



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Stratifying antimalarial compounds with
similar mode of action using machine
learning on chemo-transcriptomic profiles

Ashleigh van Heerden

Student number: 14020590

2019

Submitted in partial fulfilment of the degree:

Magister Scientiae

Biochemistry, Genetics and Microbiology

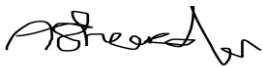
In the Faculty of Natural and Agricultural

Sciences

© University of Pretoria

i. Submission declaration

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

SIGNATURE: 

DATE: 2020/02/06

ii. Plagiarism statement

DECLARATION OF ORIGINALITY UNIVERSITY OF PRETORIA

The Department of Biochemistry places great emphasis upon integrity and ethical conduct in the preparation of all written work submitted for academic evaluation. While academic staff teach you about referencing techniques and how to avoid plagiarism, you too have a responsibility in this regard. If you are at any stage uncertain as to what is required, you should speak to your lecturer before any written work is submitted. You are guilty of plagiarism if you copy something from another author's work (eg a book, an article or a website) without acknowledging the source and pass it off as your own. In effect you are stealing something that belongs to someone else. This is not only the case when you copy work word-for-word (verbatim), but also when you submit someone else's work in a slightly altered form (paraphrase) or use a line of argument without acknowledging it. You are not allowed to use work previously produced by another student. You are also not allowed to let anybody copy your work with the intention of passing it off as his/her work. Students who commit plagiarism will not be given any credit for plagiarised work. The matter may also be referred to the Disciplinary Committee (Students) for a ruling. Plagiarism is regarded as a serious contravention of the University's rules and can lead to expulsion from the University. The declaration which follows must accompany all written work submitted while you are a student of the Department of Biochemistry, Genetics and Microbiology. No written work will be accepted unless the declaration has been completed and attached.

Full names of student: Ashleigh van Heerden

Student number: 14020590

Topic of work: Dissertation

Declaration:

1. I understand what plagiarism is and am aware of the University's policy in this regard.
2. I declare that this Dissertation is my own original work. Where other people's work has been used (either from a printed source, Internet or any other source), this has been properly acknowledged and referenced in accordance with departmental requirements.
3. I have not used work previously produced by another student or any other person to hand in as my own.
4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.

SIGNATURE



iii. Acknowledgements

I would like to acknowledge the following individuals and institutions:

My supervisor, Prof. L Birkholtz, for her insight and innovation that has aided me throughout this project, along with all the conferences and opportunities she had provided me with.

My co-supervisor, Rudi van Wyk, for his continual guidance and patience that helped me improve and comprehend bioinformatics and machine learning for this project.

The NRF, for the NRF-Grant holder linked bursary, provided by Prof. L Birkholtz, that has helped me financially to concentrate all my focus onto my studies and the completion of this project.

The scientists who have made their data publicly available which has allowed this study to be conducted.

The Lord Jesus Christ, who has given me hope and a future, whose faithful guiding hand has never forgotten or forsaken me. When I was weary and downtrodden He renewed my strength and motivation.

iv. Summary

Malaria is a terrible disease caused by a protozoan parasite within the *Plasmodium* genus, claiming the lives of hundreds of thousands of people yearly, the majority of whom are children under the age of five. Of the five species of *Plasmodium* causing malaria in humans, *P. falciparum* is responsible for most of the death toll. An increase in malaria cases was detected between the years 2016 to 2017 according to the World Malaria Report of 2017, despite control efforts. The rapid development of resistance within *P. falciparum* against antimalarials has led to the use of artemisinin combinational therapy as the current gold standard for malaria treatment. Yet decreased parasite clearance demonstrates that using combination therapy is insufficient in maintaining current antimalarials' effectiveness against these resistant parasites. Hence, novel compounds with a mode of action (MoA) different than current antimalarials are required. Though phenotypic screening has delivered thousands of promising hit compounds, hit-to-lead optimisation is still one of the rate-limiting steps in pre-clinical antimalarial drug development. While knowing the exact target or MoA is not required to progress a compound in a medicinal chemistry program, identifying the MoA early can accelerate hit prioritization, hit-to-lead optimisation and preclinical combination studies in malaria research. In this study, we assessed machine learning (ML) approaches for their ability to stratify antimalarials based on transcriptional responses associated with the treatments. From our results, we conclude that it is possible to identify biomarkers from the transcriptional responses that define the MoA of compounds. Moreover, only a limited set of 50 genes was required to build a ML model that can stratify compounds with similar MoA with a classification accuracy of $76.6 \pm 6.4\%$. These biomarkers will help stratify new compounds with similar MoA to those already defined with our strategy. Additionally, the biomarkers can also be used to monitor if the MoA of a compound has changed during hit-to-lead optimisation. This work will contribute to accelerating antimalarial drug discovery during the hit-to-lead optimisation phase and help the identification of compounds with novel MoA.

Table of contents

i. Submission declaration	I
ii. Plagiarism statement	II
iii. Acknowledgements	III
iv. Summary	IV
Chapter 1: Introduction	1
1.1 Malaria	1
1.2 The life cycle of the malaria parasite	1
1.3 Current malaria control strategies	3
1.3.1 Vector control	3
1.3.2 Vaccines	3
1.3.3 Overview of current antimalarial chemotherapeutics	4
1.4.1 Limitations in antimalarial drug discovery	6
1.5 Target and MoA identification strategies in drug discovery	8
1.7 The use of transcriptome datasets to identify a drug's MoA and targets	10
1.7.1 Gene expression correlations	10
1.7.2 Hierarchical clustering	11
1.7.3 Network analysis for MoA fingerprinting and target identification	12
1.7.4 Machine learning for MoA fingerprinting and classification	14
1.8 Rationale using <i>P. falciparum</i> transcriptome for MoA deconvolution	15
Hypothesis	19
Objectives	19
Chapter 2: Methodology	21
2.1 Identifying predictive biomarker genes for mode of action stratification	21
2.1.1 Quality control filtering	22
2.1.2 Merging and pre-processing (normalisation) of GEP datasets	23
2.2 Generating a predictive model	23
2.2.1 Employment of machine learning on GEPs	23
2.2.2 Selection of ML algorithms to be investigated	25
2.2.2.1 Multinomial logistic regression	26

2.2.2.2 Support vector classification	26
2.2.2.3 Random forest	26
2.2.2.4 Gradient boosting machine	27
2.2.2.5 Artificial neural networks	27
2.2.3 Evaluating different machine learning algorithms in stratifying antiplasmodial compounds similar MoA	28
2.2.3 Filtering criteria to identify predictive biomarker genes	30
2.3 Validation of rational feature selection through a comparison to algorithm inferred biomarkers	31
2.4 Optimisation of the minimum number of features for robust MoA stratification	32
Chapter 3: Results	33
3.1 Identifying predictive biomarker genes for MoA stratification	33
3.1.1 Data acquisition and quality control filtering	33
3.1.2 Evaluating different machine learning algorithms on the 2463-gene database	37
3.1.3 Biomarker gene selection	39
3.2 Building predictive biomarker models	43
3.2.1 Evaluating different machine learning algorithms for the 174-gene biomarker database	43
3.2.2 Validation of feature selection in the biomarker database	44
3.2.3 Optimisation of the number of features in our MLR biomarker model	46
3.2.4 Interrogation of the top 50 features from the MLR biomarker model as indicators of MoA	47
Chapter 4: Discussion	51
References	56
Appendix A	67
A.1 Machine learning theory	67
Hyperparameter testing	67
A.1.1 Principle of multiclassification support vector machines	67
A.1.2 Principle of multinomial logistic regression	69
A.1.3 Principle of random forest	70
A.1.4 Principle of gradient boosting machines	71
A.1.5 Principle of artificial neural networks	72
Appendix B	76

List of Figures

Figure 1: The <i>Plasmodium falciparum</i> life cycle.	2
Figure 2: Typical drug discovery pipeline with high-throughput phenotypic screening.	7
Figure 3: Simplified example of a network.	12
Figure 4: Supervised and Unsupervised machine learning algorithms.	14
Figure 5: Gene expression responses to drug treatment related to a compound's MoA and/or chemical structures.	16
Figure 6: Discernible expression patterns across treatments.	18
Figure 7: Principle of training a compound MoA stratification model using transcriptional responses of genes.	24
Figure 8: Method for selecting the best ML algorithm for MoA stratification.	25
Figure 9: Assessing model performance using K-fold cross-validation and untrained test data.	29
Figure 10: Feature selection filtering process to identify biomarker genes with unique predictive features.	30
Figure 11: Quality control summary of compound-treated <i>P. falciparum</i> GEP datasets acquired.	33
Figure 12: Merging and normalisation of accepted datasets to form our 2463-gene database	35
Figure 13: Normalisation strategies applied to the 2463-gene database.	36
Figure 14: Robustness and accuracy of different ML algorithms ability in stratifying treatments with similar MoA within our 2463-gene database.	38
Figure 15: Summary of feature selection and merging of datasets to form the 174-gene biomarker database.	39
Figure 16 : Filtering of DEGs to DEGs pervasive over time.	40
Figure 17: Identification of pervasive DEGs unique to individual treatments.	41
Figure 18: Pervasive DEGs compared to pervasive DEGs that are unique to treatments.	42
Figure 19: Robustness and accuracy of different ML algorithms ability in stratifying treatments with similar MoA using the 174-gene biomarker database.	43
Figure 20: ML-inferred features vs rationally selected features.	45
Figure 21: Influence of limiting the number of genes used for training on MoA stratification of MLR models.	46

Figure 22: Compound origin of the top 50 biomarker genes used in the final MLR model for MoA stratification.	48
Figure 23: Novelty of top 50 rationally selected biomarker genes	50
Figure 24: Principle of hyperparameter tuning.....	67
Figure 25: Principle of SVM classification.....	68
Figure 26: Support vector machine kernel function to separate nonlinear data	69
Figure 27: Multinomial logistic regression algorithm	70
Figure 28: Random forest employing bootstrap aggregation and multiple decision trees.	71
Figure 29: Simple artificial neural network	72
Figure 30: Performance of algorithms investigated for MoA classification built using either biomarker genes or all genes in the database.	77
Figure 31: Performance in MoA accuracy of minimodels made from biomarker or database genes using random forest and multinomial logistic regression algorithms (h2o R package).	78

List of Tables

Table 1: Antimalarial mode of action and parasite's resistance mode of action.....	4
Table 2: Properties of current target candidate profiles	5
Table 3: <i>P. falciparum</i> compound treated GEP datasets used in this study	21
Table 4: Filtering criteria of GEP datasets	22
Table 5: Final database generated from 6 datasets spanning 20 compound treatments .	34
Table 6: Top 50 biomarker genes encoded product	49

List of Abbreviations

ACT	Artemisinin combination therapy
ANN	Artificial neural network
DARTS	Drug affinity responsive target stability
DEG	Differentially expressed gene
DSEA	Drug-set Enrichment Analysis
FC	Fold change
FDR	False discovery rate
GBM	Gradient boosting model
GEO	Gene Expression Omnibus
GEP	Gene expression profile
GRN	Gene regulatory network
GSEA	Gene set enrichment analysis
H2L	Hit-to-lead optimisation
HDA	Histone deacetylases
HDD	High dimensional dataset
HPT	Hyperparameter tuning
IRS	Indoor residual spray
ITN	Insecticide-treated nets
LO	Lead optimisation
LR	Logistic regression
ML	Machine learning
MLR	Multinomial logistic regression
MMV	Medicines for malaria venture
MoA	Mode of action
PCA	Principal component analysis
PPIN	Protein-protein interaction network
RF	Random forest
SVC	Support vector classification
SVM	Support vector machine
TCP	Target candidate profile
TPP	Target product profile
WHO	World Health Organization

Chapter 1: Introduction

1.1 Malaria

Malaria is a disease that has resulted in the deaths of millions of people throughout human history [1, 2]. This disease is vector-borne and caused by a protozoan parasite within the *Plasmodium* genus, which is transmitted to humans by the female *Anopheles* mosquito [3]. Only 5 species within this genus are responsible for malaria within humans, namely, *P. knowlesi*, *P. vivax*, *P. ovale*, *P. falciparum*, and *P. malariae*, where infection with *P. falciparum* is the most severe [4, 5], due to the rapid development and sequestration of this species in humans that cause cerebral malaria and result in a coma or death [6, 7].

Tremendous progress has been made in decreasing clinical incidences of malaria by 40% within Africa from the year 2000 to 2015. This has been achieved through proper implementation of control strategies and the effective use of current antimalarial drugs [8]. However, according to the WHO Malaria Report of 2019, it has become evident that despite these efforts there has been an increase in malaria cases from 2016 to 2018 – 217 million cases in 2016 compared to the 228 million cases in 2018 [9]. Together with this, South Africa, which forms part of the 21 E-2020 countries that WHO identified as having the potential to eliminate malaria by 2020, reported increases in indigenous cases since 2015 [10]. This increase in malaria cases was also experienced by 10 other countries and there is a concern that this trend may derail the progress made in eliminating the disease from these countries.

Most cases and deaths occurred within the African continent and 78% of these deaths were children under the age of five [11-13]. Immune-compromised individuals and pregnant women also have a larger health risk when acquiring malaria [14]. In pregnant women, malaria infection can lead to an increased risk of abortion, stillbirth, pregnancy-related complications and low birth weight because of poor nutrient exchange between mother and child through the placenta [15, 16]. It is even suggested that malaria may directly contribute to almost 25% of all maternal deaths in regions where malaria is endemic [14].

1.2 The life cycle of the malaria parasite

Malaria is acquired when an infected female *Anopheles* mosquito injects her proboscis into a human for a blood meal as shown in Figure 1. The sporozoite forms of the parasite, present within the mosquito's salivary glands, are injected into the human bloodstream and transported to the liver where they infect hepatocytes [17].

Within the hepatocytes, sporozoites undergo asexual exo-erythrocytic schizogony to form multiple merozoites, that are released into the bloodstream upon the rupture of the hepatocyte [17, 18]. These merozoites invade erythrocytes and initiate intra-erythrocytic schizogony, where the parasite develops from ring stages to metabolically and transcriptionally active trophozoite stages [19]. During the subsequently schizont stage, synchronous nuclear division occurs [20, 21], followed by a single final cell division and cytokinesis, that results in the release of multiple daughter merozoites into the bloodstream [21-24]. These merozoites reinitiate the intraerythrocytic development cycle by infecting uninfected erythrocytes.

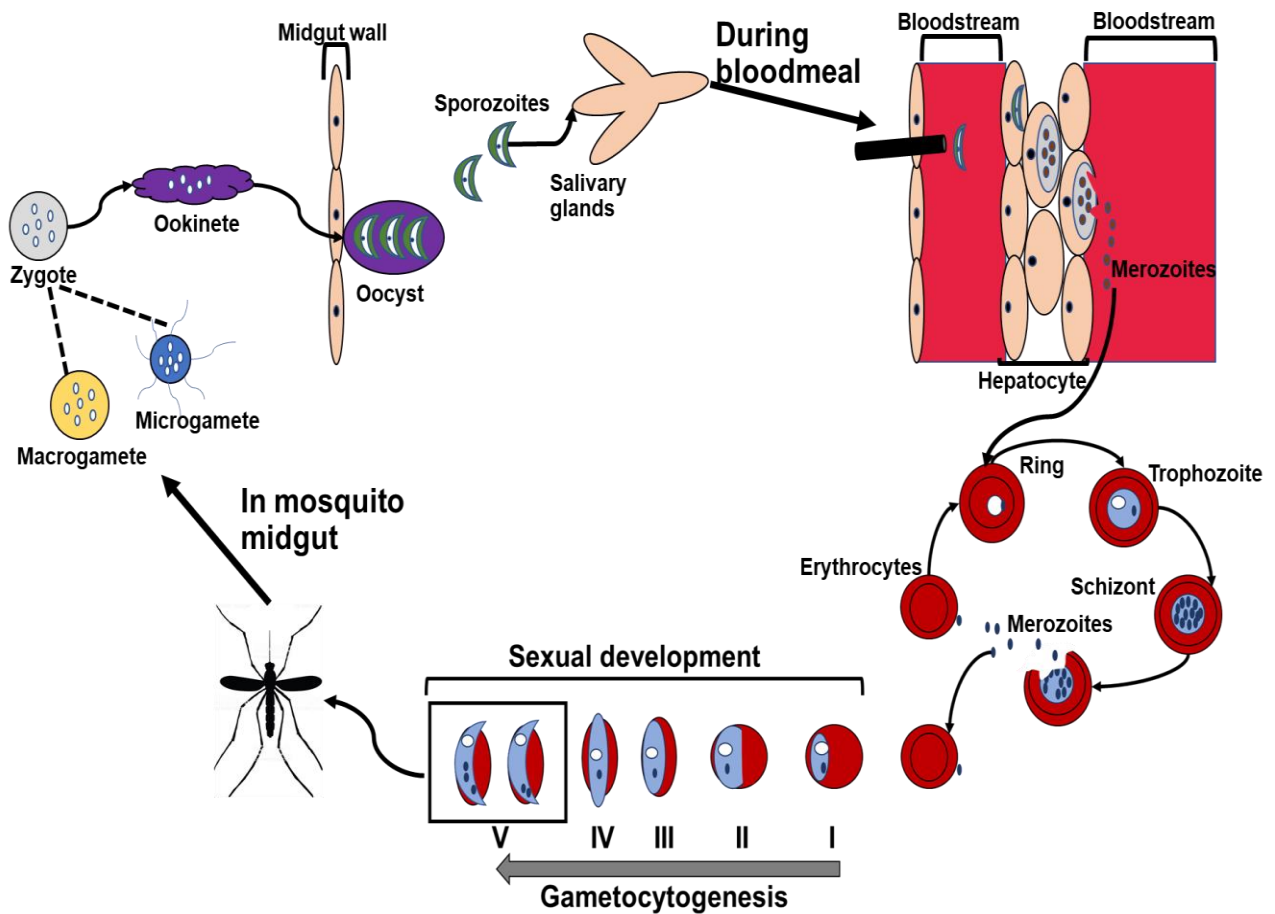


Figure 1: The *Plasmodium falciparum* life cycle.

P. falciparum infection initiates when sporozoites are released into the human bloodstream from an infected female mosquito's salivary glands during feeding. Sporozoites are then migrated to the liver where they invade hepatocytes to form liver schizonts. Infected hepatocytes rupture releasing merozoites into the bloodstream that invade erythrocytes and form the ring stage. The ring stage develops into a trophozoite and then a schizont during asexual multiplication within the infected erythrocyte. Schizonts rupture releasing merozoites that infect erythrocytes. A portion of invading merozoites are sexually committed and develop into the sexual blood stages to produce stage V mature male and female gametocytes that are transmitted to a mosquito during feeding. Mature gametocytes upon ingestion develop into micro- and macrogametes that fuse to form a zygote and develops into an ookinete. This mobile ookinete 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.

The continual rupture of erythrocytes and clearance of infected erythrocytes by the spleen in the asexual intra-erythrocytic cycle causes the majority of symptoms of malaria such as anaemia [14]. Cerebral malaria, the deadliest form of the disease, results from the cytoadherence of erythrocytes to vascular walls causing sequestration of infected erythrocytes in small blood vessels [25].

A small portion (<10%) of invading merozoites do not develop into the asexual stages, but rather are committed to the sexual development of the parasite to form gametocytes with five distinctive stages in *P. falciparum* [18]. Stage V, the final stage of gametocyte development, is essential for the transmission of the parasite from an infected human to the female *Anopheles* mosquito. After ingestion of stage V gametocytes by the mosquito, the mature gametocytes receive environmental signals and mosquito factors within the mosquito midgut. These signals activate the male and female gametocytes to mature into male and female gametes in a process known as gametogenesis. The fusion of male and female gametes give rise to a zygote [26-28], that develops into a mobile ookinete which migrates to the midgut wall that it penetrates to mature into an oocyst. Another asexual multiplication occurs within the oocyst to produce sporozoites that are released upon the rupture of the oocyst [17]. The released sporozoites travel to the salivary glands of the mosquito where it can be transferred during the mosquito's next blood meal.

1.3 Current malaria control strategies

1.3.1 Vector control

Vector control strategies targeting the *Anopheles* mosquito include indoor residual spray (IRS) and insecticide-treated nets (ITN), as well as the improving environmental or urban/rural water irrigation as a form of larva management [13]. Both IRS and ITN are cost-effective and ITN, in particular, has decreased the incidence of malaria transmission by 50% and reduced child mortality by 55% in sub-Saharan Africa [29]. Unfortunately, resistance emergence within the vector towards the insecticides used in ITN and IRS as well as feeding behavior adaptations, threaten the progress made in preventing infection [30, 31]. This has spurred the search for novel insecticides and the development of new innovative control strategies to combat the change in mosquito feeding behavior [32].

1.3.2 Vaccines

One breakthrough for malaria control strategies is the development of a malaria vaccine(s), that could prevent the development of the disease within vaccinated individuals and thus limit the spread of the disease. The most successful candidate was the RTS, S/AS01

Malaria Vaccine which is a pre-blood stage vaccine [33]. Clinical trials phase III results, however, showed that the vaccine had very little efficacy and may not yet on its' own be able to be used as a malaria control for the eradication of the disease [34, 35].

