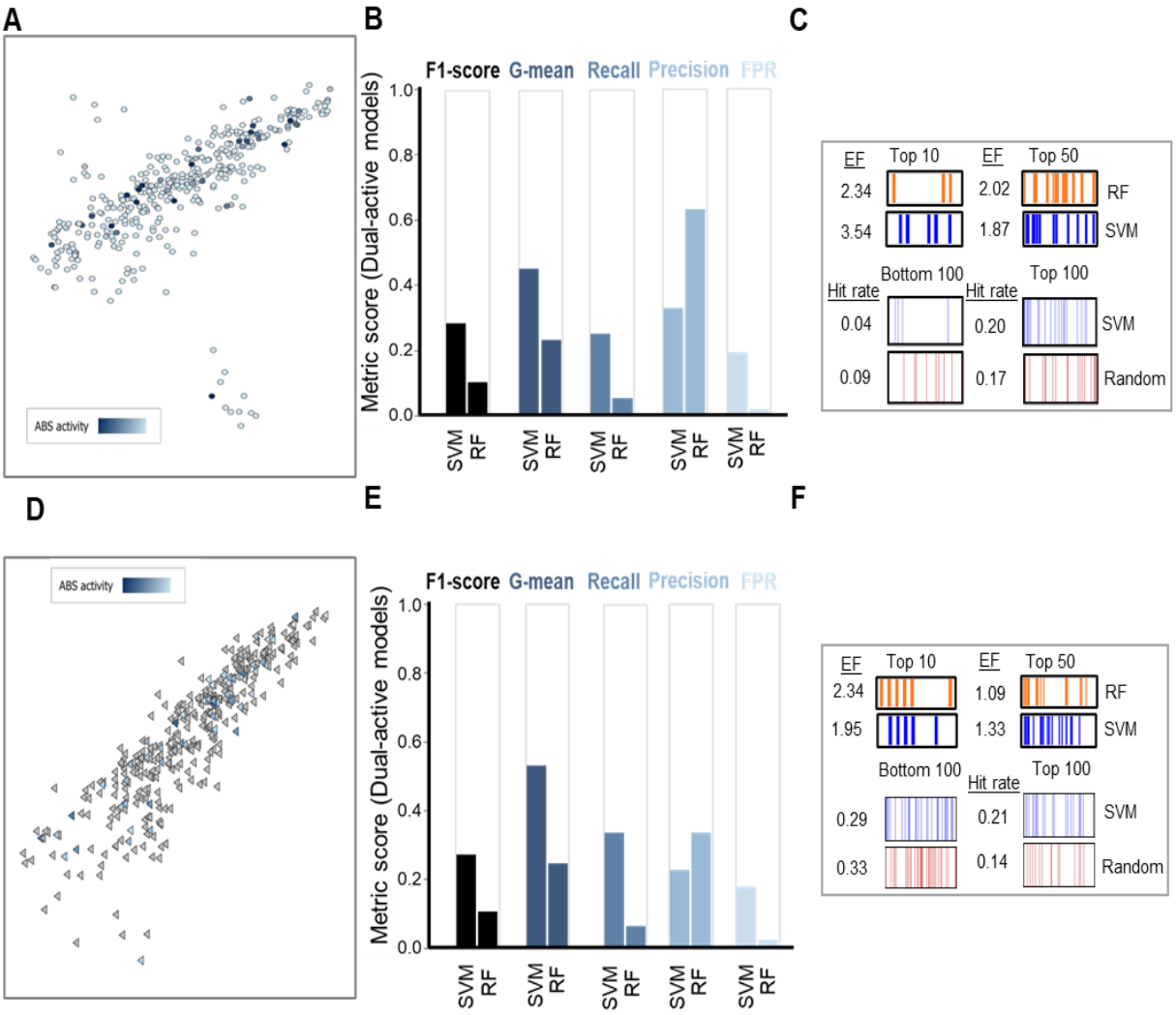


Figure 3.12: Model performance of different ML algorithms in identifying compounds with dual-activity within novel diverse chemical spaces.
Model robustness was evaluated by exposing the top two dual-activity prediction models to extreme datasets from the PRB box (A) and Pathogen box (D). Each dataset is individually distinct, chemically diverse (displayed within the context of the launched drug chemical space (available on StarDrop v 7.3.0), with heatbars indicating potency) and had differential activity against ABS and gametocyte stages. Dual-activity prediction models trained on ECFP descriptors were evaluated for their activity predictions within the PRB box (B) and the Pathogen box (E) for F1-scores (model performance exposed to imbalanced data) G-mean scores (ability to optimize sensitivity and specificity), recall, precision, and false positive rate (FPR). The hit rate of the best-performing model for correctly predicting dual-activity within these chemical spaces (C and F) was compared to random selection. The enrichment factor (EF) of models was also calculated for the top 10 and top 50 compounds to determine how effective models were in prioritizing active compounds.

Within these novel chemical spaces, the SVM and RF dual-activity models differentiated more clearly from each other when predicting dual-active compounds, with the SVM model outperforming RF on all metrics, excluding precision (Figure 3.12B & E). Investigation into the RF model predictions revealed that high precision was obtained by severely limiting the number of compounds classified as having dual-activity, which is undesirable for the purpose of our models. Though SVM tended to have a higher FPR (0.17 *vs.* 0.02), RF failed to identify dual-active compounds with a recall below <0.20, and SVM tends to shift and prioritize dual-active compounds to the top of the list more than RF for both the top 10

(3.54 *vs.* 2.34) and top 50 (1.87 *vs.* 2.02) compounds (Figure 3.12C). Together with this, the hit rate of the top 100 hits from the SVM model (20%) exceeded the hit rate of random selection (17%) and the PRB box (13%) (Figure 3.12C).

Table 3.10: Comparison of dual-activity prediction models trained on either ECFP and MACCS and their performance in predicting within novel diverse chemical spaces

Model	G-Mean	FPR (FP/FP+TN)	Sensitivity (TP/TP+FN)	Specificity (TN/TN+FP)	F1-Score	Accuracy
Dual-activity prediction model on PRB box						
SVM class-weighted (ECFP)	0.525	0.172	0.333	0.828	0.266	76.5
RF (ECFP)	0.240	0.017	0.059	0.983	0.100	86.5
GBM (ECFP)	0.487	0.135	0.275	0.865	0.250	79.0
LR class-weighted (ECFP)	0.593	0.309	0.510	0.691	0.281	66.8
SVM class-weighted (MACCS)	0.521	0.341	0.412	0.659	0.220	63.0
RF (MACCS)	0.444	0.160	0.235	0.840	0.202	76.0
GBM (MACCS)	0.501	0.246	0.333	0.754	0.221	70.0
LR class-weighted (MACCS)	0.571	0.550	0.725	0.450	0.264	49.0
Dual-activity prediction model on Pathogen box						
SVM class-weighted (ECFP)	0.449	0.185	0.247	0.815	0.281	66.5
RF (ECFP)	0.226	0.011	0.052	0.989	0.095	74.1
GBM (ECFP)	0.444	0.170	0.237	0.830	0.277	67.3
LR class-weighted (ECFP)	0.467	0.359	0.340	0.641	0.291	56.1
SVM class-weighted (MACCS)	0.492	0.289	0.340	0.711	0.317	61.3
RF (MACCS)	0.408	0.104	0.186	0.896	0.252	70.8
GBM (MACCS)	0.408	0.152	0.196	0.848	0.242	67.6
LR class-weighted (MACCS)	0.567	0.463	0.598	0.537	0.414	55.3

To ensure no bias within the final model selection, we evaluated the performance of dual-activity prediction models trained on MACCS in predicting the external datasets. Like ABS activity prediction models, dual-activity models trained on MACCS tended to have a higher FPR than that of the same algorithm trained on ECFP (Table 3.10). This indicated that ECFP was more suited for model building to allow better dual-active prediction accuracy when exposed to novel and diverse chemical spaces. Although the precision of the SVM models tended to be low, closer examination of compounds randomly selected from the PRB box and comparison of them to the training set (Figure 3.13) revealed that the chemical similarity distribution is lower than that of the test set compounds. Considering the diverse nature of the external validation datasets, some of these compounds may fall outside of the applicability domain of our dual-activity prediction models. We observe that PRB compounds

3 and 5 both show low chemical similarity distributions towards the training set of dual-activity, and both were wrongly predicted as inactive when they show gametocytocidal activity (Figure 3.13B, Table 3.11).

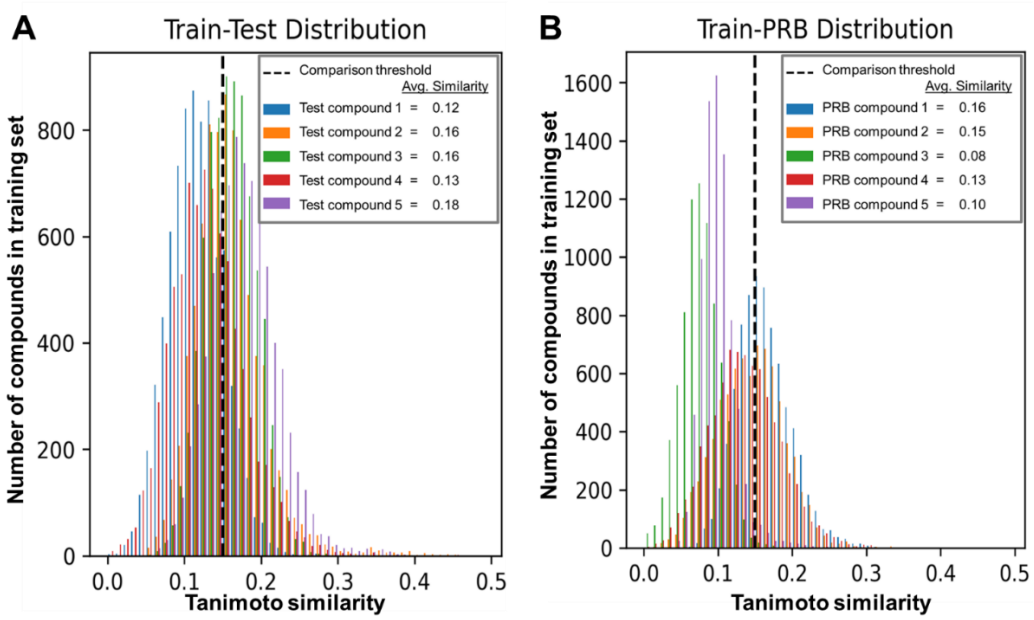


Figure 3.13: Tanimoto similarity distribution of PRB box or test set compounds on dual-activity model training set. Tanimoto similarity distribution plots were generated for each of the five compounds randomly selected from the (A) dual-active test set and PRB Box (B) based on their structural similarity to the training set used for the dual-activity prediction models. Average Tanimoto similarity of compounds over the training sets are indicated within the label The predicted activity of the randomly selected compounds is shown in Table 3.11.