1.3.3 Overview of current antimalarial chemotherapeutics

Currently, the treatment of malaria is limited to the use of antimalarial drugs and continues to be the primary artillery against the parasite while other preventative measures such as an effective vaccine are optimized. However, the majority of the historic or currently used antimalarials such as antifolates, quinolones, artemisinins [36, 37] are threatened by resistance development and their use is currently limited (Table 1). From Table 1, it is apparent that most of the current antimalarials' MoA act on the heme metabolism of the parasite, despite our current antimalarials having a chemical space that shows that compounds separate into several dominant structural groups [38, 39].

Table 1: Antimalarial mode of action and parasite's resistance mode of action

Class	Antimalarial	Mode of action	Mode of resistance
Artemisinin	Artemisinin and derivatives	Production of toxic heme-adducts	Kelch13 mutation [40]
Antimicrobial	Tetracycline	Inhibition of protein synthesis, but antiplasmodial action not clear [41]	Not known, but there is a number of hypothesizes [41]
Antifolate	Pyrimethamine	Inhibit plasmodial dihydrofolate-reductase [42, 43]	Mutation in dihydrofolate reductase binding site [42, 43]
	Sulfadoxine–pyrimethamine	Inhibition of plasmodial dihydropteroate synthase [44]	Mutation in dihydropteroate synthase gene [44, 45]
Quinoline derivatives	Halofantrine	Inhibit downstream growth of parasite [46]	Not clear [47]
	Atovaquone	Inhibits mitochondrial electron transport in the cytochrome bc complex [48]	Nucleotide polymorphisms in cytochrome B gene [48]
	Mefloquine	Cytosolic mode of action and production of toxic heme adducts [49]	Amplification of <i>pfmdr 1</i> gene that accumulates drug in digestive vacuole away from the cytosol site of action [49]
	Quinine	Production of toxic heme adducts	Amplification of <i>pfmdr 1</i> gene; production of an efflux transporter [47]
	Chloroquine (CQ)	Production of toxic heme adducts	Use of the PfCRT protein, a chloroquine efflux transporter [50, 51]
	Primaquine	Production of reactive oxygen species [52]	Not known

Antimalarial resistance development has prompted the use of artemisinin-based combination therapies (ACT) as a first-line treatment. With ACTs, antimalarials with a different mode of action (MoA) are used in combination to lower the rate of resistance emergence and spread [53]. This is because it is less probable to develop two different resistance mechanisms rather than one where the mutation or fitness cost won't be lethal to the parasite [54]. Unfortunately, even these precautions have not prevented the development of delayed parasite clearance within infected individuals treated with ACTs and could be an indicator of resistance formation against ACTs [55]. The current threat of resistance even against ACTs has resulted in the continued search for new chemical molecules with novel MoA that can be developed into new antimalarials, as discussed below.

1.4 Antimalarial drug discovery in the era of elimination

The discovery of new antimalarials is guided by target candidate profiles (TCP), which describe the ideal type of antiplasmodial molecules needed (Table 2), as identified by the Medicines for Malaria Venture [56]. These TCPs can be combined into different target product profiles (TPP) to serve as combination therapies for chemoprotection and case management that will aid in elimination [57]. TPP-1 for example, which is designed for case management, requires a combination of TCP-1 molecules, with the incorporation of TCP-5 molecules to reduce transmission and TCP-3 molecules to prevent relapse [57].

Table 2: Properties of current target candidate profiles

Type of TCP	Activity or intended effect
TCP-1	Molecules that clear asexual blood-stage parasites
TCP-3	Molecules active against <i>P. vivax</i> hypnozoites
TCP-4	Molecules active against hepatic schizonts
TCP-5	Molecules targeting the transmission of sexual parasite stages
TCP-6	Molecules targeting insect vector

The one essential factor for all new compounds irrespective of TPP and TCP targeted is that the compound has to present a novel MoA [58], such that these compounds do not show cross-resistance against currently circulating parasites.

Additionally, exposing parasites to new compounds with novel MoA should delay resistance development. Lastly knowing the MoA of a candidate can provide useful information in assessing cross-resistance of a combination as well as help understand and predict the combined effect of candidates [59]. With a clear goal set of the type of candidate molecules needed in antimalarial drug discovery, the search and identification of such

potential candidates need to be quickly identified during the drug discovery process to help accelerate drug development.

1.4.1 Limitations in antimalarial drug discovery

Although there are two routes of drug discovery, namely target-based and phenotypic screening, target-based drug discovery is confounded in malaria research since more than half of *P. falciparum* genes lack functional annotation and few *Plasmodium* proteins have known 3D structures [60-62]. By contrast, phenotypic screening has been highly successful and delivered thousands of hit compounds with good cell permeability and nanomolar whole-cell activity against multiple life cycle stages of the parasite [63-66].

With phenotypic screening, a large diverse chemical library is screened against sexual and asexual stages of *P. falciparum* and compounds with activity in the nanomolar range are identified as hits (Figure 2). Other assays are conducted to infer the compound's selectivity for the parasite and the parasite stage which the compound acts on to aid in identifying promising compounds that belong to the desired TCPs (Table 2). These promising compounds undergo hit-to-lead (H2L) optimisation that relies on establishing a structure-activity relationship (SAR) and subsequent medicinal chemistry guided changes to produce potent lead druggable compounds (Figure 2). Resultant lead compounds are modified during lead optimisation (LO) to further increase their pharmacodynamics and -kinetic properties. From the final leads the most promising candidates are selected to undergo pre-clinical studies such as combinational studies and *in vivo* humanized mouse model studies that assess blood-stage antiplasmodial compounds.

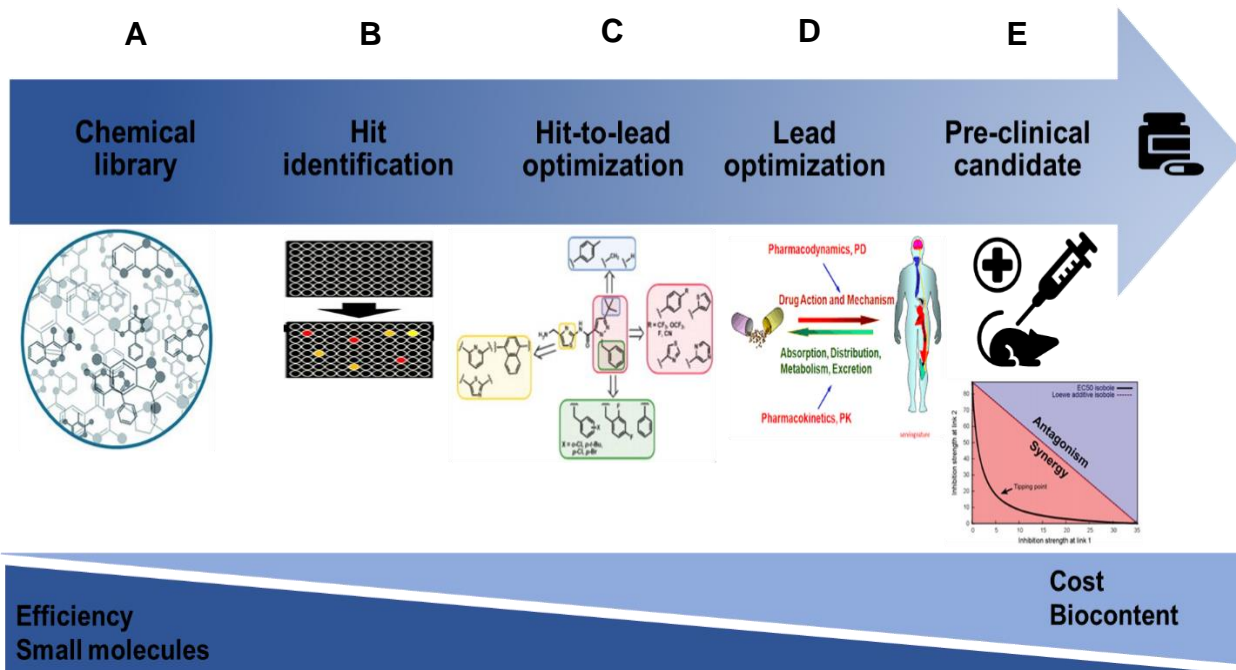


Figure 2: Typical drug discovery pipeline with high-throughput phenotypic screening.

(A) Thousands of chemically diverse compounds are tested for activity against the sexual and/or asexual stages of *P. falciparum*. (B) Hit compounds are identified which show *in vitro* nanomolar activity against the parasite. After hit identification, the efficiency and the amount of compounds within the pipeline decreases substantially. During this phase, other costly biological assays are required to infer the stage activity and selectivity of the hit compounds towards the parasite to increase the biological knowledge (bio-content) of the hit compounds and aid in identifying promising hit compounds for hit-to-lead optimisation. (C) During hit-to-lead optimisation, SAR and medicinal chemistry are used to identify chemical modifications of promising hit compounds that increase its' potency/activity against the parasite. During this phase, there is a continuous cycle of modifying hit compounds and assessing the *in vitro* activity of chemically modified derivatives from the promising hit compound until the best derivatives are selected as lead compounds. (D) Resultant leads of hit-to-lead optimisation are further modified to increase pharmacodynamics and -kinetic properties of the lead compounds and undergo time-consuming MoA studies to aid in selecting leads for pre-clinical candidates. (E) These candidates undergo pre-clinical studies that further develop the bio-content of candidates for clinical trials such as combinational studies and *in vivo* humanized mouse model studies.

Phenotypic screening has the advantage of identifying multiple hits without any knowledge on their MoA or target. Therefore, in the antimalarial drug discovery pipeline, MoA is typically determined for compounds during the LO phase to ensure that a compound's drug target or MoA is known before it is evaluated as a preclinical candidate. This is to assist downstream combination studies as well as to prevent further investment in leads that have a low probability of succeeding in clinical trials.

In the antimalarial drug discovery pipeline (Figure 2), the rate-limiting steps that decrease the efficiency of compounds progressing through the pipeline occur within the H2L and LO phase of drug discovery. During H2L optimisation, the medicinal chemist repeatedly modifies hits to increase their potency and use SAR to guide them in their chemical modifications. However, this process is fraught with the possibility that changes in potency resulting from chemical modifications are due to stronger binding of the compound to its' target or alternatively to a change in MoA. Not only this, but compounds with undesired

MoA may only be discovered during the LO phase whereby the cost invested in such leads is already wasted.

To address these challenges and help accelerate antimalarial drug discovery, a MoA identification method is required which is cheap and can be adapted for high-throughput applications to be used during the H2L phase. Such a method will help medicinal chemists in monitoring change in MoA during the derivatization of hits and define the chemical space of hits during H2L optimisation. It will also help reduce costs by eliminating the investment in compounds with undesired MoA.

1.5 Target and MoA identification strategies in drug discovery

The Malaria Drug Accelerator Consortium (MaDA) is a collection of laboratories in academia and industry that are at the forefront of antimalarial drug discovery that aims to identify novel assayable drug targets and the MoA of promising leads within the antimalarial drug discovery pipeline [67]. To help speed up drug discovery of lead compounds they have employed and developed multiple methods for target and MoA identification, some of which will be discussed below.

To identify the targets of a compound, multiple direct and indirect methods for target identification had been developed over the years. Protein-based pull-down assays such as affinity chromatography and drug affinity responsive target stability (DARTS) are an example of such direct methods which rely on the binding affinity between a compound and its' targets. The limitations of DARTS and affinity chromatography is that low-affinity targets or proteins in low abundance are not detected [68, 69]. This is particularly true for target identification studies in antimalarial drug discovery where protein abundance within *P. falciparum* varies during the parasite's development [70, 71]. Alternatively, indirect forward genomic methods have also been developed to help identify drug targets. This typically includes exposing parasites to sub-lethal concentrations of a compound until the parasite develops resistance. Whole-genome sequencing of the resistant mutants is subsequently used to identify gene mutations that led to resistance development towards the compound, which is typically found within the drug target itself [69, 72]. One major limitation is that not every resistance mutation relates to the target [69, 73] but can simply be an indication of the resistance mechanism itself. An example of this is the chloroquine and the chloroquine resistance transporter that helps in exporting the drug out of the vacuole and is not the target of chloroquine action [74].

The above target identification strategies are typically not high throughput. Other target-based investigative tools (e.g. gene function manipulation through CRISPR Cas for instance) are not scalable and will not provide a full descriptor of the global effect of a compound on an organism, including identifying 'off-target' effects [75]. Additionally, these direct and indirect approaches for target identification has only been applied for the asexual proliferative stages of the parasite, whereas forward genetics resistance induction would not be relevant for other non-proliferative stages like gametocytes.

In instances where target identification is not possible, a compound's MoA may still be informative to allow progression through the drug discovery pipeline. Additionally, this allows evaluation of pluripharacology where compounds have complex MoA and the combined effect of a compound acting on multiple targets play an important part in the MoA of a compound that leads to the desired phenotype [75, 76]. Thus, MoA studies of a compound aim to highlight the key biochemical pathways affected due to compound exposure [77]. MoA provides information on the overall cell processes effected by the compound which leads to the observed phenotype, but the actual drug targets are not necessarily identified [78].

Most MoA characterisation methods rely on -omics approaches to obtain a global perspective of a compound action, whereby the MoA footprint of a compound can be determined to function as a descriptor of the compound's induced phenotype. Metabolomics, for example, indirectly allows one to infer a compound's MoA based on the principle that if a compound inhibits a metabolic enzyme, the respective substrates would accumulate, whereas the opposite is true of the products [69]. Global unbiased approaches allow the entire metabolome to be compared between treated and control samples, and this is typically performed using mass spectrometry or nuclear magnetic resonance [69, 79]. Deviations from the controls in certain substrates or products can help in identifying metabolic pathways that were affected by the treatment. This method is very sensitive and has already been used to identify the MoA of 40 antiplasmodial compounds with the highest priority in 400 compounds from the Open Access MMV Malaria Box which previously had an unknown MoA against *P. falciparum* [80, 81]. Recently, the targeted pathways of 110 out of the 169 compounds within the MMV Box were predicted using metabolomic profiling [82]. There are, however, limitations to using metabolomics for MoA studies early in the drug discovery pipeline, as some compound's whose MoA is unrelated to metabolism may be more difficult to ascertain or detect [69, 80]. Moreover, only a handful of metabolites can be quantitatively detected and may only be active during a specific stage within the parasite life cycle. This also makes between stage comparison difficult as some stages of the

parasite are more metabolically active than others [83]. Lastly, metabolomics is extremely time and resource-intensive, which precludes its routine use in guiding the profiling of compounds through the drug discovery pipeline.

Alternatively, the use of transcriptomic data has been successfully used in studies of cancer in elucidating the MoA of compounds and some studies have even been able to refine the use of transcriptomic data to identify a list of possible targets the compounds may act on. The rationale is based on the interaction between the compound and its' target(s), by binding to the target, the activity or function of the target is affected. Ultimately, homeostasis and normal cellular functions are affected which influence the expression of genes involved with these cellular processes and homeostatic control. As a consequence, the transcriptome becomes perturbed and cause particular genes to be differentially expressed (DE) [84]. These DE genes may function to help elevate the effect of the compound or be a direct effect of the compound on a cellular pathway. Thus the drug perturbed transcriptome can be exploited to aid in elucidating the MoA of a compound [85].

1.7 The use of transcriptome datasets to identify a drug's MoA and targets

Although transcriptomics allows for an intensive global overview of a drug-treated cellular state, it also produces gene expression profiles (GEPs) with high dimensionality as a result of over-viewing a large set of genes. This high dimensionality datasets can complicate the determination of patterns for MoA fingerprinting or target identification. To overcome this, many methods have been developed to allow one to extract the relevant data from transcriptional responses, ranging from hierarchical clustering and gene expression correlations to more advanced methods for inferring MoA and targets using network analysis and machine learning (ML).

1.7.1 Gene expression correlations

Gene expression patterns can be inferred by using correlations tools such as Gene Set Enrichment Analysis (GSEA) to highlight corresponding gene sets that are significantly different between two different biological conditions [86]. Although this tool is usually used to identify differentially expressed genes (DEGs) between control and treatment samples, this method can also be applied to different compound treatments to highlight gene sets that differ in their gene expression signatures between different compounds.

Gene expression profiles of cancer cell lines before and after treatment with different compounds have been used to identify shared pathways affected by different compounds. Using this, compounds that affect similar pathways could be classed together as it is likely

that such compounds share similar MoA. This was accomplished by developing a computational method called Drug-set Enrichment Analysis (DSEA), that highlights significantly modulated pathways within a drug-set, relative to other drugs within the database. In this way, transcriptional responses that may not relate to the compound induced phenotype are 'normalized' so that only pathways that are significantly affected are analysed. The significance by which a pathway is affected by a compound is then assessed using a similar approach to GSEA, the only difference being that an enrichment score is computed for drug ranks to a set of genes in a pathway and not gene ranks. Drugs ranks are drugs distributed within a row according to how they affect a pathway, with each row representing a different pathway. In this regard, a pathway is defined as a set of genes that are functionally related. The drug ranks are compared on how they affect pathways relative to other drugs and an enrichment score generated [87]. With DSEA, the analysis becomes more powerful and robust as the number of drugs included in the analysis database is increased.

DSEA might be more relevant to MoA prediction of a compound within *P. falciparum* than identifying drug targets since DSEA focuses on highlighting the pathways affected that results in a certain phenotype. However, one limitation of using this within *P. falciparum* is that the majority of the parasite's genome is not functionally annotated and this, in turn, limits the number of known pathways within the parasite to investigate.

1.7.2 Hierarchical clustering

Clustering can be used to arbitrarily group compounds together based on similar gene expression patterns in their GEPs. This is very useful in identifying genes that can function as a MoA fingerprint or descriptor. For example, within the TB research field, Murima *et al.* were able to develop a microfluidic medium high-throughput format for transcriptional profiling of compounds that used only 90 biomarker genes to effectively stratify and deconvolute a compound's MoA [88]. This was done through hierarchically clustering the GEPs of *Mycobacterium tuberculosis* treated independently with a variety of compounds. The GEPs of compounds with very similar transcriptional responses clustered together, as it is likely that these compounds target the same pathways which result in similar downstream transcriptional responses that led to the phenotype observed. From the hierarchical clusters, they identified biomarker genes that represented these clusters which were able to aid in MoA stratification of compounds active against *M. tuberculosis* [88].

One limitation using hierarchical clustering is that some DEGs may become irrelevant during clustering and once a cluster has been merged it is impossible to retrace the steps

made during clustering [89]. Although unsupervised clustering may be adequate for identifying biomarkers for a MoA fingerprinting, clustering analysis alone has shown to produce poor results when trying to predict a condition or class from GEPs [90]. Thus, clustering is inadequate for solving classification problems, such as identifying compounds with similar MoA.

1.7.3 Network analysis for MoA fingerprinting and target identification

Network analysis allows a more contextual overview of gene expression patterns. Networks can either be directed (e.g. gene regulatory networks or GRNs) or undirected networks (e.g. co-expression networks), where the direction of interaction between two nodes is either known or not known, respectively. This is not only limited to genes but can also be applied to metabolomics and proteomics and can even allow the integration of these different approaches (Figure 3). In the case of GRNs not only are the DEGs analysed in the network but also their effect on other genes that they are connected to in the network. Network analysis can thus be a powerful tool to analyse the influence of DEGs on other genes and highlight affected pathways due to treatment with a compound, thereby helping in MoA and target identification.

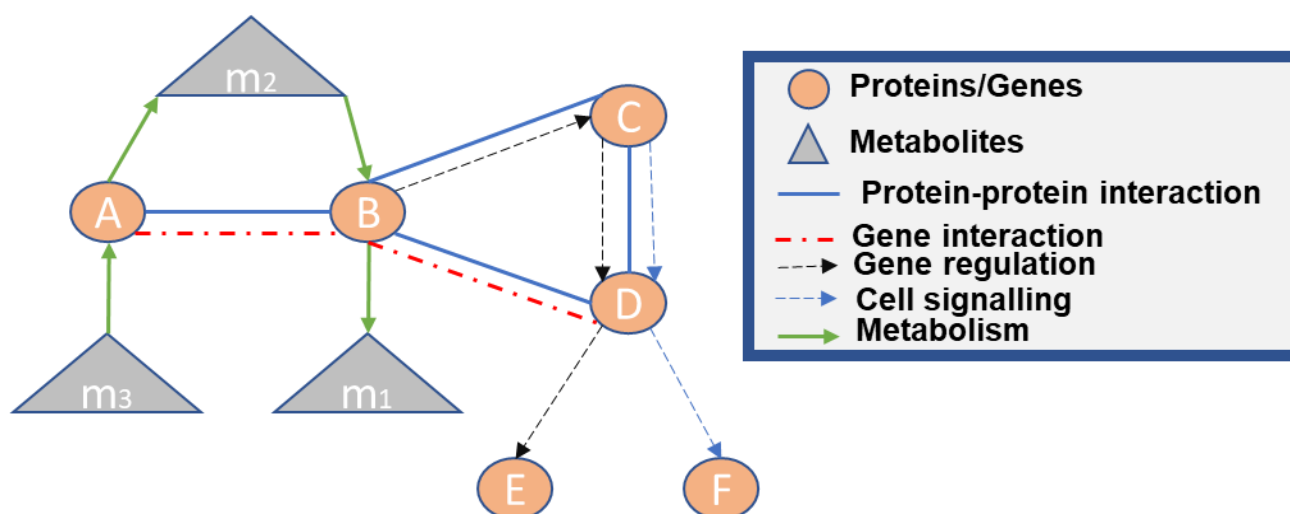


Figure 3: Simplified example of a network.

This is a combined directed and undirected network, where lines (edges) with arrowheads indicate the direction of interaction between the different biological entities (nodes) such as metabolites (triangles) and proteins or genes (circles).

As an example of a GRN, in cancer research, Woo *et al.* applied network analysis to clarify the MoA of compounds using transcriptome data as a result of global gene dysregulation due to treatment. The resultant DeMAND algorithm evaluates dysregulation between each

interacting gene pair within the GRN and produces a probability density before and after treatment with a compound. Thus, all perturbations are considered in the transcriptome and not only DEGs. The algorithm then gives an output in the form of a ranked list containing all the network genes and the statistical significance of their dysregulation. DeMAND successfully predicted the possible target proteins of ~70% of compounds with a ~20% false discovery rate (FDR) and these targets can be further investigated to identify the MoA of a compound [91]. There are, however, some limitations to this DeMAND algorithm as it may not be fully utilized or employed within antimalarial drug discovery. This is because currently there are only limited GRNs of *P. falciparum* available and most of these do not cover more than half of the parasite's genome. The DeMAND algorithm is dependent on high-quality context-specific GRNs and although the algorithm can still use non-context-specific networks for analysis this will result in a higher FDR and incorrect predictions. Since there isn't much known of the global GRN during the parasite's asexual or even sexual development, an approach similar to the DeMAND algorithm may not yet be applicable to *P. falciparum*.

As an alternative to GRNs, protein-protein interaction networks (PPINs) can be used when evaluating gene expression perturbations to uncover a compound's targets and the pathways influenced. A study within the cancer field using PPINs was able to predict 22% of known targets within the 1st percentile. By using direct contacts, genetic interactions, and functional relationships, PPINs were used to provide downstream relationships between drug targets and other proteins. Pathways affected by the compound was highlighted by using a network topological distance measure to extract the shortest path between perturbed genes and known targets, as this may explain the phenotype that resulted from drug treatment. In the case of *P. falciparum*, the PPINs available may not be suitable for MoA identification, as most information about the parasite's proteome is less detailed than that of the parasite's transcriptome [92].

Although network analysis may be an attractive approach, due to the complexity of the parasite and that more than 59% of coding genes are differentially expressed when treated with compounds, this can lead to generating a very interconnected network [93]. Such networks are usually difficult to unravel in order to obtain useful information, especially when the majority of the parasite's genome lacks functional annotation. One alternative suggestion is that ML algorithms can be a useful alternative to deconvolute these complex transcriptional drug signatures to identify pathways on which a compound act. One benefit these algorithms have over network analysis methods is there is no prerequisite to define in detail the structure of interaction between genes [94].

1.7.4 Machine learning for MoA fingerprinting and classification

ML is a computational method that uses statistics and mathematics algorithms on large datasets to determine imperceptible patterns in the said dataset. Based on the algorithms' training on the dataset, it has the ability to make dependable statistical predictions about similar data [95]. There are two types of ML approaches, namely, supervised and unsupervised learning (Figure 4). Supervised learning is a type of ML where the algorithm is trained with a labelled training dataset whereby the algorithm then can make predictions on an unlabelled, unevaluated dataset [96]. With supervised learning, the category/classes or properties in the training set need to be predefined/labelled [97].

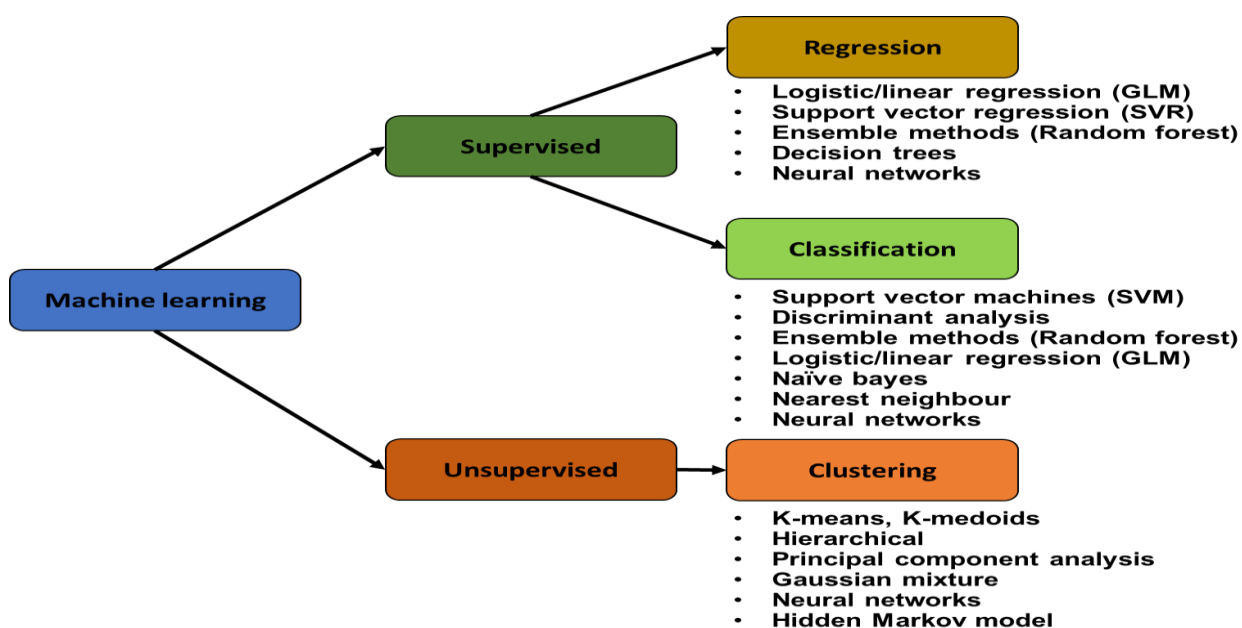


Figure 4: Supervised and Unsupervised machine learning algorithms.

Supervised learning can be grouped into two sections, classification and regression algorithms such as support vector machines and linear regression. Unsupervised learning primarily consists of clustering algorithms such as k-means and hierarchical clustering.

In unsupervised ML methods, patterns are found in a dataset without the dataset needing to be labelled and is an alternative ML method when a labelled training set isn't available [96]. Unsupervised ML methods, however, perform poorly with classification problems and mainly focus on defining or understanding the relationship between data points and makes no assumptions about the data structure. As a result, unsupervised learning algorithms may identify spectrally similar classes within the data that may be due to the asexual stages of the parasite or other factors rather than the MoA of a compound [98]. This, in turn, can result in arbitrary MoA classification when using unsupervised learning. In this regard, supervised classification algorithms are more suitable for MoA classification.

Due to the promising results of ML in complex biological conditions, some pharmaceutical companies are employing ML to help drive drug discovery in a cheaper, efficient and more effective way [99, 100]. Similarly, some cancer studies employed the use of ML to predict drug response using drug-induced gene expression data and/or single-cell phenotype images [101, 102]. In fact, ML has already been used in a few studies in the malaria field to help identify synergistic compound combinations [103]. However, the use of ML in malaria research is relatively new and has not yet been utilized fully in the extent of antimalarial drug discovery compared to cancer studies.

To accelerate antimalarial drug discovery, a ML model that can stratify the MoA of compounds based on transcriptional responses will be extremely useful if it can be employed during H2L optimisation. It will be very advantageous as this predictive stratification model generated for *P. falciparum* will be able to identify gene expression patterns shared between compounds with similar MoA and stratify such compounds together. Not only will such a model identify compounds with novel MoA early in preclinical development, but it will help monitor any change in MoA of a compound during chemical modification in H2L and LO optimisation. Such a model will be beneficial in guiding medicinal chemists on how the chemical modifications change the MoA of a compound as well as aid in defining the chemical space of such compounds.

1.8 Rationale using *P. falciparum* transcriptome for MoA deconvolution

Prior evidence exists as proof of principle that *P. falciparum* transcriptome data can be used to obtain some differential transcriptional responses to particular compound classes [104]. In this study by Siwo *et al.*, they analysed the gene expression of *P. falciparum* towards 31 chemically and functionally different compounds that targeted different pathways within the parasite. For each compound, they generated a genome-wide response index by calculating the response index for each gene after treatment with a compound. To compare the global transcriptional response of each compound they correlated the genome-wide response index of different compounds and employed hierarchically clustering. From clustering, they found that similar transcriptional responses were shared between different compounds some of which have similar chemical features (Figure 5). However, not all compounds with the same transcriptional responses necessarily shared the same chemical features and were more likely to affect the same pathway within a cell. With PCA they determined that chemical similarity and the MoA of a compound play a role in evoking similar transcriptional responses [104].

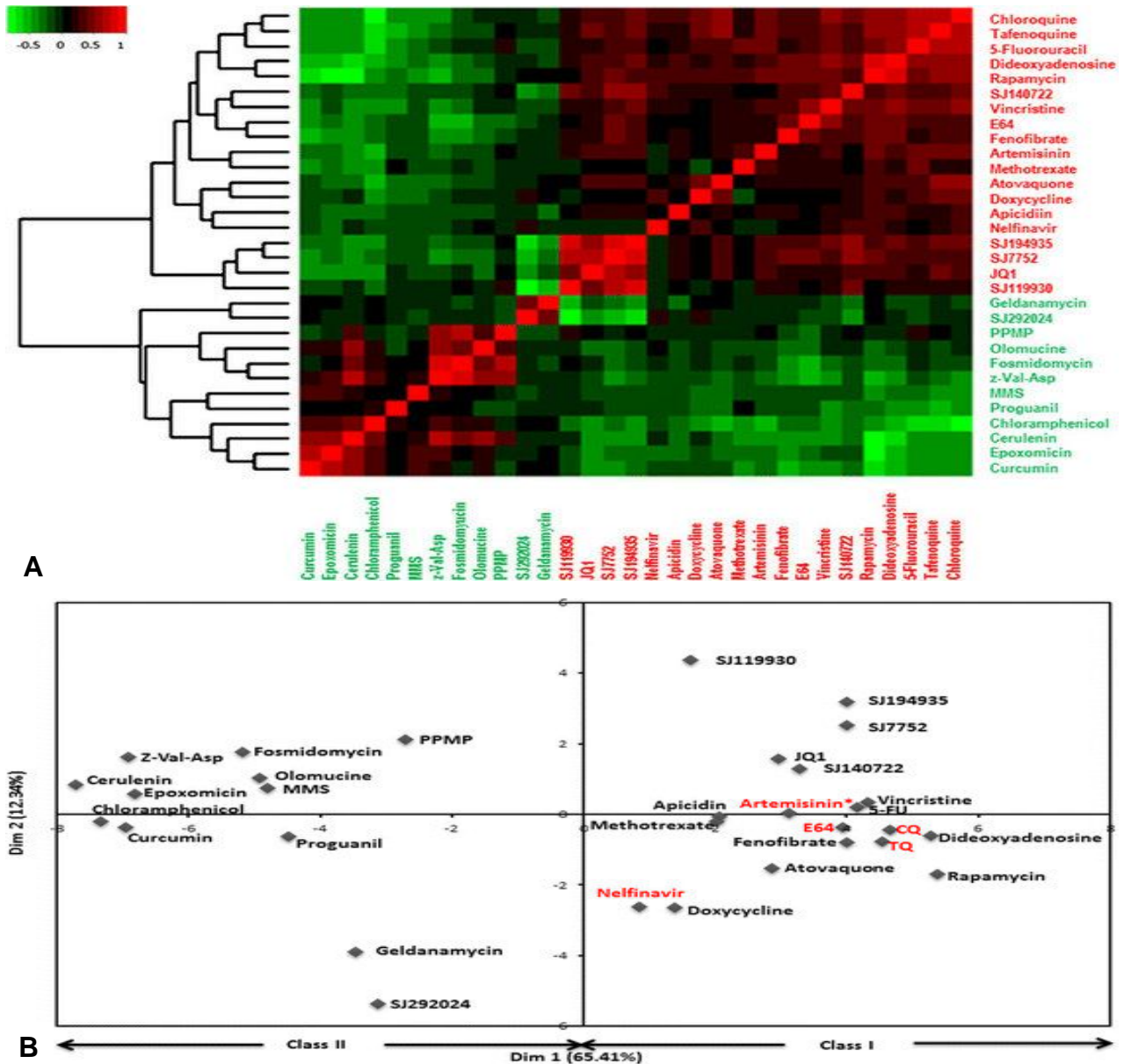


Figure 5: Gene expression responses to drug treatment related to a compound's MoA and/or chemical structures.

A genome-wide response indexes was calculated for each compound and used for (A) hierarchical clustering. Two classes were observed (red and green). Within both classes, compounds were positively correlated to compounds within its class and negatively correlated to compounds in the other class. (B) From PCA it is observed that both chemical structural similarity (in red CQ and TQ compounds) and MoA (in red E64, Artemisinin, CQ and TQ) are captured within the drug perturbed gene expression of the parasite. Source:[104]

One critique of the Siwo *et al.* study, is that they used GEPs of different *P. falciparum* cell lines to try to enhance the signal-to-noise ratio. These cell lines differ in compound sensitivity to certain antimalarials and may cause increased variation between different cell line transcriptomes. Therefore, using different cell lines for averaging may cause skewed results.

Hu *et al.* also used transcriptional responses of the *P. falciparum* to integrate co-expression of genes together with domain-domain, sequence homology and yeast two-hybrid data in order to construct an interaction network able to predict protein function [93]. The authors

treated parasites with 20 different compounds and found that gene expression in response to chemical stimuli was highly reproducible and dose-dependent. They also found that when considering both proteome coverage and positive prediction rate, transcriptomic data outperformed all of the three different protein-protein interaction datasets [93]. Although the authors were able to identify DEGs for multiple treatments and could observe notable expression patterns differences between these treatments, they did not further investigate these genes for MoA stratification.

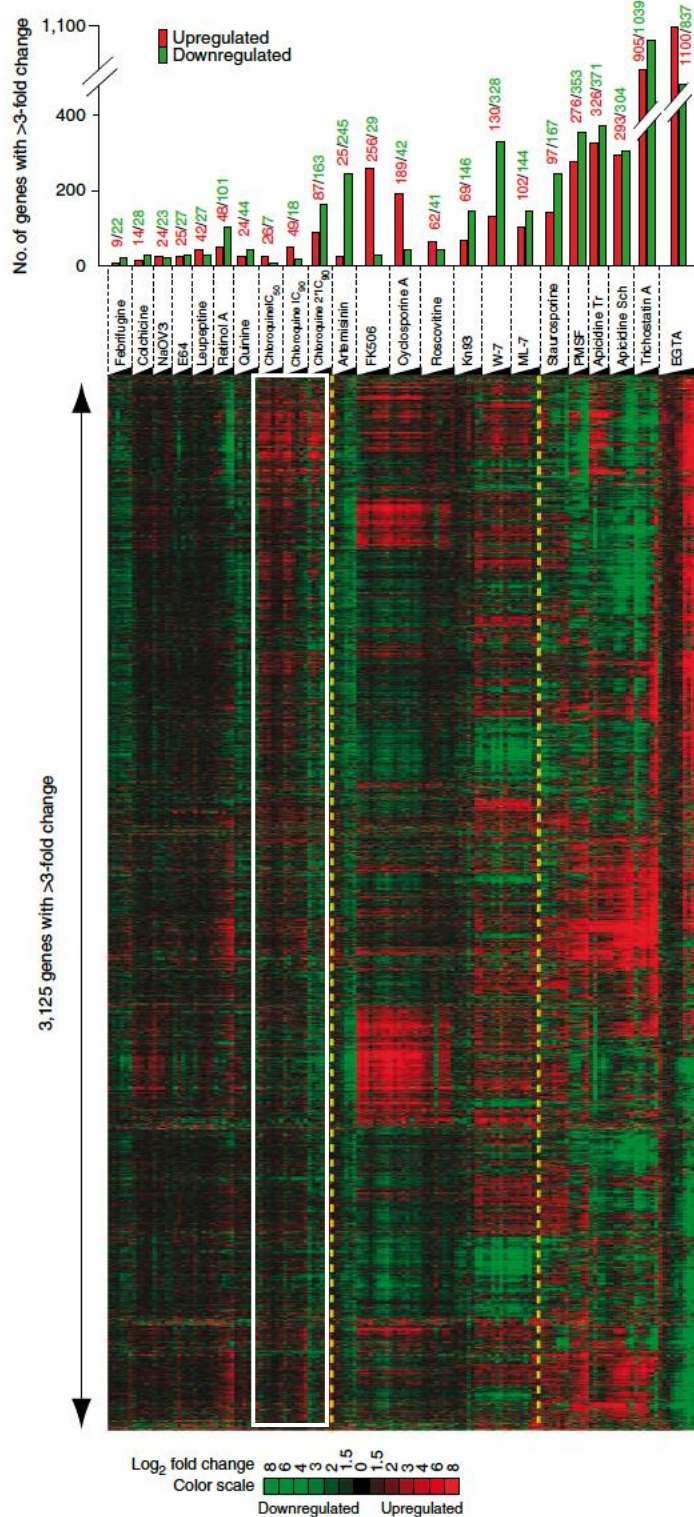


Figure 6: Discernible expression patterns across treatments.

Across different treatments, strong distinguishable expression responses are observed. With a 3-fold change in expression criteria for DEGs, a majority (3125) of the coding genes show DE in at least one treatment. From chloroquine treatments (white box) it is shown that these responses are reproducible and dose-dependent. Source: [93]

From the above two studies, it is evident that the parasite has unique gene expression patterns to such an extent that compounds having different MoA or chemical features show different transcriptomic responses. Additionally, these responses are reproducible and

specific to compounds with similar MoA. This implies that the parasite's transcriptome can be mined to determine MoA or stratify compounds to their MoA in pre-clinical studies but doing so on a genome-wide scale may not be feasible in early drug discovery. The current challenge is to devise a tool that can be used in a drug discovery program to quickly (medium- to high-throughput level), efficiently and economically stratify a compound's potential MoA, without the need to complete full transcriptome analysis of every compound. We hypothesise that a small subset of DEGs could be used as representative biomarkers of compound MoA to be used to evaluate and stratify compounds at scale.

1.9 Study aim

This project aimed to develop a ML model that can stratify compounds with similar MoA together based on transcriptional responses of a limited number of biomarker genes. This model will enable medium through-put MoA elucidation and be useful in guiding preclinical decisions to fast-track drug development during H2L optimisation as well as the identification of compounds with novel MoA.

We rationalise that these biomarker genes can be identified using a rational feature selection approach that can select genes representative of drug MoA. We hypothesise that we can use this rationale feature selection approach to identify biomarker genes that are unique for a MoA and pervasive throughout a compound's treatment and that these biomarkers can be used as predictive features to train a robust MoA stratification model.

Aim

To use transcriptomic data of drug-treated *P. falciparum* parasites to identify drug-specific biomarkers that with the help of machine learning can generate a predictive model to stratify antiplasmodial compounds with a similar mode of action together.

Hypothesis

The predictive model will be able to robustly group antiplasmodial compounds with a similar mode of action together.

Objectives

- Identify possible biomarker genes that represent a MoA

- Evaluate different classification algorithms for their ability to robustly stratifying compounds with similar MoA from transcriptomic data
- Generate a predictive model from biomarker genes
- Determine the optimal minimal number of biomarkers needed to generate a robust predictive model

Chapter 2: Methodology

To build a MoA stratification model that utilizes transcriptional responses, GEPs of drug-treated *P. falciparum* parasites were needed from which the ML models could be trained and tested on as well as to help identify biomarker genes representative of compound MoA.

2.1 Identifying predictive biomarker genes for mode of action stratification

A database was generated consisting of GEPs of *P. falciparum* parasites treated with different compounds. These GEPs were obtained from different open sources such as the NCBI GEO database and the detailed information on all of these datasets is presented in Table 3.