To ensure no bias within the final model selection, we evaluated the performance of dual-activity prediction models trained on MACCS in predicting the external datasets. Like ABS activity prediction models, dual-activity models trained on MACCS tended to have a higher FPR than that of the same algorithm trained on ECFP (Table 3.10). This indicated that ECFP was more suited for model building to allow better dual-active prediction accuracy when exposed to novel and diverse chemical spaces. Although the precision of the SVM models tended to be low, closer examination of compounds randomly selected from the PRB box and comparison of them to the training set (Figure 3.13) revealed that the chemical similarity distribution is lower than that of the test set compounds. Considering the diverse nature of the external validation datasets, some of these compounds may fall outside of the applicability domain of our dual-activity prediction models. We observe that PRB compounds 3 and 5 both show low chemical similarity distributions towards the training set of dual-activity, and both were wrongly predicted as inactive when they show gametocytocidal activity (Figure 3.13B, Table 3.11).

Table 3.11: Activity predictions of compounds with low chemical similarity to dual-activity model training set

Compound	Activity	Predicted activity	Predicted probability
Dual-activity SVM model			
Test set compound 1	Dual-activity/Gametocytocidal	Inactive	0.021
Test set compound 2	Dual-activity/Gametocytocidal	Active	0.959
Test set compound 3	Dual-activity/Gametocytocidal	Active	0.537
Test set compound 4	Dual-activity/Gametocytocidal	Active	0.983
Test set compound 5	Dual-activity/Gametocytocidal	Active	0.962
PRB compound 1	Dual-activity/Gametocytocidal	Active	0.950
PRB compound 2	Dual-activity/Gametocytocidal	Active	0.974
PRB compound 3	Dual-activity/Gametocytocidal	Inactive	0.302
PRB compound 4	Dual-activity/Gametocytocidal	Inactive	0.011
PRB compound 5	Dual-activity/Gametocytocidal	Inactive	0.233
Dual-activity RF model			
Test set compound 1	Dual-activity/Gametocytocidal	Inactive	0.137
Test set compound 2	Dual-activity/Gametocytocidal	Active	0.578
Test set compound 3	Dual-activity/Gametocytocidal	Inactive	0.202
Test set compound 4	Dual-activity/Gametocytocidal	Active	0.933
Test set compound 5	Dual-activity/Gametocytocidal	Inactive	1.000
PRB compound 1	Dual-activity/Gametocytocidal	Inactive	0.200
PRB compound 2	Dual-activity/Gametocytocidal	Active	0.513
PRB compound 3	Dual-activity/Gametocytocidal	Inactive	0.338
PRB compound 4	Dual-activity/Gametocytocidal	Inactive	0.075
PRB compound 5	Dual-activity/Gametocytocidal	Inactive	0.144

3.4) Discussion

To eliminate malaria, dual-active compounds which target both the ABS of the parasite, responsible for the clinical symptoms, and the parasite's transmissible stages to limit the disease's spread are very attractive antimalarials. Due to the unique biology of the parasite, compounds active towards the proliferative ABS stage might not have the same activity towards the differentiating gametocyte stages. The identification of gametocytocidal compounds is further complicated by the long turnaround time required to screen compounds against all gametocyte stages, especially against the mature transmissible Stage V parasites, which can take days to assay. Additionally, the cost associated with culturing and assaying during phenotypic screening makes identifying gametocytocidal compounds more costly and laborious than ABS phenotypic screening. Tools that prioritise the screening of compounds which are most likely to show activity against ABS and gametocyte stages would aid immensely in accelerating transmission-blocking drug discovery. Through the use of ML, this has been done for ABS via DeepMalaria and MAIB [80, 82] as well as liver-stage phenotypic screens [81] by leveraging the chemical space information of phenotypic screens conducted against the parasite. Nevertheless, to our

knowledge, no study has fully used the chemical space information of phenotypic screens conducted against gametocytes and ABS to predict dual-active compounds with ML.

Leveraging the use of ML algorithms capable of training on imbalanced data, we successfully trained models adept at predicting compounds with activity against both ABS parasites and gametocytes with good accuracy, precision, and recall. Throughout model building and evaluation, it became apparent that cluster-based undersampling of inactive compounds paired with ML algorithms suited for imbalanced data was ideal for dual-activity model building, as oversampling resulted in lower sensitivity in predicting active compounds. The use of oversampling may result in the model failing to generalise and recognise patterns within novel active compounds and may, as a consequence, cause the model to become fixated on patterns associated with oversampled active compounds due to overrepresentation of such features within training data[162, 163]. Additionally, the use of ML algorithms more adept at training on class imbalanced data while also trying to limit class bias within training data proved to be more informative in building ML models for predicting ABS and/or dual-activity, negating the use of more complex ML approaches such as deep neural network. Similar observations have been seen in different fields involving image classification when comparing deep learning to more standard ML methods such as SVM [164, 165]. Interestingly, single classifiers like SVM outperformed ensemble models within representative chemical spaces in our data, which could be considered smaller than the typical big data in that ensemble methods such as GBM and RF generally perform better. Most noteworthy, our best-performing SVM and RF models for ABS inhibition activity prediction were on par with DeepMalaria [82] that used neural network-based models, with SVM obtaining higher recall abilities (95% *vs*. 87.75%) but slightly less precision (2.40% *vs*. 3.54%) with similar accuracy (60% *vs*. 60.06%) when tested on the same untrained data.

Although our dual-active models within this study obtain low precision, identifying dual-active compounds through phenotypic screening is very scarce. Considering the time and cost associated with screening such compounds, high recall is desirable despite identifying some false positives. Additionally, our models have proven that they can sort dual-active compounds to the forefront of potential candidate lists, as observed by the enrichment factor of the models. These models have limitations in that the error in classifying the activity of compounds may become more error-prone as the chemical similarity between query compounds and the training data decreases. Such compounds may fall outside the chemical

applicability domain of our models, however, we aim to keep these models up to date with the latest curated phenotypic screening data made available. Interestingly, the concerns regarding model performance due to training data containing chemical libraries with varied activity thresholds were consolidated by the model's performance on external datasets. This confounding aspect can still be further investigated in future to determine whether better models could be obtained when trained on a smaller database employing a consistant activity threshold.

3.5) Conclusions

In summary, cluster-based undersampling and ML algorithms suited for training on class imbalanced data proved to generate better models than TL and deep learning. The models we generated performed well within representative and novel chemical spaces and could prioritise active compounds to the forefront of screening lists. Our models will aid immensely in fast-tracking transmission-blocking antimalarial drug discovery to reach malaria elimination goals by reducing the time and cost allocated towards compounds with a low probability of having dual-activity against both ABS parasites and gametocytes. Furthermore, to aid infectious disease drug discovery, we have made these models freely available and open source within the Ersilia Model Hub repository [166].

CHAPTER 4

CHEMICAL FEATURES PREDICTIVE OF ACTIVITY AGAINST ABS AND DUAL-ACTIVITY

The work in this chapter has been published as follows:
Heerden, A.v., Turon, G., Duran-Frigola, M., Pillay, N., and Birkholtz, L.-M. (2023). Machine learning approaches identify chemical features for stage-specific antimalarial compounds. ACS Omega. 2023 Nov 7;8(46):43813-43826. doi: 10.1021/acsomega.3c05664. eCollection 2023 Nov 21

ML has provided powerful tools to fast-track drug discovery by prioritising promising candidates with the highest probability of showing activity. These ML models also hold invaluable chemical information via feature importance scores that indicate which features the model uses highly in making such predictions. This information can further aid in understanding how to develop better drugs and what chemical modifications during derivatisation would improve or reduce the potency of compounds or allow additional stage-specific activity. Most ML models allow the user to do feature analysis, where the features are weighted according to their importance in aiding predictions. Others, however, apply data transformation techniques to identify patterns within data, and such approaches make it difficult to calculate the actual weight such features hold before such transformations.

In Chapter 3, we generated models capable of predicting ABS and dual-activity with good accuracy and recall. Such models may contain invaluable information on stage-specific activity's chemical features. Hence, we aimed to utilise the models to identify chemical features that are statistically significant and important for the model in compound activity prediction.

4.2) Methods

4.2.1) Ethics

All computational work, the saving, modifying and storing of data has ethics approval (University of Pretoria Ethics Application NAS345/2020).

4.2.2) Enrichment analysis on ECFP features of active and inactive compounds

Although feature importance analysis would highlight important ECFP features that are predictive of compound activity or inactivity, it would not distinguish which features are predictive of compound activity and which are predictive of inactivity. Alternatively, we identified ECFP features significantly enriched within active/inactive compounds (ABS inhibiting or dual-active), intending to use this information to define whether a predictive ECFP feature was important for activity/inactivity prediction during feature importance analysis. Enrichment of EFCP features towards active *vs* inactive compounds was identified using the Z-test on two sample proportions. ECFP features were defined as significantly enriched if a Z-score >2.5 (p-value < 0.01) was obtained within active compounds compared to inactive compounds. Significantly enriched features identified within active compounds were then ranked according to the feature importance score obtained from the best-performing activity prediction model.

4.2.3) Feature importance analysis

Highlighting the ECFP features important for predicting ABS or dual-activity was complicated due to our best-performing model, SVM, employing a radial basis function (RBF) kernel transformation technique. This data transformation technique complicated feature analysis in calculating a feature's weight within model predictions. To allow interpretability, RF models trained on ECFP with good precision and the ability to distinguish active and inactive compounds within representative chemical spaces against ABS and/or gametocytes were selected to perform feature importance analysis instead of SVM. During feature analysis, the permutation importance of RF models was calculated and extracted to prevent the inclusion of impurity-based features that inflate the importance of features not predictive of compound activity. Recursive feature elimination (RFE) via scikit-learn [145] was also performed to determine which features were the most necessary for compound activity predictions. Additionally, to highlight any features important for ABS and/or dual-activity, shapley values were calculated via the SHAP python package [167], however, due to the density of the data provided, accurate shapley values could not be calculated.

4.2.4) Extraction of top 100 features for models

The top 100 ECFP features were identified from RF models and compared to the top 100 ECFP features identified in other activity prediction models to highlight any overlap between the top ECFP features of different models. Theoretically, if features are predictive of compound activity, such features would also be present as the top features within different models and would be categorized as important regardless of the different algorithms employed in pattern recognition. Hence, we wanted to identify and highlight such chemical features which could be important to consider for chemical modifications during derivatization.

4.3) Results

4.3.1) ECFP features enriched within active or inactive compounds

Within both ABS and dual-active databases, the enrichment of ECFP features (500 in total) within active and inactive compounds was calculated using the Z-test. As a result, no overlap of compounds was observed for compounds enriched in active or inactive compounds for a specific stage-activity (ABS inhibition or dual-activity). A threshold of Z-score >2.5 (p-value < 0.01) was used to define whether compounds were significantly enriched for compound activity/inactivity. Using this, within the ABS database, 383 ECFP features were identified as enriched within ABS inhibiting compounds, whereas 67 ECFP features were identified as enriched in inactive compounds against ABS (Figure 4.1A). Similarly, within the dual-active database, 266 ECFP features were identified as enriched within dual-active compounds and 52 ECFP features were identified as enriched in inactive compounds (Figure 4.1B). Interestingly, considerably fewer ECFP features were enriched for dual-activity than those observed in the ABS-inhibiting space (266 features compared to 383 for ABS activity). This could be a result of the limited chemical space which had been phenotypically screened against gametocytes compared to ABS and/or that the chemical space for dual-activity is much narrower and limited.