Table 3: *P. falciparum* compound treated GEP datasets used in this study

Ref	Compound(s)	Strain	Time points	Stage ^a	Controls	[Drug]	Replicates	GEO no	Date accessed
[105]	Thiostrepton~	3D7	24 h	R	DMSO	IC ₅₀	3	GSE28701	2019/02
Own	MMV390048 or MMV642943	3D7	24 or 48 h	R	UT	10 x IC ₅₀	1	GSE100692	2019/02
[106]	Cisplatin, Etoposide, Methyl methanesulphonate (MMS), Pyrimethamine ~	3D7	6 h	R	Reference pool	IC ₅₀ and IC ₉₀	3	GSE72580	2019/02
[107]	dihydroartemisinin (DHA)*	K1	1-3 h	T	Not clear	IC ₅₀	5	GSE62136	2019/03
[108]	Choline kinase inhibitor, hexadecyltrimethylammonium bromide*~	K1	72 h	R	UT	IC ₅₀	3	GSE54775	2019/03
[109]	Dcompound1, novel dihydroorotate dehydrogenase (DHODH) inhibitor*	Dd2	Not clear	Not clear	Dd2	IC ₅₀	3	GSE35732 GSE37306	2019/03
[110]	ACT-213615	3D7	1, 2, 4, 6, and 8 h	T	DMSO	IC ₅₀	Not clear	GSE39485	2019/03
[111]	Trichostatin A (TSA), suberoylanilide hydroxamic acid (SAHA) and 2-aminosuberic acid derivative (2-ASA-9)~	3D7	2 h	T	DMSO	IC ₉₀	2	GSE25642	2019/02
[112]	Pyronaridine, CQ*	K1	4 h and 24 h	T	UT	IC ₅₀	3	GSE31109 GSE30867 GSE30869	2019/02
[113]	Cyclohexylamine	3D7	18, 25 and 30 hpi	Both	UT	IC ₉₉	2	GSE18075	2019/02
[63]	DL- α -difluoromethylornithine (DFMO)	3D7	19, 27 and 34 hpi	Both	UT	5x IC ₅₀	2	GSE13578	2019/03
[114]	Dehydrobrachylaenolide	3D7	2, 6, and 12 h	Both	DMSO	IC ₉₉	2	GSE29874	2019/03

[93]	ML7, W7, KN7, Staurosporine, KN93, Cyclosporine A, FK506, Roscovitine A, Quinine, Chloroquine, Febrifugine, artemisinin, Na3VO4, Colchicine, Retinol A, PMSF, E64, Leupeptine, Apicidin, Trichostatin A, EGTA	3D7	1,2,4,6,8 and 10 h	Both	UT	IC ₅₀ and IC ₉₀	1 and other treatments had 2	GSE19468	2018/06
[115]	Ionomycin	3D7	30 min, 1, 2, 4 and 6 h	S	Reference pool	10x IC ₅₀	1	GSE33869	2019/02

A: R= rings, T = trophozoites; S = schizonts; UT= untreated parasites, hpi = hours post invasion, h= hours, min= minutes

2.1.1 Quality control filtering

Raw GEP data was used which contains a considerable amount of noise and technical variability that was introduced during labelling, hybridization, and scanning. Each GEP dataset was assessed individually for quality and usefulness using a filtering criterion set that was established as outlined in Table 3.

Table 4: Filtering criteria of GEP datasets

Criteria	Accepted	Rejected
Controls	Untreated parasites under same conditions as compound-treated parasites	Different conditions compared to compound-treated
Gene coverage	~75% coverage of <i>P. falciparum</i> genes	<60% coverage of <i>P. falciparum</i>
Mode of action	Known in <i>P. falciparum</i>	Unknown in <i>P. falciparum</i>
Time series	If there are ≥ 2 time points available for comparison to other compound treatments	Compound treatments that have no time points or replicates
Concentrations	IC ₅₀ and higher concentrations	Concentrations below IC ₅₀
Parasite strain	Treatments and controls need to be the same strain, preferably NF54 or 3D7	Resistant strains or clinical isolates will not be considered as transcriptional responses may vary due to strain differences and not compound treatments

To ensure that biomarker genes function as a representative of the compound's MoA, genes were extracted whose expression is perturbed when compared to their expression under normal circumstances. Thus, the GEPs needed to have suitable controls that reflect the state of the untreated parasite. These controls also needed to be from the same parasite population and strain as the compound-treated parasites and under the same conditions to ensure that the variances in gene expression were not due to environmental or population differences.

2.1.2 Merging and pre-processing (normalisation) of GEP datasets

Since normalisation strategies may differ between datasets, the raw GEP data was used and each dataset was assessed with regards to standard of quality for comparison. Since GEPs measure the expression level of genes using probes poor hybridizations, sample quality, and salt concentrations can lead to dry spots where probes for genes show too low or no signal [116]. As a result, low probe quality can cause genes to show no signal and thereby lower the gene coverage of the GEP of a dataset. Hence, GEPs were filtered for gene coverage based on probe quality and GEP datasets which had a gene coverage below 3180 genes (i.e. below 60%) were rejected. The accepted datasets were merged to form our GEP database, but since the GEPs are obtained from different authors and are performed on different GEP platforms (e.g. Affymetrix and Illumina), between-array normalisation was needed to allow for comparison between datasets.

Different normalisation methods were assessed for their ability to allow comparisons between different GEPs in our database. This was also an essential pre-processing step before the data can be used to build a model using machine learning algorithms, as some algorithms require the data to be normalized. Different microarray between-array normalisation strategies were assessed using the limma package [117].

For each compound and their respective controls within the database, boxplots were drawn before and after normalisation to aid in evaluating each normalisation strategy for the acquired data. From this, the best normalisation strategy was selected, and between-array normalisation was performed on our database.

2.2 Generating a predictive model

2.2.1 Employment of machine learning on GEPs

To build our antiparasitodal MoA stratification model, algorithms that can solve multiclass classification problems were identified (Section 2.2.2). These algorithms were then investigated for their ability to build a robust and stable MoA stratification model from transcriptomic data using the principle explained in Figure 7. For these classification algorithms, it was required that the input data used for training to be labelled. Thus, each input data, i.e. time points of a compound treatment, were labelled according to that compound's respective MoA. In other words, when training our model on the GEPs, the

gene expression within the treatment time point is the input data points, the genes are the features and the labels are the compound's MoA.

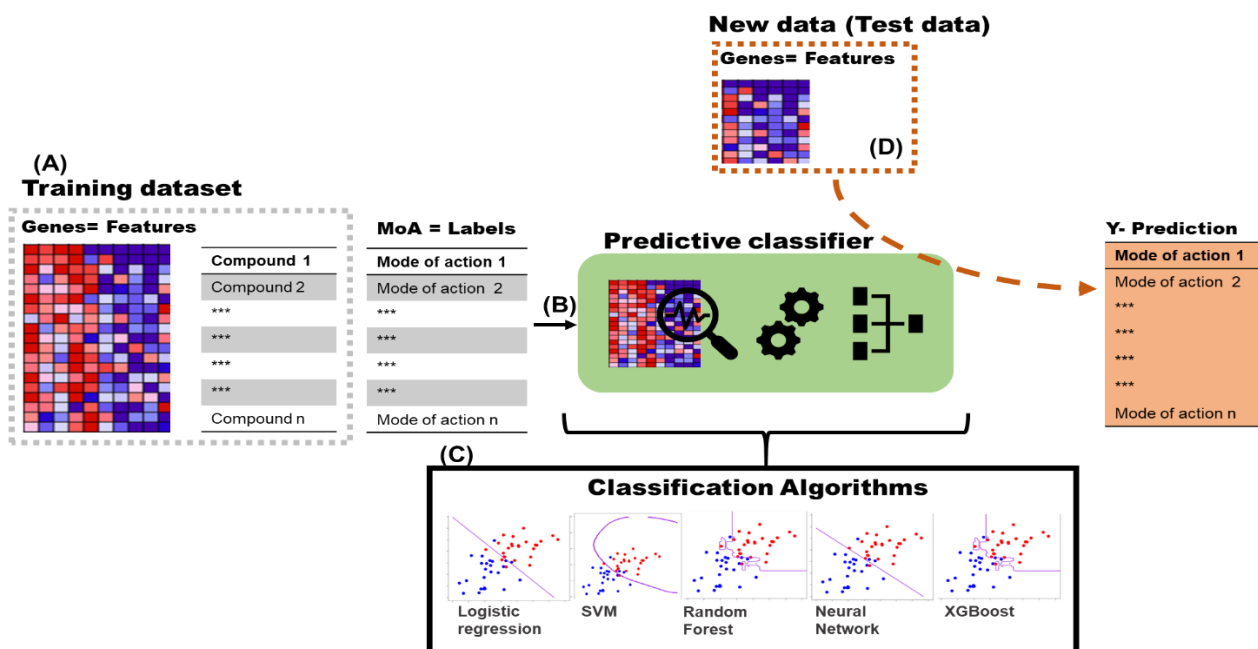


Figure 7: Principle of training a compound MoA stratification model using transcriptional responses of genes.

From our treated *P. falciparum* GEP database, a subset of the database will be used for training the algorithm and will thus be called a training dataset. (A) The training dataset will contain genes that function as the training features. To ensure a MoA stratification model is built, the compounds used in treatments have to be relabelled according to their MoA corresponding to literature. (B) The training dataset is then used to train a classifier i.e. model to recognize patterns in gene expression for a specific MoA. (C) Here different classification algorithms can be used to build a classifier. (D) The remaining subset of the database is then used to assess the accuracy of the classifier.

For downstream applications and analysis, each algorithm had to construct two ML models, one using all the genes within our GEP database and the other only using the biomarker genes identified from our rational feature selection approach (Figure 8).

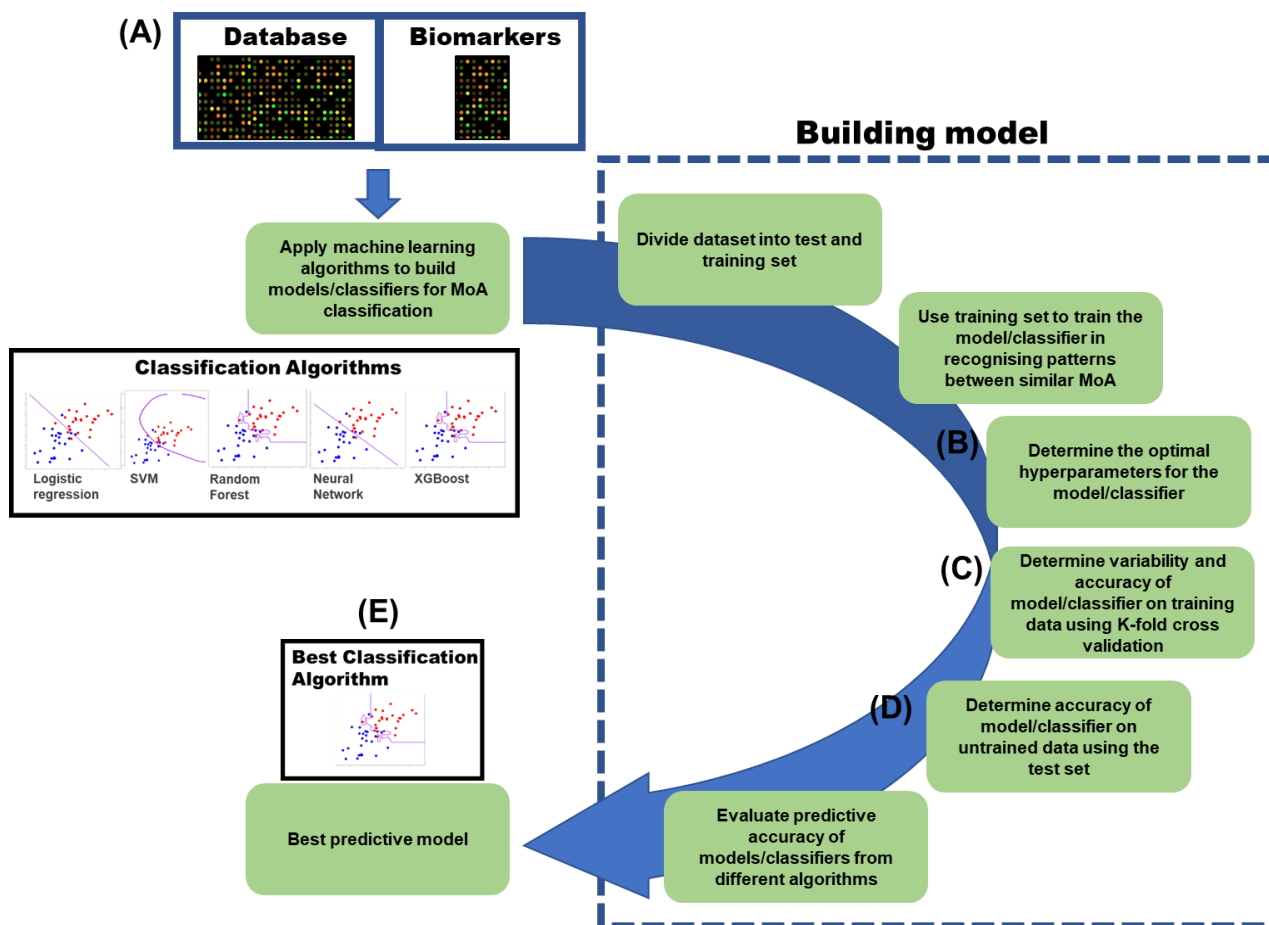


Figure 8: Method for selecting the best ML algorithm for MoA stratification.

(A) For each classification algorithm, two models were built, one using all the genes within our database as features, the other using the biomarker genes identified as features. During model building, the optimal hyperparameters are determined (B) for each model before 10-fold cross validation (C) was implemented on the final model using the optimal hyperparameters. (D) Each of the two final models from each algorithm was then evaluated for their MoA stratification performance on the test set. (E) The ML algorithm with the best accuracy and performance from (C) and (D) was then selected.

During the process of building these ML models, each was optimized and assessed for their MoA stratification performance (Figure 8) before the best algorithm was chosen that performs well on both the full set of genes in our database as well as our biomarkers.

2.2.2 Selection of ML algorithms to be investigated

Since the MoA of compound treatments is known within our database, we provided the labels to these treatments and used supervised algorithms to help build a model including multinomial logistic regression (MLR), support vector machines (SVM), random forest (RF), artificial neural networks (ANN) and gradient boosting machines (GBM) to address our MoA multiclassification problem (detailed in Appendix A). Before an algorithm's model can be evaluated on its' performance, the optimal architecture of the model was identified through

hyperparameter tuning (Appendix A). Hyperparameters dictate a ML model's architecture and the default hyperparameters used in an algorithm may not be the optimal architecture of the model to address our MoA classification problem [118]. Hyperparameters were fine-tuned individually for each algorithm as explained below and in Appendix A: Table A.1.

2.2.2.1 Multinomial logistic regression

Multinomial logistic regression (MLR) solves multiclassification problems by building generalized linear models that provide outputs in the form of the estimated probabilities of belonging to a category or class [119]. To build a MLR model for multiclassification the h2o R package was used by employing the `h2o.glm()` function, which is abbreviated for generalized linear model and can be used for both binary and multiclassification problems [120]. There are no hyperparameters for multinomial logistic regression and as such no hyperparameter tuning was required when building our MLR models [121].

2.2.2.2 Support vector classification

Support vector classification builds classifiers that utilize hyperplanes to separate and help distinguish members of different classes. In the case of multiclassification problems, multiple classifiers are built and the outcome of each serves as a vote on which class the data belong to and the majority vote of the multiple classifiers determines the class. To build such a multiclass SVM model the e1071 R package was used to train and fine-tune the hyperparameters for our SVM models [122]. With the e1071 R package the performance measure used to identify the optimal hyperparameters from the ranges shown in Appendix A: Table A.1 was classification error. During hyperparameter tuning, no assumptions were made regarding the data space of our database and hence various kernels tricks (sigmoid, polynomial, linear and radial) were also investigated for their MoA stratification performance.

2.2.2.3 Random forest

Random forest (RF) is an ensemble classifier that builds multiple decision trees, where within each decision tree the data is repeatedly split by the branches of the tree until a decision is made on which class the input data belongs to. From the multiple decision trees, multiple votes are obtained on which class the input data belong to and the majority class vote becomes the predicted class of the model. RF algorithms are useful as they do not require data to be normalised. RF models for multiclassification were built using the `RandomForest`, as well as `h2o` R packages [120, 123]. Hyperparameter tuning was implemented using the e1071 R package for `RandomForest` and an internal grid search

function was used for hyperparameter tuning for the h2o package [122]. Classification error was used to identify the optimal hyperparameters, however, for the mtries hyperparameter, the out-of-bag error was used to identify the optimal hyperparameter value. Since the hyperparameter, mtries, specifies how the data is to be subsampled during bootstrap aggregating when training the RF model, a mean prediction error i.e. out-of-bag error, can be calculated from samples not included in the training. Hyperparameters which had the lowest error was selected as the optimal architecture for the model. With the h2o package, hyperparameters were selected according to the Logloss value, also known as logarithmic loss. A high Logloss value indicates that the classification accuracy of the model is poor and vice versa [124]. Thus, optimal hyperparameters were chosen which had the lowest Logloss value.

2.2.2.4 Gradient boosting machine

Similar to RF, gradient boosting machines (GBM) builds multiple decision trees, however, based on the class vote of the decision tree built the subsequent tree is built to address the shortcomings of the previous decision tree and this is repeated until no further improvement is obtained or the number of trees specified has been reached. For multiclassification GBM models the xgboost R package, as well as the h2o R package was used [120, 125]. Hyperparameter tuning was implemented using the caret R package for xgboost and the internal grid search function was used for hyperparameter tuning for the h2o package [120, 126]. The caret package used classification accuracy as a performance measure in selecting the optimal hyperparameters, whereas the h2o package used Logloss to select the optimal hyperparameters.

2.2.2.5 Artificial neural networks

Artificial neural networks (ANN), analyses input data within its' hidden layers and may apply statistical functions within these hidden layers to better make decisions upon the data. The output layer of the ANN takes the information from the hidden layers and converts it into probabilities of the input data belonging to a class, whereby the highest probability becomes the predicted class. The h2o R package with the h2o.deeplearning() function was used to develop an ANN capable of multiclassification and a grid search was done to find the optimal model hyperparameters using Logloss as a performance measure [120]. Due to computational cost and efficiency, not all the ANN hyperparameters and/or large ranges could be investigated (see Appendix A: Table A.1).

2.2.3 Evaluating different machine learning algorithms in stratifying antiplasmodial compounds similar MoA

When it comes to big data problems and complex data structures it is not always clear which ML algorithm will perform the best. Hence, each of the above algorithms was assessed in their ability to correctly stratify compounds to their respective MoA using all the genes within our database as well as the biomarkers genes identified as predictive features. The reason for identifying and using these biomarkers is to enable efficient screening, which is the end-goal, by evaluating few genes rather than whole transcriptomes.

Model overfitting was evaluated by splitting the database into an 80:20 ratio, where 80% of the dataset was used to train the model and the 20% used as a test set to analyse the performance of the model predictions on untrained data. In parallel, K-fold cross-validation (K = 10) was performed except in this case the training data was used and subsampled to build the model in order to obtain a more precise estimate of the model's performance (Figure 9). K-fold cross-validation is valuable in that it does an internal validation of the data being used to train the model. It also can give an indication of a classifier's stability whereas the test set validation is an external validation and uses data the model has not been trained on. Many hyperparameter tuning algorithms use K-fold cross-validation to determine which parameters are more accurate. For our model assessment, both K-fold cross validation and the test data was used to analyse the performance of ML models in MoA stratification.

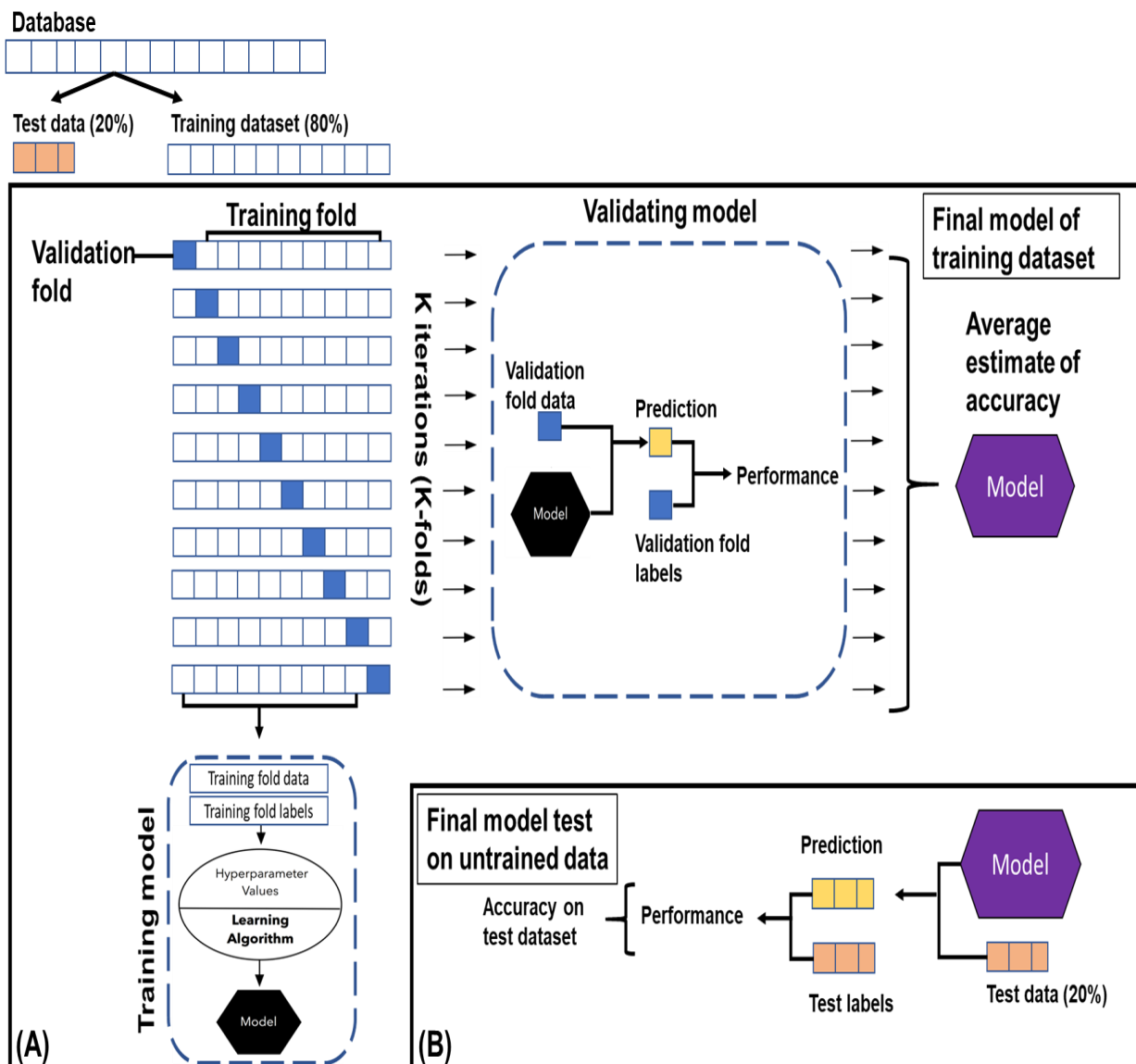


Figure 9: Assessing model performance using K-fold cross-validation and untrained test data.

Our GEP database of compound-treated *P. falciparum* parasites was split into a test and training dataset. (A) The training dataset will undergo 10-fold cross-validation whereby the training set is split into 10 sets. One set is selected as a validation fold and the remaining sets are the training fold. The training fold trains the model (black) using the algorithm and the optimal hyperparameters selected. The model is validated by assessing the model's prediction performance on the validation fold. This is done repeatedly until each set was a validation fold. The performance measures give an average accuracy of the final model (purple) trained on all the training dataset (all 10 sets). (B) The final model (purple) accuracy is then assessed on the test data that the model has not been trained on.

The caret package was used to perform 10-fold cross-validation on the random forest and SVM models made using the randomForest and e1071 R packages, respectively [122, 123, 126]. In instances where the h2o R package was used to make models (such as the ANN, MLR, and GBM), the 10-fold cross-validation was done simultaneously without the requirement of another RStudio package [120]. The best ML algorithm was then identified based on results from the 10-fold cross-validation, model stability and classification accuracy on test data.

2.2.3 Filtering criteria to identify predictive biomarker genes

With ML, if the training data contains much noise and variability, the model built on this data may recognize this and attach significance to this noise, thereby decreasing accuracy in prediction. Therefore, feature selection is an essential part of ML as it helps reduce the dimensionality of high dimensionality datasets (HDD) such as GEP data and also improves predictive accuracy by focusing on data that is relevant for the model [127].

To identify features i.e. genes that may be useful for MoA stratification, a rational feature selection criterion was developed as shown in Figure 10. Since algorithms can identify non-informative features or noise as signal, non-informative genes that are not significantly differentially expressed after treatment was removed (Figure 10.B). Informative genes with strong signals were extracted by identifying genes that had a log fold-change (FC) in the upper or lower 5th percentile. Genes that fell within the upper or lower 5th percentile of gene expression in the GEP during any time point in a compound treatment were defined as differentially expressed (DE).

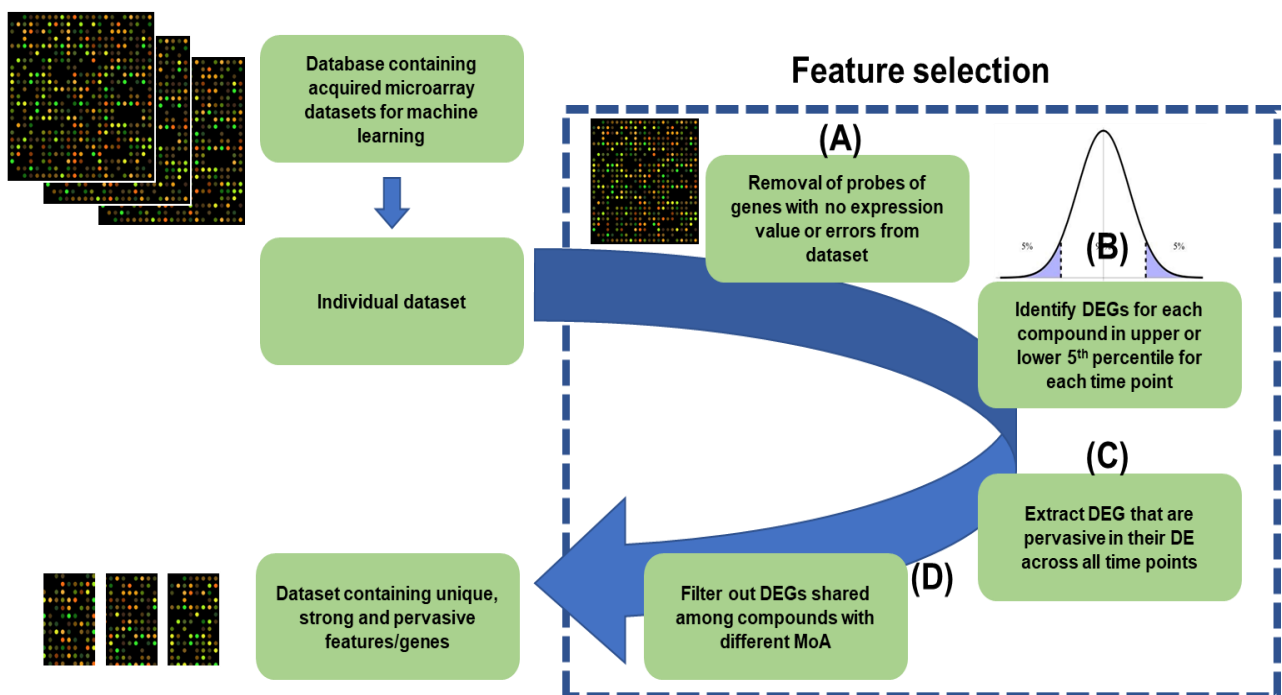


Figure 10: Feature selection filtering process to identify biomarker genes with unique predictive features.

From our database containing the accepted GEP datasets of compound-treated *P. falciparum*, individual datasets undergo feature selection whereby biomarker genes, i.e. important features for predictive modelling are identified. (A) Each dataset is pre-processed to remove gene probes with no signal. (B) After which DEGs are identified for each treatment and all their corresponding time points. (C) DEGs identified for each compound are filtered to extract DEGs that are pervasively DE across all time points for that compound. (D) The extracted DEGs with pervasive DE are then filtered to exclude DEGs shared among compounds with different MoA.

Although these genes may have a strong signal, this signal may not be continuous throughout treatment with a compound and vary across time points. Such genes may not function as predictive features as they are non-pervasive throughout various time points of a compound's treatment in a highly perturbed state. Hence, to reduce noise being introduced into the model, only genes that were continuously DE across all time points of a treatment were selected (Figure 10.C). Genes that were not continuously perturbed throughout the time points were thus not considered as strong predictive features. In this way, unwanted variability and 'noisy' genes were removed. This was done for each time point of each compound's treatment, however, to ensure our filter selection was not biased towards immediate transcriptional responses pervasively DEGs that only occurred after of treatment were also included.

Not all pervasive DEGs may be due to a compound and can be the result of general drug stress or other factors. Such DEGs are likely to be shared between compound treatments and would not be informative or useful in our MoA stratification model. To ensure such genes were excluded, the pervasive DEGs extracted were further filtered to obtain pervasively DEGs specific for a compound's treatment that is not shared with other compound treatments with different MoA (Figure 10.D).

2.3 Validation of rational feature selection through a comparison to algorithm inferred biomarkers

We subsequently evaluated if the features from our rational feature selection approach compare to features objectively identified in an unsupervised fashion by the ML algorithm as good features for MoA stratification. Since our feature selection is based on a rationalized, subjective criterion that makes assumptions on what a good predictive feature for compound MoA should be, potentially good features that do not comply with this criterion may be excluded. Hence, we wanted to validate our features by comparing the model built from them to a model built from features that the best ML algorithm identified top features for MoA stratification.

For this, the 2463-gene database model was used by the best ML algorithm to produce a ranked list of genes that the algorithm gave importance to for MoA stratification when determining compound MoA. From these top ML-inferred genes, a model was built using the same number of genes/features (174) that was present in our rationally selected biomarker model to allow proper computational comparison between the two models.

Based on results from the 10-fold cross-validation, model stability and classification accuracy on test data the best feature selection approach (ML-inferred vs rational feature selection) was identified.

2.4 Optimisation of the minimum number of features for robust MoA stratification

Since non-informative features can cause lower accuracy and instability within a ML model, to be thorough in selecting which feature selection approach was more appropriate, a sliding gene-scale method was implemented. This is because as the number of features decreases, features that are noisy or redundant will become more apparent in affecting the model's performance by lowering the accuracy of the model, whereas the opposite is true for good predictive features. This would also help with the primary goal in identifying a small subset of biomarker genes with predictive features regarding a compound's MoA that can generate a robust model with low variability and good performance in stratifying compounds together with similar MoA.

To identify the minimum number of genes most representative of compound MoA without compromising on model performance, genes identified from the rational feature selection or inferred ML method underwent a sliding gene-scale approach. Genes were ranked according to their importance in MoA stratification and from this 'minimodels' were made with each sequential model containing fewer training features than the previous model. The minimodel that performed the best in their classification accuracy, model stability and test set (untrained data) using the least number of features determined which approach was more suitable for feature selection. From this also the minimum number of features for robust MoA stratification was determined.

Chapter 3: Results

3.1 Identifying predictive biomarker genes for MoA stratification

3.1.1 Data acquisition and quality control filtering

To build a model that stratifies compounds with similar MoA together, GEPs of compound treatments whose MoA is known in *P. falciparum* were used to train the model. In total, 14 publicly available datasets were obtained, representative of responses due to 42 compound treatments (Figure 11) but filtering for known MoA resulted in the use of only 9 datasets.

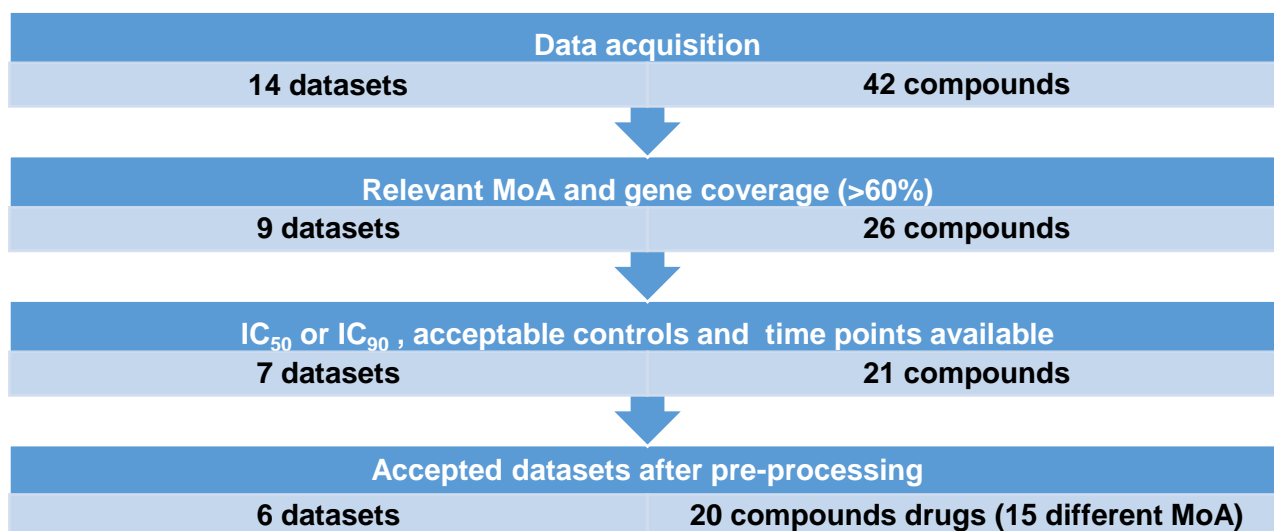


Figure 11: Quality control summary of compound-treated *P. falciparum* GEP datasets acquired.

GEP datasets of compound-treated *P. falciparum* parasites were filtered according to criteria in Table 4. During each filtering step datasets and compounds were excluded until 6 GEP datasets were obtained containing 20 compound treatments. In total, these datasets have 15 different MoA.

Subsequently, to ensure that the model will be robust enough to be applied over a broad time-frame and not constrained or biased towards a certain stage of the parasite, only datasets that contained multiple time points for compound treatments were included. This resulted in 2 of the 9 datasets excluded as they only contained one time point each. If, however, a compound in a dataset with only one time point had the same MoA as a compound in another dataset with multiple time points, these compounds were accepted. The reasoning behind this was that broad time-frame biomarkers genes identified for MoA would not show a difference in expression and be biased towards different compounds with similar MoA in different datasets. Hence, the histone deacetylases (HDA) inhibitors in the Andrew *et al.* dataset (Table 5), despite only having one time point were accepted since such inhibitors were also present within the Hu *et al.* dataset containing multiple time points.

This resulted in 7 datasets accepted, representative of 21 compound treatments and 17 different MoAs. These datasets were pre-processed for probe quality since raw GEP data

was used and this resulted in the exclusion of 1 dataset due to the gene coverage falling below 60%. The final database used for further analyses was obtained from 6 datasets and represented 20 compound treatments, spanning 15 different MoAs (Table 5).

Table 5: Final database generated from 6 datasets spanning 20 compound treatments

Compound	Mode of action	Ref	Dataset	GEO No	Time points	Gene coverage after pre-processing
W7	Calcium/calmodulin-dependent protein kinase inhibitor	[93, 110]	Hu <i>et al.</i> , 2009	GSE1 9468	4-5 time points per treatment	3705/5400 (69%)
ML-7		[93, 128]				
Staurosporine	Inhibits serine/threonine kinases, reduces merozoite invasion	[129, 130]				
Cyclosporin A	Has a strong affinity to sphingomyelin in membrane environment like parasitized erythrocytes membranes, thus aids in inhibiting merozoite invasion. Also believed to be a calcineurin pathway inhibitor.	[93, 131]				
Colchicine	Microtubule is the target, inhibits merozoite invasion	[132]				
PMSF	Serine protease inhibitor	[133]				
Leupeptin	A cysteine, serine, and threonine peptidase inhibitor which affects haemoglobin degradation	[134]				
Artemisinin	Partially understood but hypothesized to be involved in producing carbon-centered free radicals that in turn alkylate heme and proteins	[135]				
Chloroquine	Inhibits the heme polymerase enzyme	[74]				
Febrifugine	Targets <i>P. falciparum</i> prolyl-tRNA synthetase activity	[136]				
Quinine	Partially understood but accumulate in the parasite's digestive vacuole (DV) and may inhibit the detoxification of heme	[137]				
DFMO	Inhibits ornithine decarboxylase causing parasite arrest	[138]	van Brummelen <i>et al.</i> , 2008	GSE1 3578	3 time points with replicates	4050/5400 (75%)
MMV 048 and MMV 943	Inhibits <i>Plasmodium</i> phosphatidylinositol 4-kinase (PI4K)	[139]	Connacher <i>et al.</i> , 2016	GSE1 0069 2	2 time points each	4971/5400 (92%)
ACT-213615	Artemisinin derivative that has an unknown MoA which is different from other antimalarials based different transcriptional responses to that of the Hu <i>et al.</i> dataset	[110]	Brunner <i>et al.</i> , 2012	GSE3 9485	5 time points	4857/5400 (90%)
Ionomycin	Increases cytoplasmic calcium concentrations	[115]	Cheemadani <i>et al.</i> , 2014	GSE3 3869	5 time points	4495 /5400 (83%)
Trichostatin A (TSA), Suberoylanilide hydroxamic acid, 2-aminosuberonic acid derivative, Apicidin	Histone deacetylase (HDAC) inhibitors that perturb the transcriptome	[111]	Hu <i>et al.</i> , 2009	GSE1 9468	4-5 time points	3705/5400 (69%)
		[93, 140]	Andrews <i>et al.</i> , 2012	GSE2 5642	1 time point	4364/5400 (80%)

Table 5 summarises the 6 datasets (3500-4000 genes each) that were included in our transcriptional database for further analyses. The 20 compound treatments included

spanned a variety of MoAs, ranging from cell signalling (Ca²⁺/calmodulin protein kinase inhibitors and S/T kinase inhibitors) to heme metabolism and transcription (heme polymerase enzyme inhibitors and histone deacetylases inhibitors). Most of the datasets had a high gene coverage, except in the case of the Hu *et al.* dataset (69% gene coverage) but since this dataset contained most of the compounds in our database and the gene coverage after merging these arrays was still acceptable, it was retained. Most treatments in these datasets had 3-5 time points to allow for feature selection of biomarkers present over a broad time-frame.

The database consisting of the 6 datasets (20 compounds) averaged around 3500-4000 genes each. To allow comparison between datasets, only genes within GEPs shared between the datasets were used. This resulted in a reduction of genes in the final database to 2463 (Figure 12).

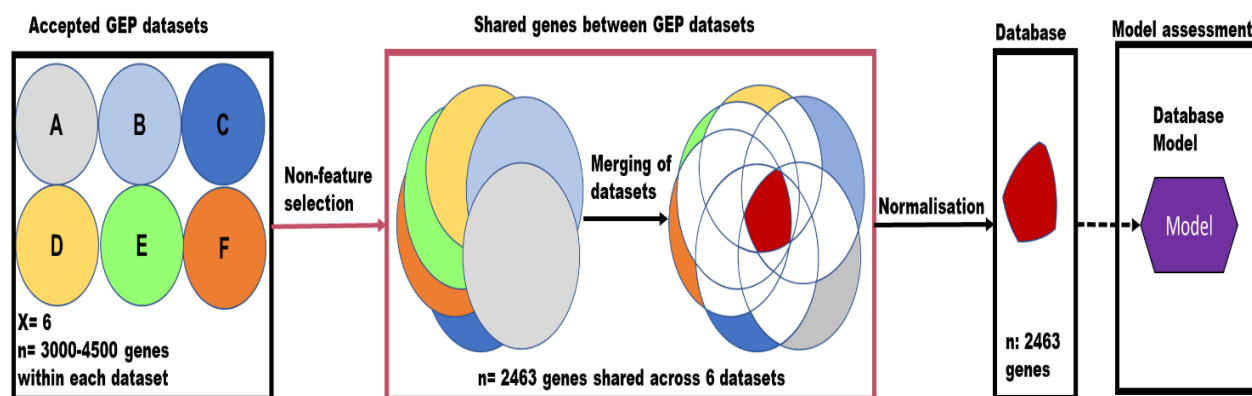


Figure 12: Merging and normalisation of accepted datasets to form our 2463-gene database

Following data acquisition, accepted datasets containing the whole GEP of these 20 compound treatments underwent non-feature selection (pink) to identify genes shared among GEPs that will allow the merging of the 6 datasets to form our transcriptional database. Genes not shared between datasets had to be excluded as genes lacking in other datasets will impede gene expression comparison between treatments and prevent ML model building. After merging the resultant database underwent normalisation to allow comparison between datasets. This resulted in our transcriptional database consisting of 2463 genes that were used for model building and assessment.

After pre-processing of raw probes and merging of the 6 datasets into the single 2463-gene database, different normalisation strategies were investigated to assess which would be optimal in allowing proper comparison between treatments of different datasets. As can be seen in Figure 13.A, DEG data from compound treated parasites and their controls do not have the same distribution and can thus not be compared to other datasets. Some of the datasets within our database were more skewed compared to the other datasets and would make comparison difficult. This disparity between GEP datasets could be a result of artefacts that were introduced from hybridization or washing [141]. Hence, different normalisation strategies were investigated for their ability to correct artefacts in the datasets

and allow effective comparisons between datasets. This included quantile normalisation (Figure 13.B), medium scaling normalisation (Figure 13.C) and cyclic loess normalisation (Figure 13.D).

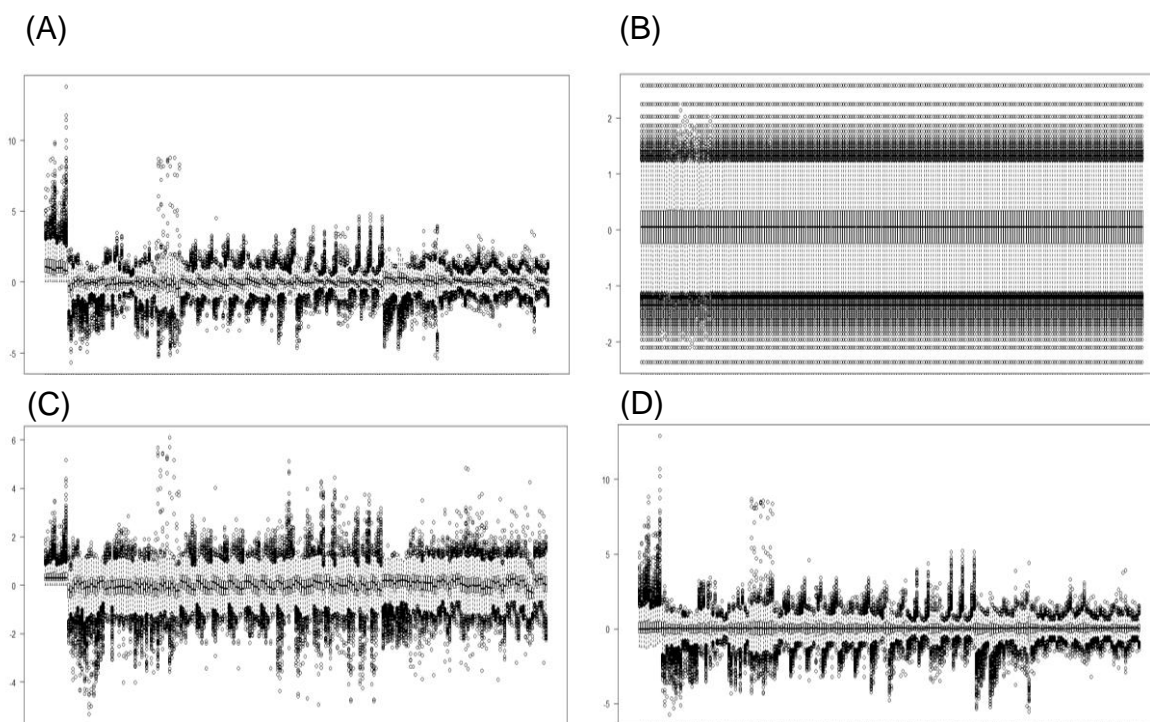


Figure 13: Normalisation strategies applied to the 2463-gene database

(A) Unnormalized vs different array normalisation strategies were implemented such as (B) quantile normalisation, (C) medium scaling normalisation and (D) cyclic loess normalisation. Data from 6 datasets (Table 5) was used, with a total of 200 time points (i.e. treatment and control time points), ranging over 20 different compound treatments.