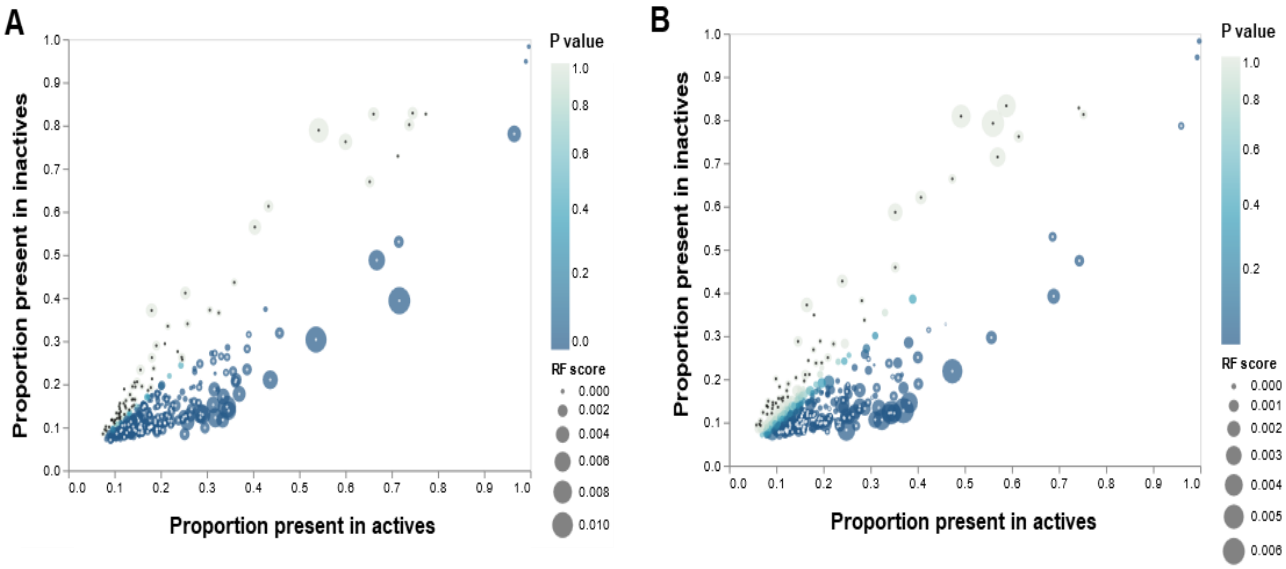


Figure 4.1: Enriched ECFP features within inactive and active compounds for stage-specific antiplasmodial action. (A) The proportion of active/inactive compounds against ABS containing a specific ECFP feature was plotted as circles. Size of the circles related to the RF permutation score of the ECFP feature within the ABS activity prediction RF model. Enrichment of an ECFP feature towards active compounds compared to inactive compounds is indicated by the p-value color obtained from the Z-test on two proportions (active vs inactive). The top 100 enriched ECFP features within active (white dots inside circles) and the top 67 enriched ECFP features within inactive (black dots inside circles) compounds were selected according to RF score and ρ-value. (B) The proportion of dual-active/inactive compounds containing a specific ECFP feature was plotted as circles, with circle size corresponding to the RF permutation score of the ECFP feature within the dual-activity prediction RF model. Enrichment of a feature towards dual-active compounds compared to inactive compounds is indicated by the ρ-value color. The top 100 enriched ECFP features within dual-active (white dots inside circles) and the top 52 enriched ECFP features within inactive (black dots inside circles) compounds were selected according to RF score and ρ-value.

4.3.2) Top predictive ECFP features enriched in active compounds

Although the above ECFP features identified may be enriched within active or inactive compounds, this could be purely due to their prevalence in the chemical libraries used in phenotypic screening and may not necessarily indicate that such features contribute to compound activity. Thus, using the models in Chapter 3, we ranked enriched features within active compounds according to their importance in model predictions for activity, i.e., feature importance. However, feature importance analysis was complicated due to SVM models employing an RBF data transformation, which hindered calculating importance scores for chemical features. However, the RF models showed a fair correlation with SVM predictions (Figure 4.2), and though RF was not the top model, valuable insights into the contribution of these features towards activity prediction could still be gleaned from such models, especially if said features are repeatedly identified as top features using different ML algorithms. Therefore, RF was used for feature importance analysis because it was the second-best model in each instance and because the predicted probability scores on test sets for SVM and RF models had a good correlation.

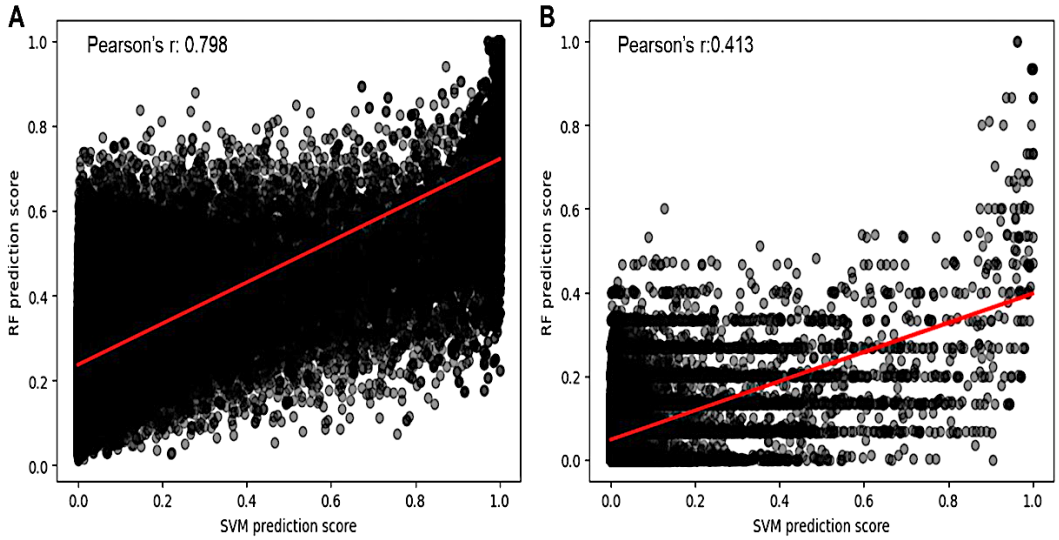


Figure 4.2: Correlation of predicted probability scores between RF and SVM on test set.
(A) Pearson correlation of predicted probability scores for ABS activity prediction SVM and RF models (trained on ECFP) on the test set. (B) Pearson correlation of predicted probability scores for dual-activity prediction SVM and RF models (trained on ECFP) on the test set.

The top 100 features were then selected based on ranked feature importance scores and whether features were enriched within ABS/dual-active compounds to provide a more comprehensive, inclusive dataset when comparing important shared features between models. More than half of the features identified within the RF models were additionally confirmed by at least one other model (Table 4.1, Figure 4.3).

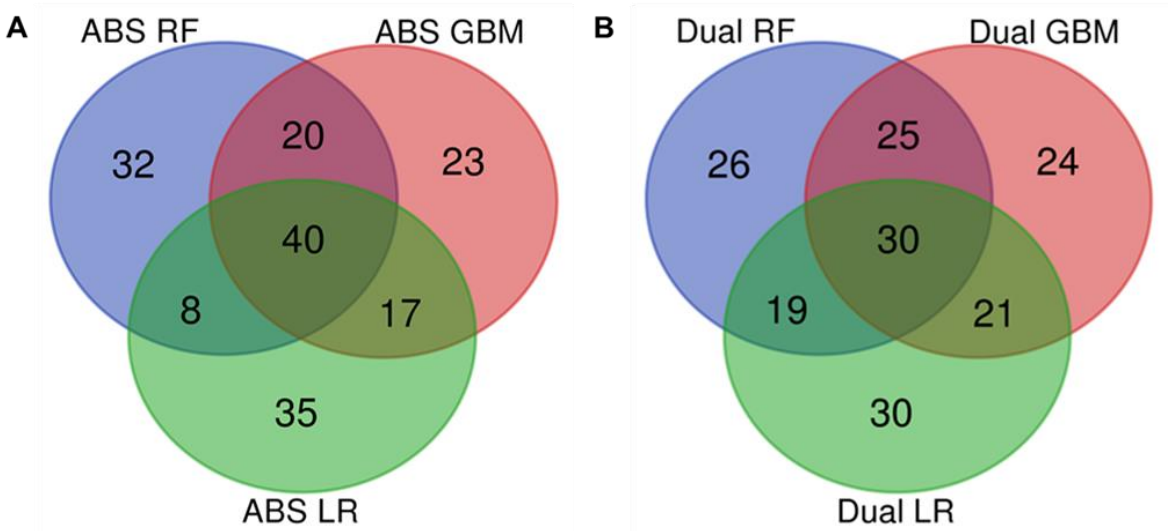


Figure 4.3: Top ECFP features unique and shared features between ABS and dual-activity prediction models.
The top 100 ECFP features were identified for (A) ABS activity prediction in RF, GBM and LR models. Similarly, the top 100 ECFP features were identified for (B) dual-activity prediction RF, GBM and LR models. Intersections indicate the number of features shared between models and are detailed in Table 4.1.

No overlap was observed between the top 100 features enriched for ABS activity and those enriched in inactive compounds against ABS (Figure 4.4A), validating the high level of specificity obtained with associating ECFP features with ABS activity. This was also seen for features in compounds with dual-activity, with no overlap between the top 100 features for dual-activity compared to the 52 features enriched in inactive compounds against gametocytes, indicating the specificity of these features for dual-activity (Figure 4.4B).