The quantile normalisation strategy was very strict and artificially removed any variability between datasets enforcing homogeneity which would be inaccurate (Figure 13.B). Quantile normalisation makes the overall distribution of expression values for each GEP identical to one another. However, it does so by transforming the highest expression value in a GEP to the mean of the highest values of all the GEPs. This is repeated until the lowest expression value in all the GEPs is transformed into the mean of the lowest values [142]. Although this strategy removes the artefact and allows for comparison, it also removes any heterogeneity between treatments and controls. By applying such strict normalisation, we would run the risk of also removing important expression patterns within compound treatments that could be useful during ML for MoA stratification. By contrast, medium scaling normalisation (Figure 13.C), which scales GEPs using the median absolute deviation was unable to solve artefacts [143]. It also did not aid in allowing comparison to other datasets and thus was not a good normalisation strategy.

Cyclic loess normalisation relies on the difference in log expression values (M) and the average of the log expression values (A) to draw a MA-plot to obtain a loess curve [144]. This loess curve is then used to apply a correction factor to the arrays being normalised [142]. When cyclic loess normalisation was implemented (Figure 13.D), the datasets are transformed in such a way as to allow comparison, including datasets with artefacts which had shown a very skewed distribution compared to other treatments in the unnormalized data (Figure 13.A). This normalisation strategy helped in correcting the artefact that may have been introduced during microarray preparation and reading as well as transformed the datasets to allow for cross comparisons. Cyclic Loess normalisation was thus employed on the 2463-gene database and the log fold-change of each treatment for each gene calculated. The normalised 2463-gene biomarker database (with 253689 individual datapoints, spanning the 103 time points) was then used to explore different ML algorithms for their efficacy and relevancy for subsequent use in MoA stratification.

3.1.2 Evaluating different machine learning algorithms on the 2463-gene database

ML models were generated using 202951 (80%) of the data points (defined as the training set) within the 2463-gene database after hyperparameter tuning (see Appendix B: Table B.1), and the algorithms were compared against one another based on accuracy scores from k-fold cross validation and performance on the test set data (the remaining 20% of the data points)(Figure 14).

Amongst the ML algorithms, models generated with both polynomial and linear kernel SVCs displayed similar variability within accuracies of $78.2 \pm 17.7\%$ and $81.67 \pm 14.19\%$, respectively, compared to MLR which showed a lower variability ($77.8 \pm 11.6\%$). The ensemble classifier set performed slightly poorer with the majority of the models displaying accuracies between 73-77%, with GBM (h2o) the most inaccurate at 62%. Models built with RF (randomForest) performed comparably well with the MLR model at $77.3 \pm 10.6\%$, with the RF (h2o) also displaying very little variability (10.06%) although quite poor accuracy (73.35%). Quite a high variability was observed (up to 20%) for the remaining ensemble classifiers. The same large variability within classifier accuracy was observed for the h2o ANN algorithm (accuracy at $74.77 \pm 14.17\%$).

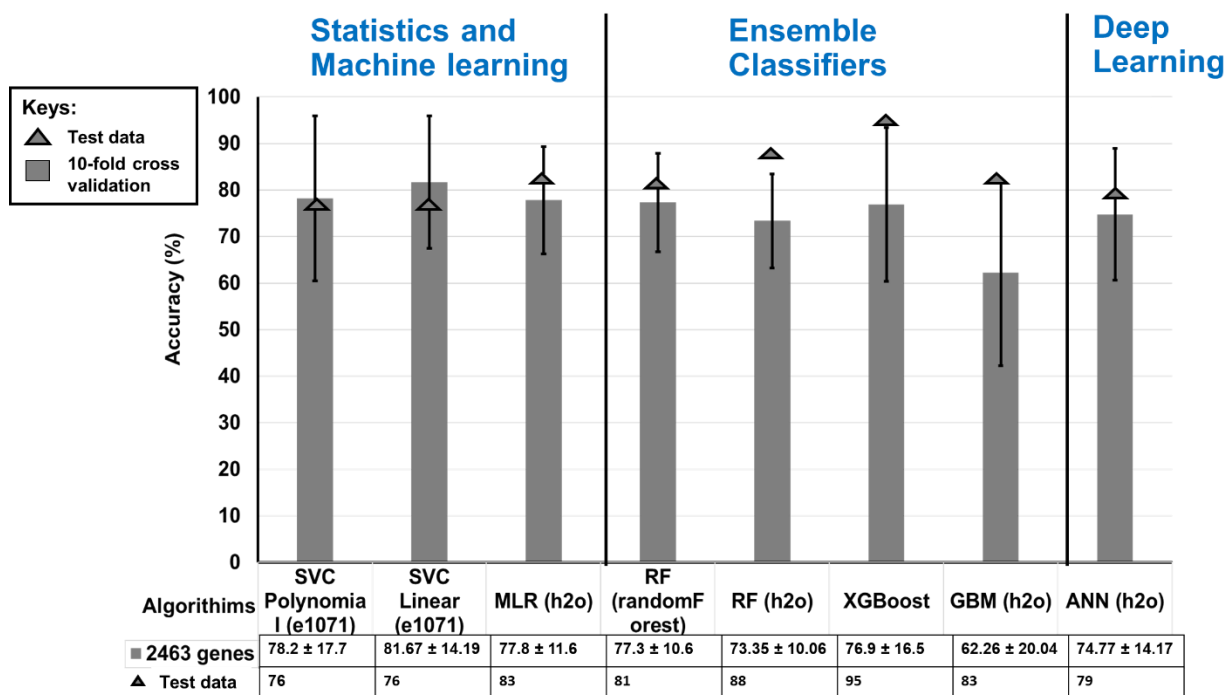


Figure 14: Robustness and accuracy of different ML algorithms ability in stratifying treatments with similar MoA within our 2463-gene database.

The accuracy of MoA class stratification of different ML algorithms is grouped according to whether they involve either statistics, ensemble classifiers or deep learning. Algorithms classifiers used all the genes in the database (2463) as training features. Classifiers were hyperparameter tuned before undergoing 10-fold cross-validation. Bars indicate the average accuracy of the classifier obtained from 10-fold cross-validation on the training data and the error bars are the standard deviation of performance measures. Triangles indicate the accuracy of the classifier in stratifying the MoA of test data. SVC= support vector classification, RF=random forest, GBM=gradient boosting machine, ANN= artificial neural networks. R packages are shown in brackets.

During the evaluation of the performance of models on the test set, both the SVC algorithms perform poorer than their average accuracy achieved on the training set, indicating that these models may have become overfitted to the training set and struggles in stratifying untrained data. However, all the other algorithms performed well with the test data. Both the MLR and RF algorithms displayed similar efficacy as evaluated by accuracy, variability and ability to perform on test data, therefore, these algorithms are useful to generate models on larger datasets that are informative to stratify compounds based on specific MoAs.

As stated before, monitoring the gene expression of 2463 genes per treatment will not allow a medium-throughput method for MoA stratification. Not only this, but most of the algorithms build models showing high variability within their accuracy (>10%) in MoA stratification. This can be resultant of including genes as training features that are 'noisy' and non-informative for MoA stratification. Including such features can compromise a model's accuracy and stability as models would give weight to such features. Hence, to exclude such non-informative genes and select biomarker genes representative of a compound's MoA, we

employed a rational feature selection approach to identify a set of biomarkers able to stratify a compound's MoA.

3.1.3 Biomarker gene selection

The previously accepted 6 datasets underwent individual feature selection to extract the minimum informative features to be used as biomarkers and to establish a 'biomarker' model (Figure 15).

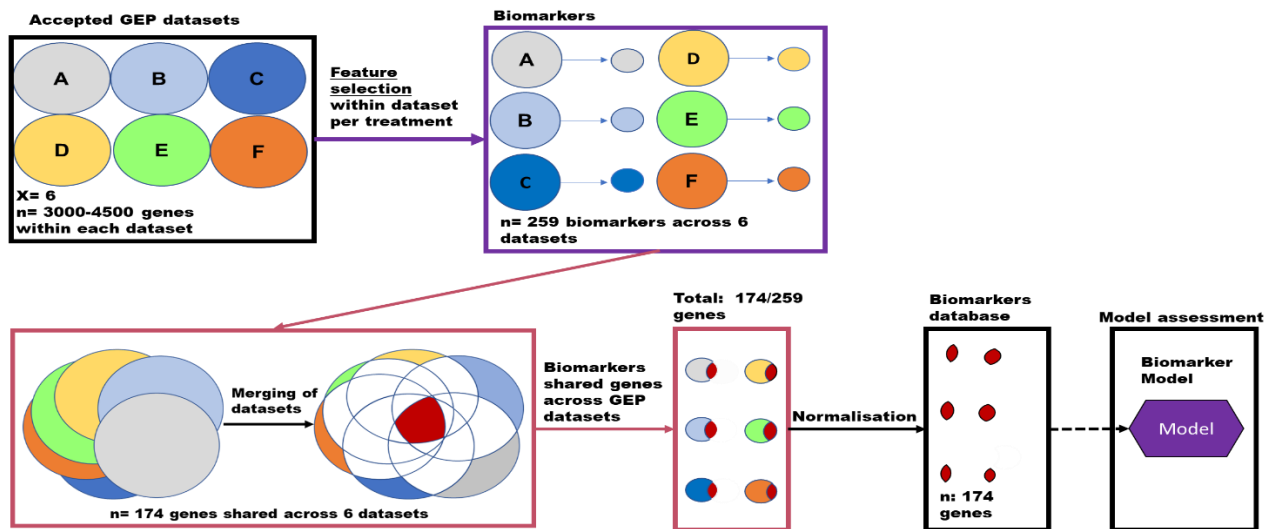


Figure 15: Summary of feature selection and merging of datasets to form the 174-gene biomarker database

Accepted datasets underwent individual feature selection (purple) to identify biomarker genes for each compound within each dataset. This yielded 259 genes showcasing unique and continual DE across treatments for 20 compounds encompassing 15 different MoA. Thereafter the 6 datasets underwent non-feature selection (pink) to identify genes shared among GEPs that will allow the merging of the 6 datasets to form our biomarker database. Genes not shared between datasets were excluded as genes lacking in other datasets will impede gene expression comparison between treatments and prevent ML model building. This resulted in our biomarker database consisting of 179 genes out of the 259 biomarker genes that were identified.

Features were defined as DEGs that were identified for each compound treatment using the expression profiles for each gene set derived from multiple time points within the individual datasets. This was performed on a total of 103 time points in the complete set of data for the 20 compounds. The DE criteria enforced for selecting these as features include that DE expression had to be present in the upper or lower 5th percentile of a compound's treatment time point(s). This resulted in identifying the majority of genes within the parasite's transcriptome (3146, Figure 16) as potentially DE within all 6 the datasets, present in at least one time point. However, to enforce continuity in DE over time, genes were only included if their DE profile was maintained during treatment time points for a compound. This second filtering step limited the number of DEGs identified for each

treatment to a total of only 338 DEGs (Figure 16, orange bars), which show continuity in DE throughout all time points for a specific treatment.

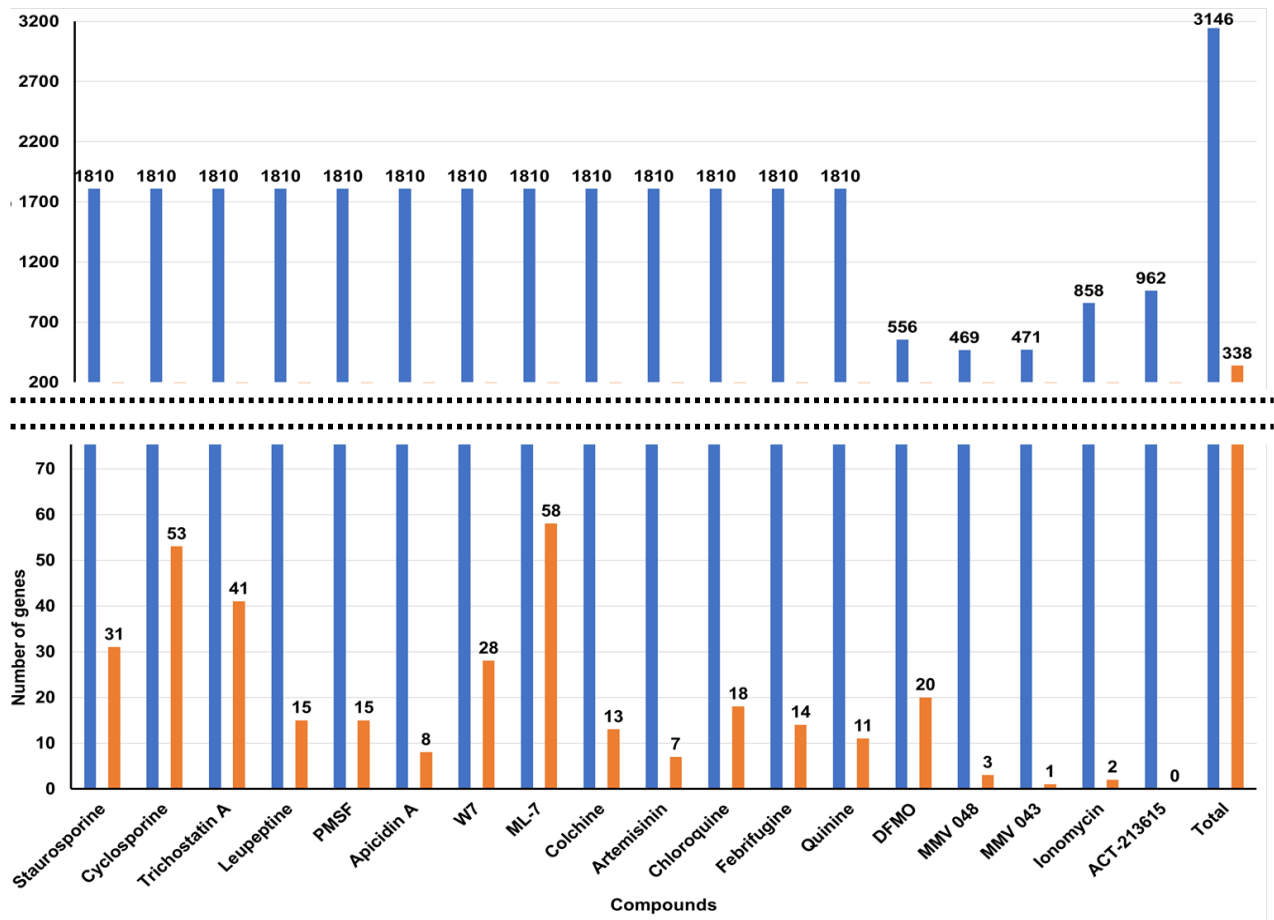


Figure 16 : Filtering of DEGs to DEGs pervasive over time.

Blue bars are the total DEGs obtained from all individual time points of each treatment. Orange bars are the genes obtained after filtering DEGs obtained for each treatment to DEGs which are pervasively DE across all time points for that treatment.

Only 18 of the 20 compounds in our database are shown in Figure 16 since the two excluded compounds (Suberoylanilide hydroxamic acid and 2-aminosuberic acid derivative) are within the Andrew *et al.* dataset which does not have multiple time points to implement our feature selection criteria for biomarker genes. However, since these compounds share similar MoA to Trichostatin A and Apicidin A, we expect similar transcriptional responses from the biomarker genes identified. The inclusion of these compounds from the Andrew *et al.* dataset, will thus assess whether the biomarkers identified from our feature selection are biased towards independent datasets containing compounds with similar MoA. In contrast to this, the artemisinin derivative ACT-213615, was included to test whether the MoA stratification model can correctly discern a compound as having a different MoA than others within the database despite ACT-213615 obtaining no biomarkers from our feature selection.

With this filtering step in selecting DEGs with continuity in DE across treatment time points, the number of genes was drastically reduced from the initial number of DEGs obtained and allowed a reduction in the dimensionality of our biomarkers that will be used to build our ML models. Therefore, the 338 pervasively DEGs were further interrogated to eliminate DEGs which were defined as ‘promiscuous’ or non-specific since they were present within multiple compound treatments. These DEGs could be indicative of general stress responses or other factors, rather than be associated with a specific MoA of individual compounds and therefore had to be removed. Amongst compounds with previously indicated similar MoAs (chloroquine, quinine, febrifugine, and artemisinin, all supposed to interfere with heme degradation and hemozoin formation), surprisingly, only 4 DEGs were shared amongst more than two compound treatments (Figure 17) and were therefore removed. This included one DEG (PF3D7_1235000) shared between febrifugine, quinine, and artemisinin which is a putative gene with a PIH1 domain-containing protein. From this, we can see that although the compounds share similar pervasive DEGs, this is not necessarily due to similar MoA.

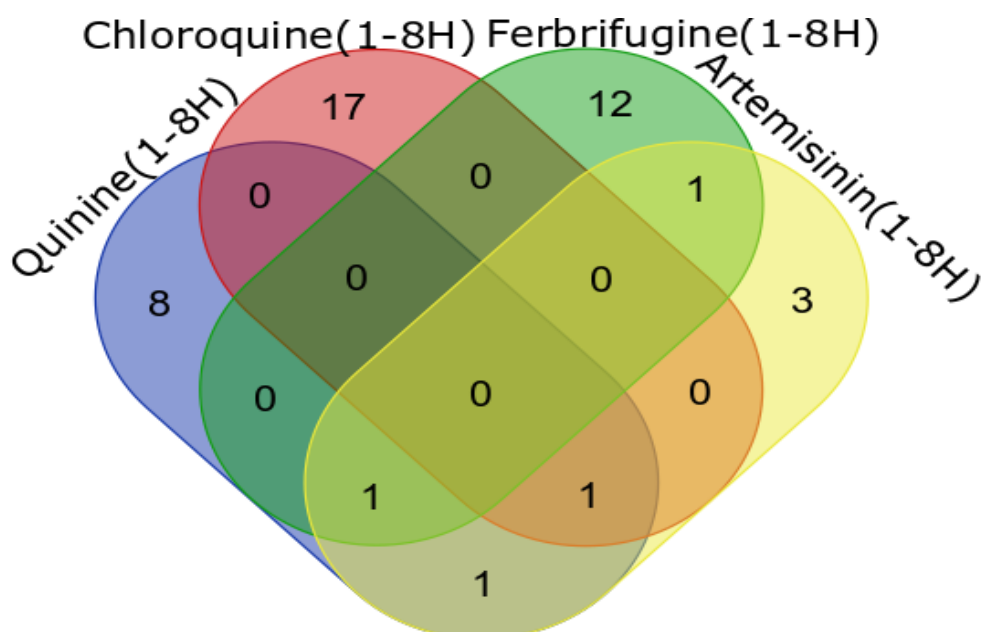


Figure 17: Identification of pervasive DEGs unique to individual treatments. The number for unique pervasive DEGs were identified for quinine (8), chloroquine (17), febrifugine (12), and artemisinin (3).

When the complete 338 DEG dataset was evaluated, the majority of the pervasive DEGs identified for a treatment were unique for that treatment (Figure 18). Only 79 DEGs were shared between the compound treatments, including genes involved in protein degradation (e.g. the ubiquitin-conjugating enzyme E2) and genes involved in transcription and

translation (such as 60S, 40S rRNA and putative zinc finger proteins), which could be indicative of general stress. The exclusion of these promiscuous genes resulted in 259 unique and pervasively DEGs that were further evaluated.