Table 4.1: Top 100 ECFP features for compound activity shared or unique among models

Models sharing ECFP features	Number of ECFP features shared	ECFP features shared
ABS activity prediction models		
GBM, LR, RF	40	200; 161; 298; 194; 78; 349; 291; 114; 153; 81; 311; 346; 495; 323; 345; 191; 178; 234; 314; 388; 326; 209; 256; 195; 21; 80; 99; 457; 277; 377; 232; 295; 50; 375; 303; 496; 239; 491; 332; 213
GBM, RF	20	90; 463; 434; 325; 226; 321; 76; 67; 236; 112; 487; 367; 216; 324; 173; 70; 484; 387; 328; 265
LR, RF	8	424; 93; 262; 176; 74; 447; 221; 310
GBM, LR	17	102; 27; 458; 431; 471; 172; 193; 72; 351; 201; 149; 302; 499; 250; 305; 59; 42
RF	32	206; 71; 125; 190; 437; 313; 231; 89; 343; 11; 275; 197; 337; 101; 459; 62; 241; 69; 119; 264; 92; 414; 167; 39; 260; 380; 155; 53; 158; 354; 399; 242
GBM	23	127; 259; 243; 35; 65; 203; 371; 145; 154; 96; 126; 37; 63; 455; 162; 410; 334; 68; 25; 46; 257; 3; 408
LR	35	18; 366; 95; 220; 20; 347; 29; 374; 58; 129; 2; 14; 131; 282; 23; 364; 159; 170; 428; 180; 341; 438; 87; 56; 470; 144; 283; 254; 105; 248; 111; 146; 456; 4; 317
GBM, LR, RF	40	200; 161; 298; 194; 78; 349; 291; 114; 153; 81; 311; 346; 495; 323; 345; 191; 178; 234; 314; 388; 326; 209; 256; 195; 21; 80; 99; 457; 277; 377; 232; 295; 50; 375; 303; 496; 239; 491; 332; 213
GBM, RF	20	90; 463; 434; 325; 226; 321; 76; 67; 236; 112; 487; 367; 216; 324; 173; 70; 484; 387; 328; 265
Dual-activity prediction models		
GBM, LR, RF	30	200; 194; 78; 434; 349; 291; 58; 81; 311; 495; 345; 191; 282; 326; 195; 21; 455; 99; 324; 377; 149; 310; 303; 496; 328; 491; 318; 257; 332; 242
GBM, RF	25	118; 55; 84; 95; 325; 203; 261; 153; 82; 178; 79; 364; 176; 256; 216; 61; 232; 167; 70; 100; 484; 392; 473; 213; 408
LR, RF	19	424; 20; 397; 65; 459; 241; 471; 218; 404; 314; 486; 227; 488; 447; 477; 342; 360; 421; 13
GBM, LR	21	127; 445; 376; 220; 337; 289; 2; 323; 236; 234; 209; 80; 379; 263; 474; 370; 116; 380; 46; 436; 456
RF	26	298; 31; 340; 76; 339; 361; 112; 187; 121; 344; 355; 479; 367; 460; 297; 334; 312; 56; 260; 1; 375; 416; 254; 155; 354; 3
GBM	24	90; 366; 233; 190; 259; 275; 321; 346; 69; 487; 43; 457; 255; 277; 91; 438; 87; 214; 293; 365; 141; 222; 25; 250
LR	30	276; 27; 161; 458; 89; 157; 29; 431; 432; 67; 327; 212; 126; 483; 37; 428; 383; 410; 201; 50; 221; 173; 499; 68; 283; 450; 85; 185; 390; 439

We further investigated if there was any overlap between the top ECFP features predictive of ABS activity and those predictive of dual-activity, as well as to determine if there were any unique features associated with stage-specific activity. As expected, about 50% of the top 100 features associated with ABS activity or dual-activity were shared (Figure 4.4C, Table 4.2).

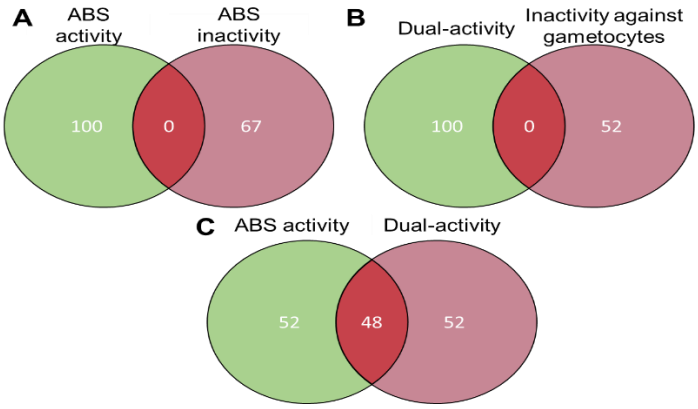


Figure 4.4: Intersection between the top ECFP features identified for ABS and dual-activity and ECFP features enriched within inactive compounds.
(A) No overlap observed between ECFP features enriched within inactive compounds against ABS and the top 100 ECFP features (enriched within ABS active compounds) and identified as important for ABS activity prediction in RF models. (B) Similarly, there was no intersection between ECFP features enriched within inactive compounds against gametocytes and the top 100 ECFP features (enriched within dual-active compounds) and identified as important for dual-activity prediction in RF models. (C) In total, 48 ECFP features that were identified as enriched within ABS inhibiting compounds and important for ABS prediction were also similarly identified as enriched within dual-active compounds and important for dual-activity prediction.

Of the top 100 features enriched in the respective active fractions, 52 features were uniquely described as either associated with sole ABS activity or dual-activity (Figure 4.4C & Figure 4.5, Table 4.3). These 104 features (52 associated with sole dual-activity and 52 associated with sole ABS activity) clearly distinguish the chemical composition of compounds required to kill multiple stages of the parasite or only to kill ABS parasites.

Table 4.2: Top 5 ECFP features enriched for ABS and dual-activity and identified as important for activity prediction in ABS and dual-activity prediction models

ECFP feature number	RF score (ABS model)	Z- score on ABS database	RF score (Dual-activity model)	Z-score on dual database	ECFP feature
349*	0.0056	57.5	0.0055	19.6	
311*	0.0052	59.1	0.0051	17.2	
303*	0.0026	39.2	0.0046	18.2	
377*	0.0020	54.4	0.0039	16.2	
200*	0.0075	55.4	0.0036	18.3	

*ECFP features were also identified as important via RFE for both ABS and dual-activity prediction

Closer inspection of the top 5 unique features (Figure 4.5) enriched in compounds able to target ABS parasites contain heterocyclic structures that are activated via oxygenation or alkylation, allowing the compound to be more reactive to electrophilic attack. Additionally, oxygenation may be associated with a compound's bioavailability by increasing the compound's solubility and uptake [168]. By contrast, the top 5 unique features enriched in compounds with dual-stage activity (Figure 4.5) are enriched for amine groups, whether it be nitrogen-containing 4-membered heterocyclic structures or the nitration of benzene and/or carbon groups. This is interesting, considering that amine groups typically create localized electron deficient sites within the compound, enabling interaction with cellular components such as amino acids and nucleic acids to be more favourable [169]. Hence, adding amine groups to 4-membered heterocyclic structures and benzene groups may be more involved in the killing/drug effect and aid in the solubility of compounds.

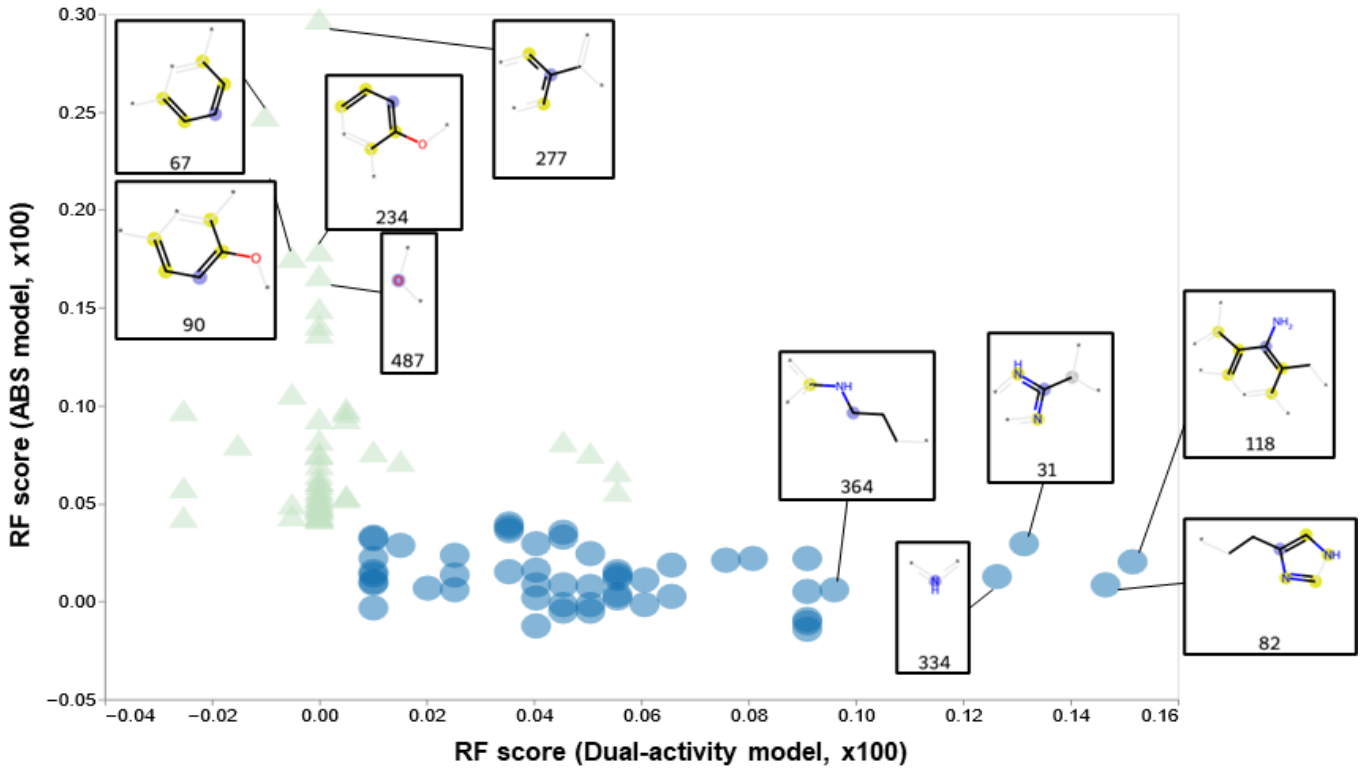
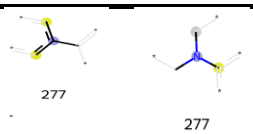
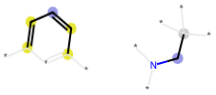
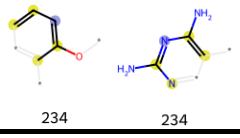
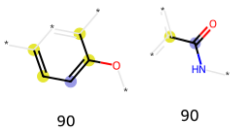
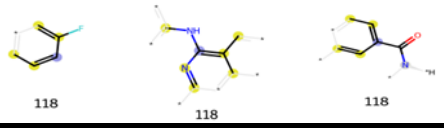
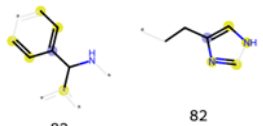
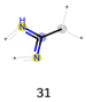
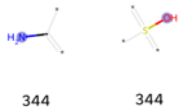
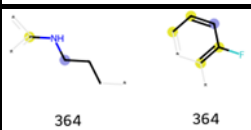


Figure 4.5: Unique enriched ECFP features associated with sole ABS activity or dual-activity.
Comparison of the unique ECFP features associated with activity against ABS (52, green) or dual stages (52, blue). For the top unique ECFP features, structural elements are indicated.

Table 4.3: Top 5 ECFP features enriched for compound activity and showing stage-specific described as either associated with sole ABS activity or with dual-activity

ECFP feature number	RF score (ABS model)	Z- score on ABS database	ECFP features for ABS activity
277	0.0030	35.9	
67[2]	0.0025	41.3	
234[1]	0.0018	35.4	
90	0.0017	42.3	
487[1]	0.0016	34.7	
ECFP feature number	**RF score (Dual-activity model)**	**Z-score on dual database**	**ECFP features for dual-activity**
118	0.0015	7.64	
82	0.0015	5.49	
31	0.0013	2.63	
344	0.0013	4.42	
364	0.0010	5.71	

[1]ECFP features were also identified as important via RFE for ABS activity prediction. [2] ECFP features were also identified as important via RFE for dual-activity prediction

Most of these unique or shared ECFP features where again reconfirmed with RFE, further highlighting the importance of these features for stage-specific activity (Table 4.4, Figure 4.6A). It should be noted that RFE also highlighted features enriched within inactive

compounds (Figure 4.6B), hence, we believe using the Z-score to determine the enrichment of features in active *vs* inactive compounds together with model feature importance scores enabled us to better differentiate between features important for compound activity and inactivity.

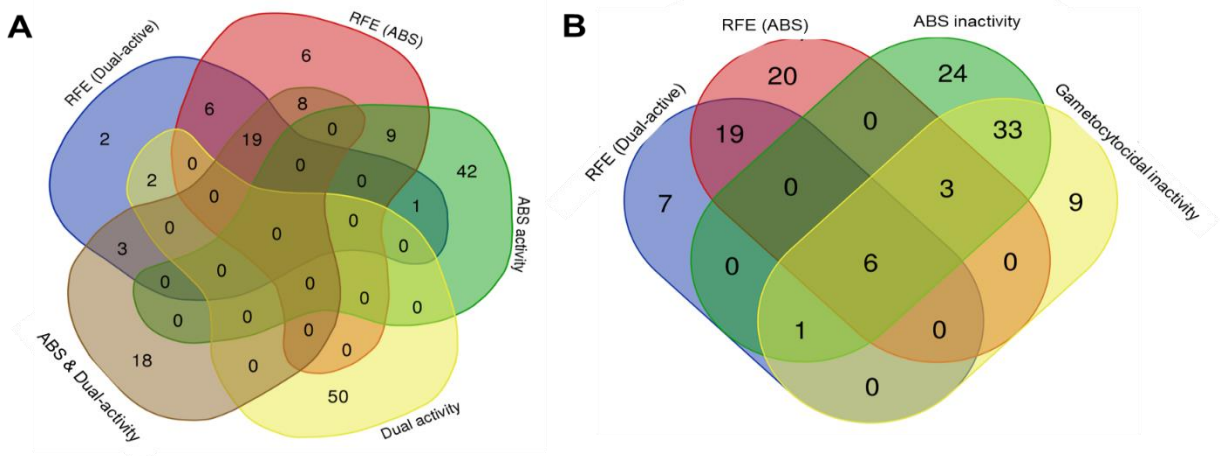


Figure 4.6: Overlap of ECFP features identified though RFE and unique/shared enriched ECFP features associated with stage-specific-activity or inactivity.
(A) ECFP features identified as important for ABS and/or dual-activity prediction via RFE are compared for any overlap to ECFP features identified as important features for ABS inhibition or dual-activity. (B) Similary, ECFP features identified as important for ABS and/or dual-activity prediction via RFE are compared for any overlap to ECFP features identified as important features for enriched within inactive compounds. Overlap between ECFP features for stage-specific activity and those identified through RFE is shown in Table 4.4.

Table 4.4: Overlap between important ECFP features associated with stage-specific-activity or inactivity and ECFP features identified through recursive feature elimination

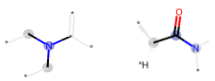
Group overlap	# of features	ECFP feature number	Unique or shared
Both ABS & dual-activity, RFE (ABS), RFE (Dual-active)	19	200, 194, 349, 81, 76, 311, 345, 191, 326, 195, 21, 99, 377, 232, 375, 303, 496, 491, 332	Shared
RFE (ABS), RFE (Dual-active)	6	272, 139, 230, 52, 490, 94	Shared
ABS activity RFE (Dual-active)	1	67	Shared
Dual-activity, RFE (Dual-active)	2	95, 149	Shared
Both ABS & dual-activity, RFE (Dual-active)	3	434, 495, 216	Shared
ABS activity, RFE (ABS)	9	161, 346, 69, 234, 487, 209, 80, 295, 50	Shared
Both ABS & dual-activity, RFE (ABS)	8	298, 78, 291, 178, 314, 70, 310, 213	Shared
RFE (Dual-active)	2	127, **189***	Unique
RFE (ABS)	6	27, 351, 302, **229***, **28***, **34***	Unique
ABS activity	42	90, 206, 71, 125, 190, 437, 463, 313, 231, 89, 343, 11, 93, 275, 197, 114, 226, 337, 321, 101, 62, 236, 323, 262, 388, 119, 457, 264, 74, 277, 92, 414, 221, 39, 173, 380, 53, 387, 239, 158, 399, 265	Unique
Dual-activity	50	118, 55, 84, 20, 31, 397, 65, 203, 261, 58, 340, 82, 339, 471, 218, 361, 187, 404, 282, 121, 79, 344, 364, 355, 479, 486, 455, 227, 460, 297, 334, 488, 61, 477, 342, 360, 312, 56, 1, 416, 100, 421, 254, 392, 318, 13, 257, 473, 3, 408	Unique
Both ABS & dual-activity	18	424 325, 153, 459, 241, 112, 176, 367, 256, 324, 447, 167, 260, 484, 155, 328 354, 242	Shared

*Bold feature numbers indicate ECFP features identified by RFE which were enriched within inactive compounds

4.3.3) Top predictive ECFP features enriched in inactive compounds

Also noteworthy was that the chemical features enriched in compounds that are inactive against ABS parasites (67) and/or gametocytes (52) tended to overlap with one another (Table 4.5). For example, ECFP feature nr. 94, 139, 299 and 287 appear as the top ECFP features for inactivity against both ABS and gametocyte stages. Generally compared to the ECFP features for activity, these features indicate that a lack of bioactivity may be associated with nitrogen being unable to bind to hydrogen or structures containing amides and branching carbon chains.

Table 4.5: Top 6 ECFP features associated with stage-specific inactivity

ECFP feature number	RF score (ABS activity model)	Z- score on ABS database	ECFP features for ABS inactivity
94	0.009	-63.7	
490	0.004	-42.3	
287	0.003	-39.2	
229	0.003	-52.1	
139	0.003	-46.2	
52	0.002	-40.9	
ECFP feature number	**RF score (Dual-activity model)**	**Z-score on dual database**	**ECFP features for gametocyte inactivity**
230	0.006	-15.9	
94	0.005	-22.0	
139	0.004	-18.0	
315	0.003	-9.1	
287	0.003	-13.6	
229	0.002	-12.5	

4.4) Discussion

Using our models in Chapter 3, we successfully identified chemical ECFP features enriched within ABS and/or dual-active compounds and predictive of stage-specific activity. We believe these features are indeed needed for activity as not only have these features been repeatedly identified through different ML algorithms and through recursive feature elimination, but they have also been shown to be significantly enriched within active compounds. The most notable commonality among enriched features for dual-activity was the nitration of heterocyclic and carbon groups that may aid in interactions with cellular targets [169]. In the same way, features enriched within ABS inhibiting compounds and important for predicting ABS inhibition showed that oxygenation was more common and may relate to solubility and drug uptake. We have also identified chemical features which enriched and predicted inactivity. With the identification of chemical features important for stage-specific activity and those associated with inactivity, we hope these features will aid medicinal chemists in guiding compound derivatisation during hit-to-lead optimisation of stage-specific and/or dual-active candidates. As such, the full list of features identified for stage-specific activity and inactivity has been made available in our recent publication, which this study has contributed towards. Consideration must be taken, however, that the features described here lack the context of connectivity to one another and it is likely that a combination of such chemical features is needed for stage-specific activity. One solution to this could be using unsupervised ML techniques such as association rule mining [170], to determine important relationships between these features and how they relate to compound activity. With this in mind, we hope to clarify how these ECFP features should be combined to achieve stage-specific activity in the future. Another limitation regarding EFCP features, which are 2D fingerprints, means that the features identified are not sensitive to the stereochemistry of compounds.

CHAPTER 5

CONCLUDING DISCUSSION

Treatment of malaria heavily relies on artemisinin combination therapies (ACT) to prevent further resistance formation, however, partial resistance via delayed clearance of the parasite has been observed within multiple areas across Southeast India and African countries and has been linked to the Kelch13 mutation [171-174]. Complete resistance towards artemisinin will have a devastating impact on treatment and control strategies if no drug substitutes can be identified. Additionally, reports of increased gametocytaemia of parasites under drug pressure [175] hints at sexual commitment of parasites to function as an escape from drug pressure and aid in resistance formation and spread. Treatments against ABS are mostly ineffective against gametocytes due to their parasite morphology and cellular biology being different from that of the ABS [176, 177]. To truly move towards malaria eradication, compounds which target the ABS and the transmissible gametocyte stages of the parasite are urgently needed.

Due to the parasite's unique life cycle and biology, the identification of gametocytocidal compounds is fraught with hurdles compared to identifying compounds with activity against ABS parasites. For one, the cultivation of large batches of gametocytes for phenotypic screening is hindered due to the low commitment rate of the parasite. Secondly, during gametocytogenesis, the parasite differentiates into multiple different forms during this process, and compounds need to be screened against each of these forms to determine whether they show activity towards a particular stage during sexual development. Lastly, in contrast to ABS screening, which can be conducted within a few days, the culturing and assaying against gametocyte stages can take days, if not weeks. As a result, the cost and time associated with screening a set number of compounds against ABS pales compared to screening the same number of compounds against the sexual forms of the parasite.

Much cost and time is wasted on the screening of compounds with no activity against gametocytes. A tool that could prioritise the screening of compounds most likely to show activity towards the gametocyte stages of the parasite would be a powerful resource to

accelerate antimalarial transmission-blocking drug discovery. Such a tool can prevent unnecessary expenditure and the reallocation of resources towards screening more promising compounds. Within various disease fields, ML has proven to be influential in accelerating drug discovery in predicting compound activity and compound prioritisation, provided enough chemical inhibition data is available. We theorised that ML could be used to develop robust models for predicting compounds with dual-activity against the *P. falciparum* parasite. The development of such models is complicateddue to limited gametocytocidal phenotypic screening data available, which results in poor performance of models. This is further exasperated by the severe class imbalance within phenotypic screening data where the majority of compounds are inactive, with few compounds having activity against the gametocyte stages of the parasite.

We believed that TL and proper sampling of inactive compounds could alleviate such limitations within gametocytocidal screening data for robust model building. To our surprise, techniques such as TL did not offer better performance over traditional ML algorithms; rather, the inherent class imbalance presented within phenotypic screening data proved to be a more daunting challenge for ML model building. By employing cluster-based undersampling to address class imbalance within the data and using ML algorithms more suited for training on imbalance data, we obtained models with good recall, accuracy, and sensitivity in predicting ABS and dual-activity and proved our hypothesis. More importantly, these models ensured that compounds with activity were enriched within the top list of compounds to be screened. This was true even in extreme datasets more chemically distinct and novel from model training data. Hence, these models can be used as a pre-screening tool to guide compound prioritisation during gametocytocidal phenotypic screening. Our models would be highly beneficial in reducing wasteful expenditure in the screening of inactive compounds against gametocytes and optimise gametocytocidal screening campaigns by reallocating such cost and time towards more promising compounds highlighted via compound prioritisation.

Additionally, these models have been made available as open access on the Ersillia Hub website to accelerate antimalarial and infectious disease drug discovery. Although the chemical space these models capture is limited, we aim to keep them updated with newly curated phenotypic screening data. Other than predicting the stage-specific activity of compounds, these models can also serve as starting points to guide generative modelling

through reinforcement learning in the hope of *de novo* generation chemical libraries targeted towards ABS or dual-activity.