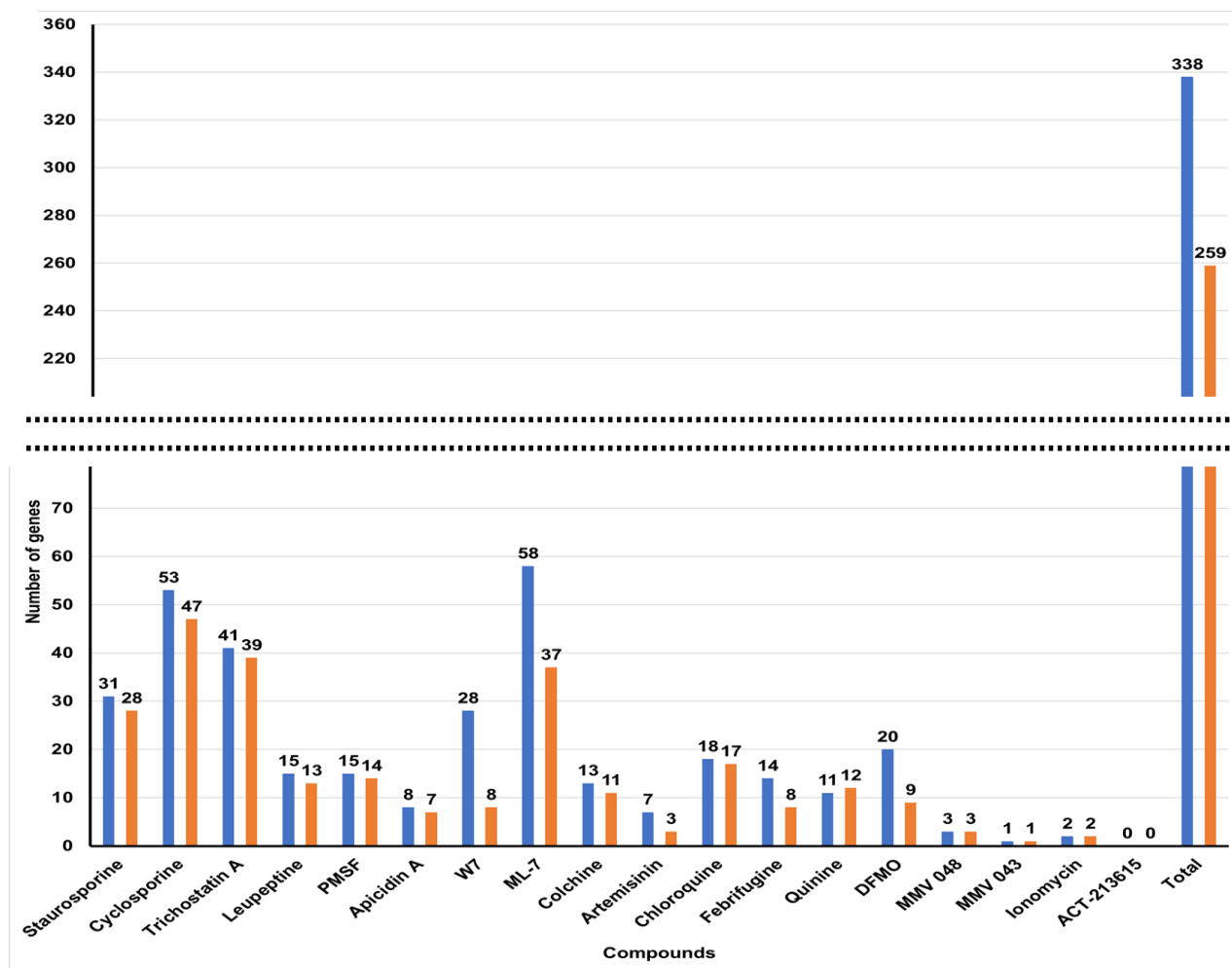


Figure 18: Pervasive DEGs compared to pervasive DEGs that are unique to treatments. The blue bars represent the pervasive DEGs identified for each treatment, whereas the orange bars are the subset of the pervasive DEGs identified for a treatment that is not shared with other treatments.

To allow subsequent use of the biomarker genes in generating a ML model, the 259 genes were further interrogated since they have to be present between all 6 datasets to allow comparison of GEPs between the different compound treatments in the 6 datasets. Therefore, the 259 genes were further reduced to 174 since incomplete gene coverage of some datasets resulted in the loss of some biomarker genes during merging (Figure 15). These 174 genes are therefore represented in each of the 20 compound treatments in the 6 datasets and cover all 103 time points in total between the 20 compound treatments. Since these genes were already within our 2463-gene database that underwent normalisation, these 174 genes and their expression under various treatments were extracted and used to form the 174-gene biomarker database. The 174-gene biomarker

database (with 17922 individual datapoints, spanning the 103 time points) was then used for building the predictive biomarker models.

3.2 Building predictive biomarker models

3.2.1 Evaluating different machine learning algorithms for the 174-gene biomarker database

Different algorithms were again used as before to generate models on a training set from the 174-gene biomarker database. This is to evaluate the performance of the different algorithms on a database with 90% fewer features (174) than that found in the previously evaluated full 2463-gene database (Figure 19). The training set was again arbitrarily generated by using 80% of the data points within the biomarker database. The remaining 20% was used later as a test set. After hyperparameter tuning (see Appendix B: Table B.1), the algorithms were compared against one another based on accuracy scores from 10-fold cross validation and performance on the test set data.

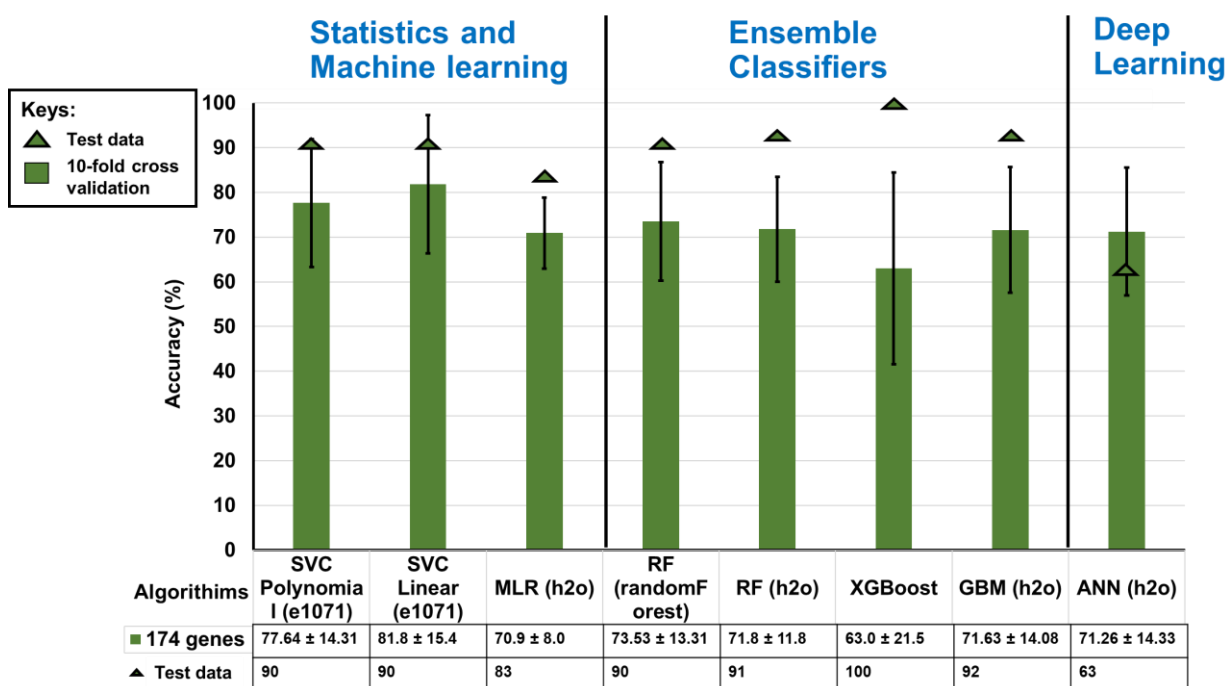


Figure 19: Robustness and accuracy of different ML algorithms ability in stratifying treatments with similar MoA using the 174-gene biomarker database.

The accuracy of MoA class stratification of different ML algorithms is grouped according to whether they involve either statistics, ensemble classifiers or deep learning. Classifiers were hyperparameter tuned before undergoing 10-fold cross-validation. Bars indicate the average accuracy of the classifier obtained from 10-fold cross-validation on the training data and the error bars are the standard deviation of performance measures. Triangles indicate the accuracy of the classifier in stratifying the MoA of test data. SVC= support vector classification, RF=random forest, GBM=gradient boosting machine, ANN= artificial neural networks. R packages are shown in brackets.

Biomarker models were generated using both polynomial and linear kernel for SVCs (Figure 19), with high variability and accuracies of $77 \pm 14.31\%$ and $81.8 \pm 16.4\%$, respectively. By contrast, the MLR, although slightly less accurate at 70.9%, was very robust with very little variability observed ($\pm 8.0\%$) (Figure 19). Within the ensemble classifier set, the majority of the models had fair accuracy $\sim 73\%$, except for XGBoost at 63%. However, in all instances, high variability was observed, ranging between 11-21%. The same large variability was observed for the h2o ANN algorithm (accuracy at $71.26 \pm 14.33\%$).

To evaluate the performance of the models on untrained data, the remaining 20% (3584 data points) from the 174-gene biomarker database was again used as a test set. From this, only the h2o ANN algorithm for deep learning performed poorly, indicating overfitting of the model to the training set due to the inability of the model to generalize and recognize patterns in untrained data (63%). All the other algorithms performed well with the test data. Although ensemble classifiers did very well in predicting the test set, these algorithms do not perform well during 10-fold cross validation. Since this cross validation splits the training data to get a better estimate of the model's accuracy, it is shown that these algorithms are not reliable when the number of samples is reduced. Considering the lack of GEPs of compound-treated *P. falciparum* parasites, it is unlikely that we would be working with big datasets and adequate sample sizes in the future. Because of this, we require an algorithm that does not show such high variability in accuracy when presented with low sample numbers for testing and training, like that of the MLR algorithm which has shown to be better suited in handling such sample constraints.

Taken together, the MLR algorithm generated the most effective model for the 174-gene biomarker database based on its' good accuracy, combined with low variability and good performance on test data.

3.2.2 Validation of feature selection in the biomarker database

The MLR-based model generated on the biomarker database was therefore considered the most effective and was subsequently evaluated and validated for its ability to stratify the MoA of antimalarial compounds. However, since this model was based on a minimum number of features that were objectively selected to form a rationally selected 174-gene biomarker database, we interrogated if the MLR algorithm can extract better predictive features than those we selected if applied in an unsupervised fashion on a larger dataset.

To address this, the output of the MLR model built on the full 2463-gene database was therefore used to rank genes according to their importance in stratifying compounds to their MoA (Appendix B: Table B.2). From these, the top 174 genes were selected as a separate and new biomarker database, which we define as a ML-inferred biomarker database to compare and distinguish it from our rational selected biomarker database. The number of genes in these datasets was consciously kept the same to allow comparative computational analysis.

The ML-inferred 174-gene biomarker database was subsequently used to generate a new MLR model to compare to the MLR model from the rational selected 174-gene biomarker database for accuracy and variability (Figure 20).

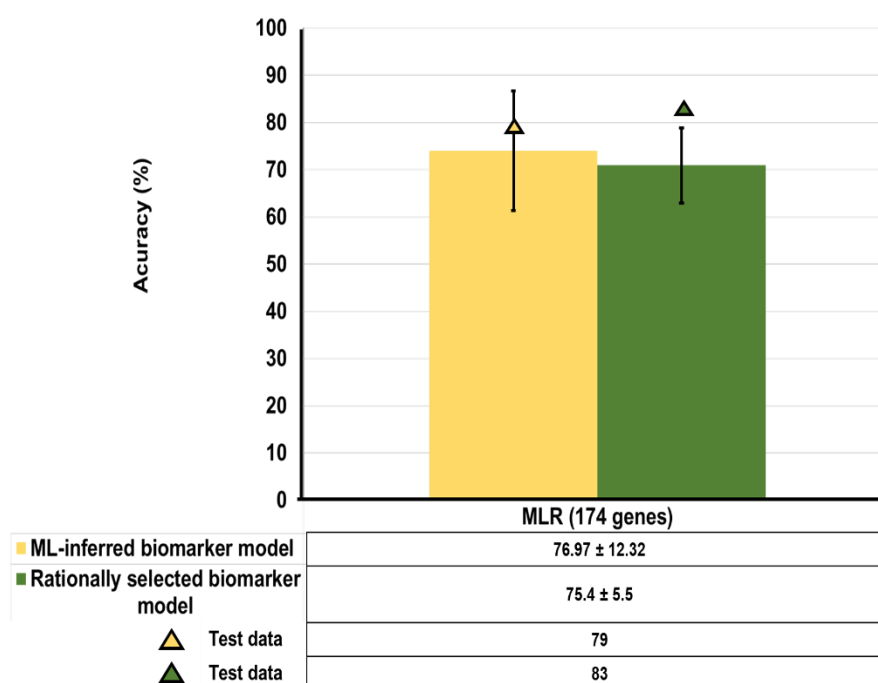


Figure 20: ML-inferred features vs rationally selected features

Biomarker MLR models for MoA stratification were generated by using features deemed as important by the MLR algorithm for MoA stratification (yellow) or using features identified through our rational feature selection criteria (green). Both models were built on 174 training features and underwent 10-fold cross validation and evaluation of performance on untrained test data. Bars indicate the average accuracy of the classifier obtained from 10-fold cross-validation on the training data and the error bars are the standard deviation of performance measures. Triangles indicate the accuracy of the classifier in stratifying the MoA of test data.

Similar accuracies were obtained for both feature selection approaches (Figure 20), however, the ML-inferred approach generated a more variable model ($76.97 \pm 12.32\%$) compared to the rationally selected biomarker model ($75.4 \pm 5.5\%$). Together with this, the rationally selected biomarker model also performed better by correctly stratifying compounds to their MoA of untrained data (83%). From this, we could validate that the

rational feature selection approach was appropriate in the selection of features important for MoA stratification. This could, however, change if the number of training features used to build the model is reduced as non-informative features can be removed and improve model performance or persist and reduce model accuracy and stability. Thus, both approaches underwent optimisation to identify the minimum number of features that can generate a model with robust MoA stratification.

3.2.3 Optimisation of the number of features in our MLR biomarker model

To further optimise the biomarker models, we used a sliding gene scale approach to build smaller models (minimodels) with sequentially fewer features to remove non-informative features and obtain the minimum feature model (Figure 21). To do this, biomarker genes from both approaches were ranked similar to that which was done with the 2463-gene database using the ML-inferred approach (Appendix B: Table B.2 and Table B.3).

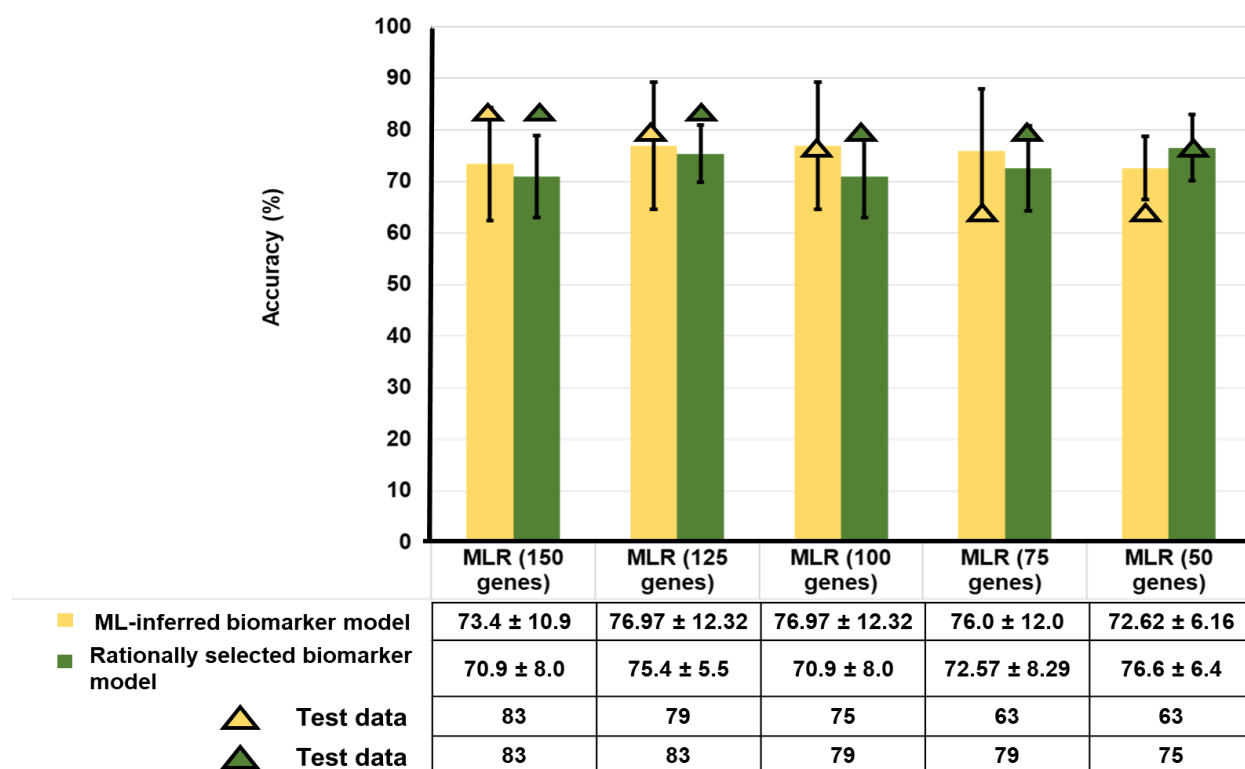


Figure 21: Influence of limiting the number of genes used for training on MoA stratification of MLR models

MLR classifiers trained on either ML-inferred features (yellow) or on rationally selected biomarker genes (green) were used to extract a list of ranked genes. Using variable importance, genes were ranked according to their importance in making classification decisions for the MLR classifier. With the ranked genes a sliding gene-scale approach was applied where the top genes were used to make minimodels with each sequential model containing a decreasing number of genes/features used to train the MLR classifier. Minimodels underwent 10-fold cross-validation and was also assessed in the accuracy of MoA stratification on test data.

From Figure 21, the average accuracy of the rationally selected biomarker minimodels is either maintained or increased above 70%, even when lowering the number of training features to 50. The variability within the minimodels seems to decrease as the number of training features are reduced, with the minimodel using the top 75 biomarker genes being an exception ($76.0 \pm 8.29\%$). When evaluating the performance of these minimodels on the test set, the minimodels obtain an accuracy of 75% or above. The opposite is observed for the ML-inferred biomarker minimodels, as the variability of the minimodels increases with reduced features (10.9% increased to 12.0%) and a decline in performance on test data is seen which indicates overfitting. This again reaffirms that the features from our rational feature selection approach are more suitable for MoA stratification than that of ML-inferred features. Although a gradual decline in performance on the test set is observed as the number of features is reduced for our rationally selected biomarker minimodels (75%), it is not to the extent as that within the ML-inferred biomarker minimodels (63%).

Based on the test set performance, as well as accuracy, model variability and the least number of training features used, the rationally selected 50-gene biomarker minimodel ($76.6 \pm 6.4\%$) was selected as the optimal minimum number of features for robust MoA stratification of compounds. This minimodel obtained good accuracy with the lowest variability within the model as well as used the least number of features without resulting in model overfitting to training data.

3.2.4 Interrogation of the top 50 features from the MLR biomarker model as indicators of MoA

The 50-feature minimodel from the rationally selected MLR biomarker model was subsequently manually interrogated (Figure 22). These top 50 genes are biomarker genes identified from 14 of the 20 compounds which account for 12 of the 15 MoAs of all the compounds in our database. Within these top 50 biomarker genes, no biomarkers for compounds such as colchicine, leupeptine, apicidin A, two HDA and ACT-213615 were present, though this did not decrease the model's accuracy in MoA stratification. Interestingly, some compounds contribute more to the overall features used in the MLR model than others. For example, from the artemisinin treatment, only two biomarker genes are utilized, whereas for the trichostatin A treatment five biomarker genes are utilized. The majority of the compound biomarkers which contribute much to the overall features of the model are inhibitors to proteins that serve important and global functions within a cell, such as kinases and deacetylases as we can see from Figure 22.