Another hurdle within antimalarial drug discovery is the trial-and-error we often observe during hit-to-lead and lead optimisation, where candidates undergo chemical modifications to increase potency and/or add additional stage-specific activity. Considering that these models have a good understanding of the chemical space for stage-specific bioactivity and inactivity, we further mined these models via feature analysis to identify predictive and enriched features for stage-specific activity and inactivity. We have made these enriched and predictive features for stage-specific activity available with the hopes that it will guide medicinal chemists during compound derivatisation to allow better potency or to add additional stage-specific activity. We also highlighted features indicative of compound inactivity to dissuade the use of such features in compound derivatisation to improve compound activity. These features do not, however, account for stereochemistry and do not highlight any association of multiple chemical features that could contribute to stage-specific activity, and we aim to address this in future studies.

In summary, the tools provided within this PhD study will contribute significantly to accelerating gametocytocidal drug discovery by aiding in compound prioritisation and derivatisation.

References

1. *Technology development driving genomics and life sciences.* Cell Genomics, 2021. **1**(1): p. 100009.
2. Joslin, J., et al., *A Fully Automated High-Throughput Flow Cytometry Screening System Enabling Phenotypic Drug Discovery.* SLAS Discovery, 2018. **23**(7): p. 697-707.
3. Bellman, R., *A new type of approximation leading to reduction of dimensionality in control processes.* Journal of Mathematical Analysis and Applications, 1969. **27**(2): p. 454-459.
4. Chicco, D., *Ten quick tips for machine learning in computational biology.* BioData Mining, 2017. **10**: p. 35.
5. Bzdok, D., N. Altman, and M. Krzywinski, *Statistics versus machine learning.* Nature Methods, 2018. **15**(4): p. 233-234.
6. Hochreiter, S., G. Klambauer, and M. Rarey, *Machine Learning in Drug Discovery.* Journal of Chemical Information and Modeling, 2018. **58**(9): p. 1723-1724.
7. Vamathevan, J., et al., *Applications of machine learning in drug discovery and development.* Nature Reviews Drug Discovery, 2019.
8. El Bouchefry, K. and R.S. de Souza, *Chapter 12 - Learning in Big Data: Introduction to Machine Learning*, in *Knowledge Discovery in Big Data from Astronomy and Earth Observation*, P. Škoda and F. Adam, Editors. 2020, Elsevier. p. 225-249.
9. Rodríguez-Pérez, R. and J. Bajorath, *Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions.* Journal of Computer-Aided Molecular Design, 2020. **34**(10): p. 1013-1026.
10. Zhang, J., et al., *Application of pharmacodynamics-based optimization to the extraction of bioactive compounds from Chansu.* Microchemical Journal, 2020. **159**: p. 105552.
11. Dhandore, A., P. Mhatre, and K. Bhole. *Prediction of Drug Events using Machine Learning*. in *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*. 2022.
12. Gombar, V.K. and S.D. Hall, *Quantitative Structure–Activity Relationship Models of Clinical Pharmacokinetics: Clearance and Volume of Distribution.* Journal of Chemical Information and Modeling, 2013. **53**(4): p. 948-957.
13. Wajima, T., et al., *Prediction of Human Pharmacokinetics from Animal Data and Molecular Structural Parameters using Multivariate Regression Analysis: Oral Clearance.* Journal of Pharmaceutical Sciences, 2003. **92**(12): p. 2427-2440.
14. Comesana, A.E., et al., *A systematic method for selecting molecular descriptors as features when training models for predicting physiochemical properties.* Fuel, 2022. **321**: p. 123836.
15. Kumari, P., A. Nath, and R. Chaube, *Identification of human drug targets using machine-learning algorithms.* Computers in Biology and Medicine, 2015. **56**: p. 175-181.
16. Chen, R., et al., *Machine Learning for Drug-Target Interaction Prediction.* Molecules, 2018. **23**(9): p. 2208.
17. Tang, W., J. Chen, and H. Hong, *Development of classification models for predicting inhibition of mitochondrial fusion and fission using machine learning methods.* Chemosphere, 2021. **273**: p. 128567.

18. Wang, Z., et al., *In Silico Prediction of Blood–Brain Barrier Permeability of Compounds by Machine Learning and Resampling Methods.* ChemMedChem, 2018. **13**(20): p. 2189-2201.
19. Liu, M., et al., *Prediction of hERG potassium channel blockage using ensemble learning methods and molecular fingerprints.* Toxicology Letters, 2020. **332**: p. 88-96.
20. Sanders, H. and J. Saxe, *Garbage in, garbage out: how purportedly great ML models can be screwed up by bad data.* Technical report, 2017.
21. Gholami, R. and N. Fakhari, *Chapter 27 - Support Vector Machine: Principles, Parameters, and Applications*, in *Handbook of Neural Computation*, P. Samui, S. Sekhar, and V.E. Balas, Editors. 2017, Academic Press. p. 515-535.
22. Rodríguez-Pérez, R., M. Vogt, and J. Bajorath, *Support Vector Machine Classification and Regression Prioritize Different Structural Features for Binary Compound Activity and Potency Value Prediction.* ACS Omega, 2017. **2**(10): p. 6371-6379.
23. Knerr, S., L. Personnaz, and G. Dreyfus. *Single-layer learning revisited: a stepwise procedure for building and training a neural network*. in *Neurocomputing*. 1990. Berlin, Heidelberg: Springer Berlin Heidelberg.
24. Frunza, M.-C., *Chapter 2I - Support Vector Machines*, in *Solving Modern Crime in Financial Markets*, M.-C. Frunza, Editor. 2016, Academic Press. p. 205-215.
25. Yahyaoui's, A., I. Yahyaoui, and N. Yumuşak, *13 - Machine Learning Techniques for Data Classification*, in *Advances in Renewable Energies and Power Technologies*, I. Yahyaoui, Editor. 2018, Elsevier. p. 441-450.
26. Lapins, M., et al., *A confidence predictor for logD using conformal regression and a support-vector machine.* Journal of cheminformatics, 2018. **10**: p. 1-10.
27. Song, D., et al., *Similarity-based machine learning support vector machine predictor of drug-drug interactions with improved accuracies.* Journal of clinical pharmacy and therapeutics, 2019. **44**(2): p. 268-275.
28. McNamara, J.M., R.F. Green, and O. Olsson, *Bayes' theorem and its applications in animal behaviour.* Oikos, 2006. **112**(2): p. 243-251.
29. Webb, G.I., *Naïve Bayes*, in *Encyclopedia of Machine Learning*, C. Sammut and G.I. Webb, Editors. 2010, Springer US: Boston, MA. p. 713-714.
30. Zheng, F. and G.I. Webb, *Semi-Naive Bayesian Learning*, in *Encyclopedia of Machine Learning*, C. Sammut and G.I. Webb, Editors. 2010, Springer US: Boston, MA. p. 889-892.
31. Nylen, E.L. and P. Wallisch, *Chapter 7 - Regression*, in *Neural Data Science*, E.L. Nylen and P. Wallisch, Editors. 2017, Academic Press. p. 189-221.
32. Arthur, D.E., et al., *Activity and toxicity modelling of some NCI selected compounds against leukemia P388ADR cell line using genetic algorithm-multiple linear regressions.* Journal of King Saud University - Science, 2020. **32**(1): p. 324-331.
33. Louis, B., V.K. Agrawal, and P.V. Khadikar, *Prediction of intrinsic solubility of generic drugs using MLR, ANN and SVM analyses.* European Journal of Medicinal Chemistry, 2010. **45**(9): p. 4018-4025.
34. Robert, B.M., et al., *Computational models for predicting anticancer drug efficacy: A multi linear regression analysis based on molecular, cellular and clinical data of oral squamous cell carcinoma cohort.* Computer Methods and Programs in Biomedicine, 2019. **178**: p. 105-112.
35. Shen, M., et al., *Development and Validation of k-Nearest-Neighbor QSPR Models of Metabolic Stability of Drug Candidates.* Journal of Medicinal Chemistry, 2003. **46**(14): p. 3013-3020.

36. Priya, S., et al., *Machine learning approaches and their applications in drug discovery and design.* Chemical Biology & Drug Design, 2022. **100**(1): p. 136-153.
37. Breiman, L., *Random Forests.* Machine Learning, 2001. **45**(1): p. 5-32.
38. Genuer, R., et al., *Random Forests for Big Data.* Big Data Research, 2017. **9**: p. 28-46.
39. Fratello, M. and R. Tagliaferri, *Decision Trees and Random Forests*, in *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, et al., Editors. 2019, Academic Press: Oxford. p. 374-383.
40. Khoshgoftaar, T.M., J.V. Hulse, and A. Napolitano, *Comparing Boosting and Bagging Techniques With Noisy and Imbalanced Data.* IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, 2011. **41**(3): p. 552-568.
41. Babajide Mustapha, I. and F. Saeed, *Bioactive molecule prediction using extreme gradient boosting.* Molecules, 2016. **21**(8): p. 983.
42. Riddick, G., et al., *Predicting in vitro drug sensitivity using Random Forests.* Bioinformatics, 2011. **27**(2): p. 220-224.
43. Alghushairy, O., et al., *Machine learning-based model for accurate identification of druggable proteins using light extreme gradient boosting.* Journal of Biomolecular Structure and Dynamics, 2023: p. 1-12.
44. Basheer, I.A. and M. Hajmeer, *Artificial neural networks: fundamentals, computing, design, and application.* Journal of Microbiological Methods, 2000. **43**(1): p. 3-31.
45. Choi, J., S. Park, and J. Ahn, *RefDNN: a reference drug based neural network for more accurate prediction of anticancer drug resistance.* Scientific reports, 2020. **10**(1): p. 1861.
46. Cáceres, E.L., M. Tudor, and A.C. Cheng, *Deep learning approaches in predicting ADMET properties.* Future Medicinal Chemistry, 2020. **12**(22): p. 1995-1999.
47. Aliper, A., et al., *Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data.* Molecular pharmaceutics, 2016. **13**(7): p. 2524-2530.
48. Chen, H., et al., *The rise of deep learning in drug discovery.* Drug Discovery Today, 2018. **23**(6): p. 1241-1250.
49. Shi, G., *Chapter 3 - Artificial Neural Networks*, in *Data Mining and Knowledge Discovery for Geoscientists*, G. Shi, Editor. 2014, Elsevier: Oxford. p. 54-86.
50. Van Der Maaten, L., E.O. Postma, and H.J. van den Herik, *Dimensionality reduction: A comparative review.* Journal of Machine Learning Research, 2009. **10**(66-71): p. 13.
51. Howley, T., et al. *The effect of principal component analysis on machine learning accuracy with high dimensional spectral data*. in *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. 2005. Springer.
52. Manelfi, C., et al., *"Molecular Anatomy": a new multi-dimensional hierarchical scaffold analysis tool.* J Cheminform, 2021. **13**: p. 54.
53. Murima, P., et al., *Exploring the Mode of Action of Bioactive Compounds by Microfluidic Transcriptional Profiling in Mycobacteria.* PLoS ONE, 2013. **8**(7): p. e69191.
54. Kim, J., S. Zhao, and S. Heber. *Finding association rules of cis-regulatory elements involved in alternative splicing*. in *Proceedings of the 45th annual southeast regional conference*. 2007.
55. Vougas, K., et al., *Machine learning and data mining frameworks for predicting drug response in cancer: An overview and a novel in silico screening process based on association rule mining.* Pharmacology & Therapeutics, 2019. **203**: p. 107395.

56. Reddy, Y., P. Viswanath, and B.E. Reddy, *Semi-supervised learning: A brief review.* Int. J. Eng. Technol, 2018. **7**(1.8): p. 81.
57. Sahoo, P., et al., *MultiCon: A Semi-Supervised Approach for Predicting Drug Function from Chemical Structure Analysis.* Journal of Chemical Information and Modeling, 2020. **60**(12): p. 5995-6006.
58. Bao, J., et al. *CVAE-GAN: fine-grained image generation through asymmetric training*. in *Proceedings of the IEEE international conference on computer vision*. 2017.
59. Doersch, C., *Tutorial on variational autoencoders.* arXiv preprint arXiv:1606.05908, 2016.
60. Olivecrona, M., et al., *Molecular de-novo design through deep reinforcement learning.* Journal of cheminformatics, 2017. **9**(1): p. 48-48.
61. Popova, M., O. Isayev, and A. Tropsha, *Deep reinforcement learning for de novo drug design.* Science Advances, 2018. **4**(7): p. eaap7885.
62. Atance, S.R., et al., *De Novo Drug Design Using Reinforcement Learning with Graph-Based Deep Generative Models.* Journal of Chemical Information and Modeling, 2022. **62**(20): p. 4863-4872.
63. Korshunova, M., et al., *Generative and reinforcement learning approaches for the automated de novo design of bioactive compounds.* Communications Chemistry, 2022. **5**(1): p. 129.
64. Gupta, A., et al., *Generative Recurrent Networks for De Novo Drug Design.* Mol Inform, 2018. **37**(1-2).
65. Grisoni, F., et al., *Designing Anticancer Peptides by Constructive Machine Learning.* ChemMedChem, 2018. **13**(13): p. 1300-1302.
66. Yu, Y., et al., *A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures.* Neural Computation, 2019. **31**(7): p. 1235-1270.
67. Gupta, A., et al., *Generative Recurrent Networks for De Novo Drug Design.* Molecular informatics, 2018. **37**(1-2): p. 1700111.
68. Hadjeres, G. and F. Nielsen, *Anticipation-RNN: enforcing unary constraints in sequence generation, with application to interactive music generation.* Neural Computing and Applications, 2020. **32**(4): p. 995-1005.
69. Liu, X., et al., *An exploration strategy improves the diversity of de novo ligands using deep reinforcement learning: a case for the adenosine A(2A) receptor.* J Cheminform, 2019. **11**(1): p. 35.
70. Dairi, A., et al., *Comparative study of machine learning methods for COVID-19 transmission forecasting.* Journal of Biomedical Informatics, 2021. **118**: p. 103791.
71. Aghdam, R., M. Habibi, and G. Taheri, *Using informative features in machine learning based method for COVID-19 drug repurposing.* Journal of Cheminformatics, 2021. **13**(1): p. 70.
72. Mohapatra, S., et al., *Repurposing therapeutics for COVID-19: Rapid prediction of commercially available drugs through machine learning and docking.* PLOS ONE, 2020. **15**(11): p. e0241543.
73. Floresta, G., et al., *Artificial Intelligence Technologies for COVID-19 De Novo Drug Design.* International Journal of Molecular Sciences, 2022. **23**(6): p. 3261.
74. Marcos, M., et al., *Development of a severity of disease score and classification model by machine learning for hospitalized COVID-19 patients.* PLOS ONE, 2021. **16**(4): p. e0240200.
75. Klaudel, B., et al. *COVID-19 severity forecast based on machine learning and complete blood count data*. in *International Conference on Diagnostics of Processes and Systems*. 2022. Springer.

76. Lorenzen, S.S., et al., *Using machine learning for predicting intensive care unit resource use during the COVID-19 pandemic in Denmark.* Scientific Reports, 2021. **11**(1): p. 18959.
77. Zare, M., et al., *A machine learning-based system for detecting leishmaniasis in microscopic images.* BMC Infectious Diseases, 2022. **22**(1): p. 48.
78. Haro, P., N. Hevia-Montiel, and J. Perez-Gonzalez, *ECG Marker Evaluation for the Machine-Learning-Based Classification of Acute and Chronic Phases of Trypanosoma cruzi Infection in a Murine Model.* Tropical Medicine and Infectious Disease, 2023. **8**(3): p. 157.
79. Garcia, F.P., G.P. Guedes, and K.T. Belloze. *Identifying Schistosoma mansoni Essential Protein Candidates Based on Machine Learning*. in *Advances in Bioinformatics and Computational Biology*. 2020. Cham: Springer International Publishing.
80. Bosc, N., et al., *MAIP: a web service for predicting blood-stage malaria inhibitors.* J Cheminform, 2021. **13**(1): p. 13.
81. Mughal, H., et al., *Random Forest Model Predictions Afford Dual-Stage Antimalarial Agents.* ACS Infect Dis, 2022. **8**(8): p. 1553-1562.
82. Keshavarzi Arshadi, A., et al., *DeepMalaria: Artificial Intelligence Driven Discovery of Potent Antiplasmodials.* Front Pharmacol, 2019. **10**: p. 1526.
83. de Souza, A.S., et al., *Quantitative Structure–Activity Relationships for Structurally Diverse Chemotypes Having Anti-Trypanosoma cruzi Activity.* International Journal of Molecular Sciences, 2019. **20**(11): p. 2801.
84. Zorn, K.M., et al., *A Machine Learning Strategy for Drug Discovery Identifies Anti-Schistosomal Small Molecules.* ACS Infectious Diseases, 2021. **7**(2): p. 406-420.
85. Tse, E.G., et al., *An Open Drug Discovery Competition: Experimental Validation of Predictive Models in a Series of Novel Antimalarials.* J Med Chem, 2021. **64**(22): p. 16450-16463.
86. Shah, P., S. Tiwari, and M.I. Siddiqi, *Integrating molecular docking, CoMFA analysis, and machine-learning classification with virtual screening toward identification of novel scaffolds as Plasmodium falciparum enoyl acyl carrier protein reductase inhibitor.* Medicinal Chemistry Research, 2014. **23**(7): p. 3308-3326.
87. Jamal, S. and V. Scaria, *Cheminformatic models based on machine learning for pyruvate kinase inhibitors of Leishmania mexicana.* BMC Bioinformatics, 2013. **14**(1): p. 329.
88. Lima, M.N.N., et al., *Integrative Multi-Kinase Approach for the Identification of Potent Antiplasmodial Hits.* Front Chem, 2019. **7**: p. 773.
89. Park, H.S., et al., *Automated detection of P. falciparum using machine learning algorithms with quantitative phase images of unstained cells.* PloS one, 2016. **11**(9): p. e0163045.
90. Fuhad, K.M.F., et al., *Deep Learning Based Automatic Malaria Parasite Detection from Blood Smear and its Smartphone Based Application.* Diagnostics (Basel), 2020. **10**(5).
91. Sahu, S., et al., *Discovery of potential 1, 3, 5-Triazine compounds against strains of Plasmodium falciparum using supervised machine learning models.* European Journal of Pharmaceutical Sciences, 2020. **144**: p. 105208.
92. Maindola, P., S. Jamal, and A. Grover, *Cheminformatics Based Machine Learning Models for AMA1-RON2 Abrogators for Inhibiting Plasmodium falciparum Erythrocyte Invasion.* Molecular Informatics, 2015. **34**(10): p. 655-664.
93. Poostchi, M., et al., *Image analysis and machine learning for detecting malaria.* Translational Research, 2018. **194**: p. 36-55.

94. Mason, D.J., et al., *Using machine learning to predict synergistic antimalarial compound combinations with novel structures.* Frontiers in pharmacology, 2018. **9**: p. 1096.
95. Cox, F., *History of the discovery of the malaria parasites and their vectors.* Parasit Vectors, 2010. **3**.
96. Josling, G.A. and M. Llinas, *Sexual development in Plasmodium parasites: knowing when it's time to commit.* Nat Rev Micro, 2015. **13**(9): p. 573-587.
97. Guttery, D.S., A.A. Holder, and R. Tewari, *Sexual Development in Plasmodium: Lessons from Functional Analyses.* PLoS Pathogens, 2012. **8**(1): p. e1002404.
98. Billker, O., et al., *Calcium and a calcium-dependent protein kinase regulate gamete formation and mosquito transmission in a malaria parasite.* Cell, 2004. **117**(4): p. 503-14.
99. Billker, O., et al., *The roles of temperature, pH and mosquito factors as triggers of male and female gametogenesis of Plasmodium berghei in vitro.* Parasitology, 1997. **115**.
100. Barnes, K.I., et al., *Increased gametocytemia after treatment: an early parasitological indicator of emerging sulfadoxine-pyrimethamine resistance in falciparum malaria.* J Infect Dis, 2008. **197**.
101. Beri, D., B. Balan, and U. Tatu, *Commit, hide and escape: the story of Plasmodium gametocytes.* Parasitology, 2018. **145**(13): p. 1772-1782.
102. van der Watt, M.E., J. Reader, and L.-M. Birkholtz, *Adapt or Die: Targeting Unique Transmission-Stage Biology for Malaria Elimination.* Frontiers in Cellular and Infection Microbiology, 2022. **12**.
103. Birkholtz, L.-M., P. Alano, and D. Leroy, *Transmission-blocking drugs for malaria elimination.* Trends in Parasitology, 2022. **38**(5): p. 390-403.
104. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* Journal of artificial intelligence research, 2002. **16**: p. 321-357.
105. Haibo, H., et al. *ADASYN: Adaptive synthetic sampling approach for imbalanced learning*. in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 2008.
106. Han, H., W.-Y. Wang, and B.-H. Mao. *Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning*. in *International conference on intelligent computing*. 2005. Springer.
107. Ahmed Saad, H., et al., *A-SMOTE: A New Preprocessing Approach for Highly Imbalanced Datasets by Improving SMOTE.* International Journal of Computational Intelligence Systems, 2019. **12**(2): p. 1412-1422.
108. Donyavi, Z. and S. Asadi, *Diverse training dataset generation based on a multi-objective optimization for semi-Supervised classification.* Pattern Recognition, 2020. **108**: p. 107543.
109. Tumuluru, P., et al. *Class Imbalance of Bio-Medical Data by Using PCA-Near Miss for Classification*. in *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*. 2023.
110. Kubat, M. and S. Matwin. *Addressing the curse of imbalanced training sets: one-sided selection*. in *Icml*. 1997. Citeseer.
111. Torrey, L. and J. Shavlik, *Transfer learning*, in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. 2010, IGI global. p. 242-264.
112. Khan, S., et al., *A novel deep learning based framework for the detection and classification of breast cancer using transfer learning.* Pattern Recognition Letters, 2019. **125**: p. 1-6.

113. Oh, K., et al., *Classification and visualization of Alzheimer's disease using volumetric convolutional neural network and transfer learning.* Scientific Reports, 2019. **9**(1): p. 1-16.
114. Maqsood, M., et al., *Transfer learning assisted classification and detection of Alzheimer's disease stages using 3D MRI scans.* Sensors, 2019. **19**(11): p. 2645.
115. Khan, N.M., N. Abraham, and M. Hon, *Transfer learning with intelligent training data selection for prediction of Alzheimer's disease.* IEEE Access, 2019. **7**: p. 72726-72735.
116. Kimura, N., et al., *Convolutional Neural Network Coupled with a Transfer-Learning Approach for Time-Series Flood Predictions.* Water, 2020. **12**(1): p. 96.
117. Cai, C., et al., *Transfer Learning for Drug Discovery.* Journal of Medicinal Chemistry, 2020. **63**(16): p. 8683-8694.
118. Azab, A.M., et al., *A review on transfer learning approaches in brain–computer interface.* Signal processing and machine learning for brain-machine interfaces, 2018: p. 81-98.
119. Yuhu, C., et al., *Weighted Multi-source TrAdaBoost.* Chinese Journal of Electronics, 2013. **22**(3): p. 505-510.
120. Eaton, E. and M. desJardins. *Set-Based Boosting for Instance-Level Transfer*. in *2009 IEEE International Conference on Data Mining Workshops*. 2009.
121. Dey, V., R. Machiraju, and X. Ning, *Improving Compound Activity Classification via Deep Transfer and Representation Learning.* ACS Omega, 2022. **7**(11): p. 9465-9483.
122. Yang, X., et al., *Transfer learning via multi-scale convolutional neural layers for human-virus protein-protein interaction prediction.* Bioinformatics, 2021. **37**(24): p. 4771-4778.
123. Iman, M., H.R. Arabnia, and K. Rasheed, *A Review of Deep Transfer Learning and Recent Advancements.* Technologies, 2023. **11**(2): p. 40.
124. Kalbfleisch, J.D. and D.A. Sprott, *Marginal and Conditional Likelihoods.* Sankhyā: The Indian Journal of Statistics, Series A (1961-2002), 1973. **35**(3): p. 311-328.
125. Gilks, W.R., S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. 1995: Taylor & Francis.
126. Long, M., et al. *Transfer feature learning with joint distribution adaptation*. in *Proceedings of the IEEE international conference on computer vision*. 2013.
127. Ravikumar, B., et al., *Chemogenomic analysis of the druggable kinome and its application to repositioning and lead identification studies.* Cell chemical biology, 2019. **26**(11): p. 1608-1622. e6.
128. Aulner, N., et al., *Next-Generation Phenotypic Screening in Early Drug Discovery for Infectious Diseases.* Trends Parasitol, 2019. **35**(7): p. 559-570.
129. Birkholtz, L.-M., et al., *Discovering new transmission-blocking antimalarial compounds: challenges and opportunities.* Trends in parasitology, 2016. **32**(9): p. 669-681.
130. Abraham, M., et al., *Probing the Open Global Health Chemical Diversity Library for Multistage-Active Starting Points for Next-Generation Antimalarials.* ACS Infect Dis, 2020. **6**(4): p. 613-628.
131. Miguel-Blanco, C., et al. *Hundreds of dual-stage antimalarial molecules discovered by a functional gametocyte screen*. Nature communications, 2017. **8**, 15160 DOI: 10.1038/ncomms15160.
132. Van Voorhis, W.C., et al., *Open Source Drug Discovery with the Malaria Box Compound Collection for Neglected Diseases and Beyond.* PLoS Pathog, 2016. **12**(7): p. e1005763.

133. Gamo, F.-J., et al., *Thousands of chemical starting points for antimalarial lead identification.* Nature, 2010. **465**: p. 305.
134. Guiguemde, W.A., et al., *Chemical genetics of Plasmodium falciparum.* Nature, 2010. **465**(7296): p. 311-315.
135. Plouffe, D., et al., *In silico activity profiling reveals the mechanism of action of antimalarials discovered in a high-throughput screen.* Proc Natl Acad Sci U S A, 2008. **105**.
136. Reader, J., et al., *Multistage and transmission-blocking targeted antimalarials discovered from the open-source MMV Pandemic Response Box.* Nat Commun, 2021. **12**(1): p. 269.
137. Nguyen, M.-T., T. Nguyen, and T. Tran, *Learning to discover medicines.* International Journal of Data Science and Analytics, 2023. **16**(3): p. 301-316.
138. Chao, X. and L. Zhang, *Few-shot imbalanced classification based on data augmentation.* Multimedia Systems, 2023. **29**(5): p. 2843-2851.
139. Duffy, S. and V.M. Avery, *Identification of inhibitors of Plasmodium falciparum gametocyte development.* Malar J, 2013. **12**.
140. Duffy, S., et al., *Screening the Medicines for Malaria Venture Pathogen Box across Multiple Pathogens Reclassifies Starting Points for Open-Source Drug Discovery.* Antimicrob Agents Chemother, 2017. **61**(9).
141. Landrum, G., *RDKit: Open-source cheminformatics.* https://www.rdkit.org. 2006.
142. McInnes, L. and J. Healy, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.* 2018.
143. Chawla, N.V., et al., *SMOTE: synthetic minority over-sampling technique.* J. Artif. Intell. Res., 2002. **16**: p. 321-357.
144. Dai, W., et al. *Boosting for transfer learning*. in *Proceedings of the 24th international conference on Machine learning*. 2007.
145. Pedregosa, F., et al., *Scikit-learn: Machine Learning in Python.* J Mach Learn Res, 2011(12): p. 2825–2830.
146. Chollet, F., *keras*. 2015.
147. Abadi, M., et al., *TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow. org (2015).* URL https://www. tensorflow. org, 2015.
148. O'Malley, T., et al., *Keras Tuner. 2019.* Available online: github. com/keras-team/kerastuner (accessed on 2 April 2022), 2023.
149. Park, S. and H. Park, *Combined oversampling and undersampling method based on slow-start algorithm for imbalanced network traffic.* Computing, 2020. **103**(3): p. 401-424.
150. Krawczyk, B., *Learning from imbalanced data: open challenges and future directions.* Prog Artif Intell, 2016. **5**(4): p. 221-232.
151. Al-Stouhi, S. and C.K. Reddy, *Transfer Learning for Class Imbalance Problems with Inadequate Data.* Knowl Inf Syst, 2016. **48**(1): p. 201-228.
152. Kuang, J., et al., *Class-Imbalance Adversarial Transfer Learning Network for Cross-Domain Fault Diagnosis With Imbalanced Data.* IEEE Transactions on Instrumentation and Measurement, 2022. **71**: p. 1-11.
153. Wang, J., et al. *Balanced Distribution Adaptation for Transfer Learning*. in *2017 IEEE International Conference on Data Mining (ICDM)*. 2017.
154. Weiss, K.R. and T.M. Khoshgoftaar. *Comparing Transfer Learning and Traditional Learning Under Domain Class Imbalance*. in *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2017.
155. Xu, B., et al. *Cross-Project Aging-Related Bug Prediction Based on Joint Distribution Adaptation and Improved Subclass Discriminant Analysis*. in *2020*

IEEE 31st International Symposium on Software Reliability Engineering (ISSRE). 2020.
156. More, A.S. and D.P. Rana. *Review of random forest classification techniques to resolve data imbalance*. in *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*. 2017.
157. Brown, I. and C. Mues, *An experimental comparison of classification algorithms for imbalanced credit scoring data sets.* Expert Syst Appl, 2012. **39**(3): p. 3446-3453.
158. Korkmaz, S., *Deep Learning-Based Imbalanced Data Classification for Drug Discovery.* J Chem Inf Model, 2020. **60**(9): p. 4180-4190.
159. Esposito, C., et al., *GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning.* J Chem Inf Model, 2021. **61**(6): p. 2623-2640.
160. Bengfort, B. and R. Bilbro, *Yellowbrick: Visualizing the scikit-learn model selection process.* Journal of Open Source Software, 2019. **4**(35): p. 1075.
161. Fakoor, R., et al., *Fast, accurate, and simple models for tabular data via augmented distillation.* Adv Neural Inf Process Syst. , 2020. **33**: p. 8671-8681.
162. Zhang, C., et al. *An imbalanced data classification algorithm of improved autoencoder neural network*. in *2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)*. 2016.
163. Haibo, H. and E.A. Garcia, *Learning from Imbalanced Data.* IEEE Trans Knowl Data Eng, 2009. **21**(9): p. 1263-1284.
164. Sakr, G.E., et al. *Comparing deep learning and support vector machines for autonomous waste sorting*. in *2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET)*. 2016.
165. Liu, P., et al., *SVM or deep learning? A comparative study on remote sensing image classification.* Soft Comput, 2017. **21**(23): p. 7053-7065.
166. Turon, G. and M. Duran-Frigola, *Ersilia Model Hub: a repository of AI/ML models for neglected tropical diseases (v0.1.16).* 2023.
167. Lundberg, S.M. and S.-I. Lee, *A unified approach to interpreting model predictions.* Advances in neural information processing systems, 2017. **30**.
168. Karich, A., et al., *Benzene oxygenation and oxidation by the peroxygenase of Agrocybe aegerita.* AMB Express, 2013. **3**(1): p. 5.
169. Nepali, K., H.-Y. Lee, and J.-P. Liou, *Nitro-Group-Containing Drugs.* J Med Chem, 2019. **62**(6): p. 2851-2893.
170. Vora, L.K., et al., *Artificial Intelligence in Pharmaceutical Technology and Drug Delivery Design.* Pharmaceutics, 2023. **15**(7): p. 1916.
171. Fola, A.A., et al., *Plasmodium falciparum resistant to artemisinin and diagnostics have emerged in Ethiopia.* Nature Microbiology, 2023. **8**(10): p. 1911-1919.
172. Tacoli, C., et al., *Artemisinin resistance-associated K13 polymorphisms of Plasmodium falciparum in Southern Rwanda, 2010–2015.* Am J Trop Med Hyg, 2016. **95**.
173. Uwimana, A., et al., *Association of Plasmodium falciparum kelch13 R561H genotypes with delayed parasite clearance in Rwanda: an open-label, single-arm, multicentre, therapeutic efficacy study.* The Lancet Infectious Diseases, 2021. **21**(8): p. 1120-1128.
174. Sene, S.D., et al., *Identification of an in vitro artemisinin-resistant Plasmodium falciparum kelch13 R515K mutant parasite in Senegal.* Frontiers in Parasitology, 2023. **2**.
175. Portugaliza, H.P., et al., *Plasmodium falciparum sexual conversion rates can be affected by artemisinin-based treatment in naturally infected malaria patients.* EBioMedicine, 2022. **83**: p. 104198.

176. Gulati, S., et al., *Profiling the Essential Nature of Lipid Metabolism in Asexual Blood and Gametocyte Stages of Plasmodium falciparum.* Cell Host & Microbe, 2015. **18**(3): p. 371-381.
177. MacRae, J.I., et al., *Mitochondrial metabolism of sexual and asexual blood stages of the malaria parasite Plasmodium falciparum.* BMC Biol, 2013. **11**.