Mode of action
Increases cytoplasmic calcium concentrations
Hypothesized to be involved in producing carbon-centered free radicals that in turn alkylate heme and proteins
Inhibits the heme polymerase enzyme
Partially understood but accumulate in the parasite's digestive vacuole (DV) and may inhibit the detoxification of heme
Targets <i>P. falciparum</i> prolyl-tRNA synthetase activity
Inhibits <i>Plasmodium</i> phosphatidylinositol 4-kinase (PI4K)
Histone deacetylase (HDAC) inhibitors that perturb the transcriptome
Inhibits ornithine decarboxylase causing parasite arrest
Calcium/calmodulin-dependent protein kinase inhibitor
Inhibits serine/threonine kinases, reduces merozoite invasion
Inhibiting merozoite invasion through binding to sphingomyelin
Serine protease inhibitor

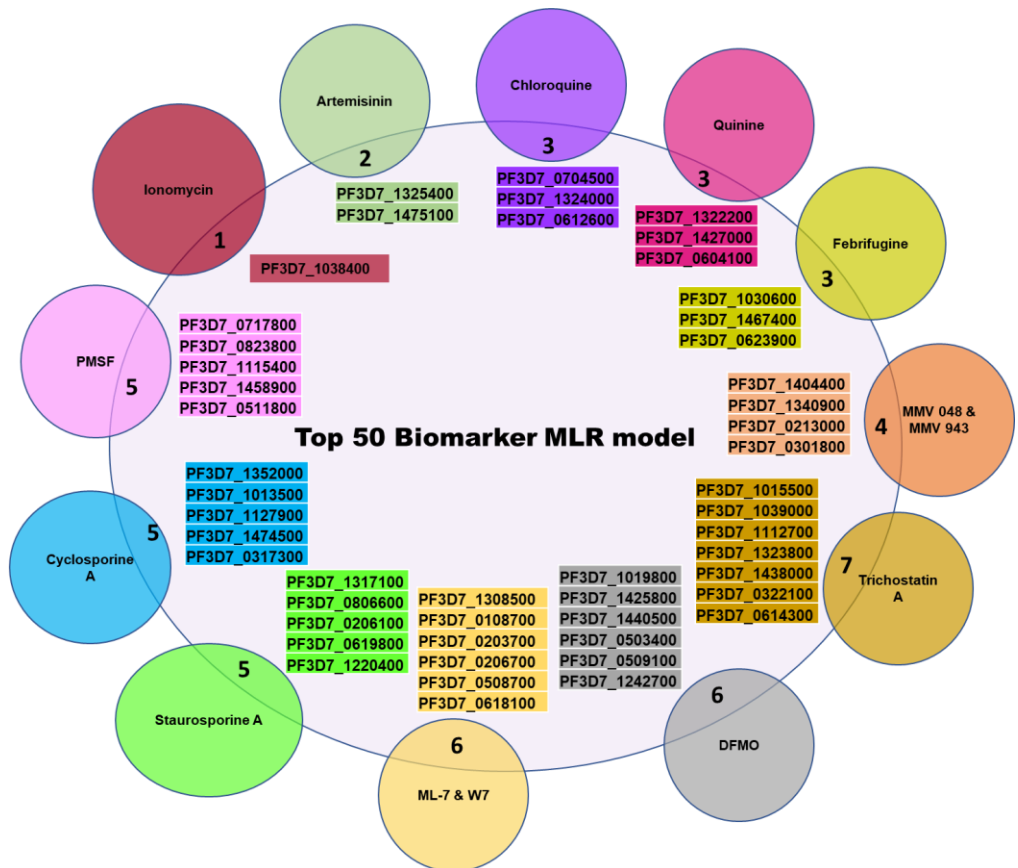


Figure 22: Compound origin of the top 50 biomarker genes used in the final MLR model for MoA stratification.

The contribution of each compound's biomarker genes to the overall 50 features that aids in MoA stratification of the MLR model. Genes are highlighted according to the compound the genes were identified from using our feature selection as well as which MoA they represent (Table 5).

Of the top 50 biomarker genes, 42% have putative protein products ascribed to them and 30% encode a novel unknown protein to which no function can be described (Table 6). This indicates that these genes are involved in important unknown cellular processes of the parasite which make them useful for biomarkers in MoA stratification. Of the top 50 biomarker genes that were annotated, 11 are involved in ATP or DNA binding, and 8 seem to be involved in translation and transcription processes within the parasite.

Table 6: Top 50 biomarker genes encoded product

Gene ID	Gene product description according to PlasmoDB	GO Functions
PF3D7_0108700	secreted ookinete protein, putative	N/A
PF3D7_0203700	protein MAK16, putative	N/A
PF3D7_0206700	adenylosuccinate lyase	N6-(1,2-dicarboxyethyl)AMP AMP-lyase (fumarate-forming) activity;catalytic activity
PF3D7_0508700	pre-mRNA-processing ATP-dependent RNA helicase PRP5, putative	ATP binding;nucleic acid binding
PF3D7_0618100	conserved <i>Plasmodium</i> protein, unknown function	N/A
PF3D7_1308500	conserved <i>Plasmodium</i> protein, unknown function	ATP binding;kinase activity;phosphotransferase activity, carboxyl group as acceptor
PF3D7_0322100	mRNA-capping enzyme subunit beta	polynucleotide 5'-phosphatase activity
PF3D7_0614300	major facilitator superfamily-related transporter, putative	N/A
PF3D7_1015500	nucleotidyltransferase, putative	RNA binding;nucleotidyltransferase activity;polynucleotide adenylyltransferase activity
PF3D7_1039000	serine/threonine protein kinase, FIKK family	ATP binding;protein kinase activity
PF3D7_1112700	conserved <i>Plasmodium</i> protein, unknown function	ATP binding;binding
PF3D7_1323800	vacuolar protein sorting-associated protein 52, putative	N/A
PF3D7_1438000	eukaryotic translation initiation factor eIF2A, putative	translation initiation factor activity
PF3D7_0623900	ribonuclease H2 subunit A, putative	RNA-DNA hybrid ribonuclease activity
PF3D7_1030600	tRNA N6-adenosine threonylcarbamoyltransferase	N/A
PF3D7_1467400	50S ribosomal protein L22, apicoplast, putative	structural constituent of ribosome
PF3D7_0206100	cysteine desulfuration protein SufE	N/A
PF3D7_0619800	conserved <i>Plasmodium</i> membrane protein, unknown function	protein binding
PF3D7_0806600	kinesin-like protein, putative	ATP binding;microtubule binding;microtubule motor activity
PF3D7_1220400	debranching enzyme-associated ribonuclease, putative	N/A
PF3D7_1317100	DNA replication licensing factor MCM4	ATP binding;DNA binding;DNA helicase activity
PF3D7_1325400	conserved <i>Plasmodium</i> protein, unknown function	ATP binding;actin binding;calmodulin binding;motor activity
PF3D7_1475100	conserved <i>Plasmodium</i> protein, unknown function	N/A
PF3D7_0503400	actin-depolymerizing factor 1	actin binding
PF3D7_0509100	structural maintenance of chromosomes protein 4, putative	ATP binding;protein binding
PF3D7_1019800	tRNA methyltransferase, putative	N/A
PF3D7_1242700	40S ribosomal protein S17, putative	structural constituent of ribosome
PF3D7_1425800	conserved <i>Plasmodium</i> protein, unknown function	nucleotide-binding
PF3D7_1440500	allantoicase, putative	allantoicase activity
PF3D7_0317300	conserved <i>Plasmodium</i> protein, unknown function	N/A
PF3D7_1013500	phosphoinositide-specific phospholipase C	phosphatidylinositol phospholipase C activity;phospholipid binding;phosphoric diester hydrolase activity;protein binding
PF3D7_1127900	conserved <i>Plasmodium</i> protein, unknown function	motor activity
PF3D7_1352000	GTP-binding protein, putative	GTP binding
PF3D7_1474500	splicing factor 3A subunit 1, putative	RNA binding
PF3D7_0612600	cytoplasmic tRNA 2-thiolation protein 1, putative	tRNA binding

PF3D7_0704500	serine/threonine protein kinase, putative	ATP binding;protein kinase activity
PF3D7_1324000	conserved <i>Plasmodium</i> protein, unknown function	asparagine synthase (glutamine-hydrolyzing) activity
PF3D7_0604100	AP2 domain transcription factor	N/A
PF3D7_1322200	conserved <i>Plasmodium</i> protein, unknown function	binding;microtubule binding
PF3D7_1427000	conserved <i>Plasmodium</i> protein, unknown function	ATP binding
PF3D7_0511800	inositol-3-phosphate synthase	inositol-3-phosphate synthase activity
PF3D7_0717800	conserved <i>Plasmodium</i> protein, unknown function	ATP binding
PF3D7_0823800	DnaJ protein, putative	DNA-directed DNA polymerase activity;nucleic acid binding;nucleotide binding
PF3D7_1115400	cysteine proteinase falcipain 3	cysteine-type peptidase activity
PF3D7_1458900	golgi apparatus membrane protein TVP23, putative	N/A
PF3D7_1038400	gametocyte-specific protein	N/A
PF3D7_0213000	conserved protein, unknown function	N/A
PF3D7_0301800	<i>Plasmodium</i> exported protein, unknown function	N/A
PF3D7_1340900	sodium-dependent phosphate transporter	inorganic phosphate transmembrane transporter activity
PF3D7_1404400	ribosomal protein L16, mitochondrial, putative	structural constituent of ribosome

* Source: PlasmoDB (www.plasmodb.org)

To evaluate the novelty of the top 50 biomarker genes associated to specific MoAs, they were compared to genes identified by Siwo *et al.* and Hu *et al.*, as associated with specific MoAs (Figure 23). No overlap was observed for the 50 biomarkers from our study compared to the Siwo dataset, and only 13% (14) of the biomarkers we identified associated with the same compound MoA in the Hu dataset. Hu *et al.*, however, had used a 3 FC to define DEGs, whereas we defined DEGs as those within the upper or lower 5th percentile of gene expression, allowing us to identify more DEGs.

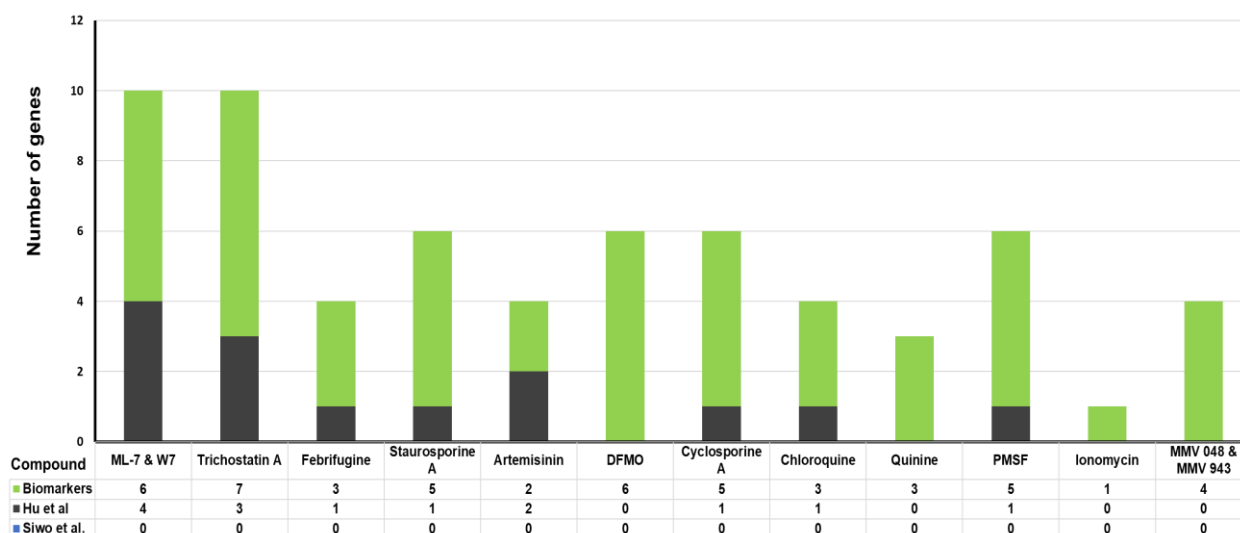


Figure 23: Novelty of top 50 rationally selected biomarker genes

The top 50 biomarkers (green) from the rationally selected biomarker minimodel were compared to DEGs associated with MoA as identified by the Hu *et al.* (dark grey) and Siwo *et al.* (blue) studies. Biomarker genes that were also identified for the same compound and MoA within these two studies are shown as stacked bars.

Chapter 4: Discussion

The progression of new antimalarial compounds is reliant on the ability to assign MoA to the compounds. Moreover, medicinal chemistry programs require an easy tracing of the MoA of a compound during H2L and LO campaigns, to correlate increased potency with the compounds' ability to still target the same drug target, or conversely, where the loss in potency is observed, to indicate if this is due to a change in MoA. Addressing this uncertainty will accelerate H2L optimisation by guiding and defining the chemical space for medicinal chemists. It is already known that *P. falciparum* has a unique just-in-time gene expression pattern when progressing through the intra-erythrocytic cycle [145]. Previous studies have revealed that the parasite's perturbed transcriptome resultant from compound treatment is highly reproducible and similar transcriptional responses are shared between antimalarials with similar chemical structure and MoA [93, 104]. The stringent control over the parasite's gene expression and the dysregulation of this expression due to compound treatments may be used to determine the MoA early in antimalarial drug discovery [146].

In this study, we investigated the use of ML and compound-induced transcriptional responses of the parasite to develop a ML model with robust accuracy in stratifying antiplasmodial compounds to their respective MoA. To accomplish this biomarker genes representative of compound MoA were needed. To our knowledge, however, there was no guideline on how to identify such genes, except by relying on algorithms for feature selection that may identify genes with no biological logic or linkage to our MoA stratification problem. Hence, we developed our own rational feature selection criteria using transcriptional response characteristics we deemed as important for good predictors of compound MoA. Using our feature selection criteria for DE we obtained 491 additional DEGs within the Hu dataset compared to what Hu *et al.* obtained using a 3-fold change cut-off limit [93]. From the rational feature selection, that we implemented to identify biomarkers representative of compound's MoA, we found that not all DEGs are pervasively expressed throughout a compound's treatment. Interestingly, the majority of the DEGs that are continually DE throughout a compound's treatment are unique to that compound treatment and can function as predictive features that represent a compound's MoA to help build a MoA stratification ML model.

The majority of the multiclassification algorithms investigated reached an average accuracy above 60% to correctly stratify compounds to their respective MoA. Most models trained on the 2463-gene database showed high variability in their accuracy of stratifying compounds to their MoA, indicating that these models are unstable for MoA stratification.

Interestingly, advanced ML algorithms employing ensemble classifiers or deep learning showed high variability within their K-fold cross validation accuracies when compared to the MLR algorithm, despite showing good performance for multiclassification problems utilizing gene expression in other cancer studies [147, 148]. One of the possible reasons for this could be that the number of hyperparameters investigated and the ranges we screened for the optimal architecture of the model was not sufficient. If the architecture of the model is not optimal, the results in MoA classification accuracy will be poor and unstable for different datasets with different perturbations.

Alternative explanations are associated with the limited features and samples within our dataset. Algorithms such as ANN and ensemble classifiers were adapted to handle 'big data' such as that in CMap which contains about 1.5 M expression profiles of over 5000 compounds treated on different cancer cell types [149]. Some algorithms are dependent on large labelled datasets for training to allow for generalization when exposed to new data in order to improve prediction accuracy as well as limit misclassification [150, 151]. In the case of our ensemble classifiers, which rely on bagging and boosting, this can lead to creating unstable classifiers since modifying or improving the model during training, through subsampling a small dataset, may not give a real representation of the data or the distribution thereof [152]. Similarly, with deep learning, such as ANN when training on small datasets can result in overfitting the network to examples in the training data, leading to the ANN performing poorly on new untrained data [153]. It is then no surprise why our GEP database, having high dimensionality as a result of the number of genes/features, but only 103 observations (time points containing 253689 data points) for training and testing, resulted in poor performance by algorithms which employed ensemble classifiers or deep learning. Not only this but using a small sample size and a large number of features, as is common with GEPs can lead to generating a classification model with faulty generalization and poor classification accuracy on untrained data [154]. By contrast, traditional ML algorithms such as MLR that rely on statistical techniques performed better on our small dataset. Preferentially, more *in silico* validations on the models would be needed but due to size constraints of our dataset, our concern was that repeated subsampling of our dataset for training and testing purposes would result in an inaccurate representation of the data. In fact, this was already observed within our random forest models, which uses subsampling, and resulted in poor generalization.

Although GEPs are useful in providing a global overview of how the parasite is affected by a compound, they contain high dimensionality within them with irrelevant and/or redundant features that limit the efficiency and generalization of ML algorithms [155]. By applying our

rational feature selection approach, we reduced the number of genes used as training features within our MoA stratification MLR model to ~7% of the genes within our 2463-gene database and obtained similar accuracy with reduced variability attributed to noisy features. This highlights that even with the best ML algorithm available to build our model, i.e. the MLR algorithm, if the training data used doesn't resemble the data it will be tested on or contains irrelevant features, the model performs poorly [156]. This is known as the 'garbage in, garbage out' ML principle, where if noisy input data is used to train a ML model this will influence the model's prediction accuracy [157, 158].

We also validated our rational feature selection approach by comparing our features to features identified by the MLR algorithm as important predictive features. Based on the accuracy measures from both the test data and 10-fold cross validation, our rational approach for feature selection selected genes that contributed more to our model's stability compared to the ML-inferred approach. When optimising for the minimal number of training features the inferred approach increased the model's variability and caused the model to become overfitted as the number of features used decreased. This indicates that genes which the MLR algorithm identifies as important may be noise or non-informative genes that are not suitable as predictive features for MoA model building. As such, it is more useful to first implement a rational approach for feature selection before using an inferred ML approach or alternative algorithms for feature selection. Since a ML model maps input data to a specific MoA class the stability of the model relies on the input data and training features.

Since this project aims to use this ML classification model in a medium-throughput manner early in drug discovery, the optimal minimum number of genes that still allow for robust prediction were identified. The results conclude that the top 50 rationally selected biomarker genes, ranked by their percent feature importance, was the optimal minimal number of genes that could be used to train our MoA stratification model. This top 50 biomarker genes aided in building a final MLR model with the most stability and highest estimated average accuracy ($76.6 \pm 6.4\%$) as well as showed good accuracy in predicting the MoA class of test data.

When comparing our final MoA stratification model to the small molecule-GO network that Siwo *et al.* used by correctly identifying the MoA of 72 % of antiparasitic compounds, our model obtained a similar accuracy. Their small molecule-GO network contained 31 compounds with a total of 16 different MoA compared to our model which was built on the GEPs of 20 compounds with 15 different MoA [104]. Although this small molecule-GO

network had similar accuracy to our ML model, it would not be feasible to use this network in a high or even medium-throughput manner in early antimalarial drug discovery. This is because their network requires the top 100 perturbed genes to be identified per compound treatment. Their method relies on a global view of the parasite's transcriptome to identify biological pathways through GO enrichment using these top 100 genes to identifying the MoA of compounds. Our final MoA stratification model, however, is able to obtain similar accuracy in identifying the MoA of compounds using only 50 biomarker genes and does not require the whole GEP to be analysed, making it more suitable for application in early antimalarial drug discovery.

Additionally, by using the top 100 compound induced genes to build their small molecule-GO network, as we have seen with both the Hu *et al.* and within our database, not all the DEGs caused by compound treatment is relevant to the compound's MoA or even unique. The inclusion of such DEGs, which may be resultant of general drug stress, is likely the reason why some of the compounds obtained a high false discovery rate within their small molecule-GO network. One advantage of our rational feature selection approach is that we were able to eliminate such genes from being incorporated into our model.

Interestingly, the most of the genes used as features within our final MoA model were novel and had been identified from compounds that target proteins affecting multiple cellular processes such as kinases and histone deacetylases. This would indicate that the biomarker genes pervasively affected by these compound treatments are genes that are involved in multiple cellular processes. From Table 6, it is highlighted that most of the top 50 biomarker genes used as features in our MoA model are in some way or another involved in ATP binding, transcription or translation while other biomarker genes have unknown functions.

Within our database, there are several histone deacetylase inhibitors present and this over-representation of a particular MoA may cause our model to be biased towards compounds with such a MoA. This can be overcome by including more chemically diverse compounds to increase the MoA diversity within the model. Nonetheless, at the beginning of this study during data acquisition, such data was lacking. Since we currently lack an abundance of GEPs for compounds with diverse MoA, we hope to in the future determine whether our MoA classification model has any class imbalances which affect the accuracy of stratifying certain MoAs and address this.

There are some limitations regarding our MoA stratification model. This MoA classification model might be biased towards slow-acting compounds as most of the time points fall within

10 h of treatment. One way to compensate this is to include additional time points that would capture responses of slow-acting compounds. However, one can also contend that compounds are constantly in a mobile cell environment and are constantly interacting with biological components and although the phenotypic response is delayed the transcriptional response may be immediate.

Lastly, the MoA classification ML model we built still requires to be validated *in vitro* before this model can be used in antimalarial drug discovery. Although this model will not define the MoA directly, it can be used as a steppingstone to help accelerate H2L optimisation by monitoring any change in MoA and guide combinational studies. Not only will such a model help in selecting compound candidates against asexuals early in drug discovery it will also reduce costs associated with H2L optimisation and MoA studies. This MoA classification model if validated *in vitro*, will be a very advantageous asset to the antimalarial community in driving antimalarial drug discovery.

References

1. Grmek MD: **-Malaria in the eastern Mediterranean in prehistory and antiquity.** *Parassitologia* 1994, **36**(1-2):1-6.
2. Bruce-Chwatt LJ: **History of malaria from prehistory to eradication.** In: *Malaria: Principles and Practice of Malariology.* Edited by Wernsdorfer WH, McGregor I. Edinburgh: Churchill Livingstone; 1988.
3. Hoffman SL, Subramanian GM, Collins FH, Venter JC: **Plasmodium, human and Anopheles genomics and malaria.** *Nature* 2002, **415**(6872):702-709.
4. Bartoloni A, Zammarchi L: **Clinical Aspects of Uncomplicated and Severe Malaria.** *Mediterranean Journal of Hematology and Infectious Diseases* 2012, **4**(1):e2012026.
5. Kwenti TE, Kwenti TDB, Njunda LA, Latz A, Tufon KA, Nkuo-Akenji T: **Identification of the Plasmodium species in clinical samples from children residing in five epidemiological strata of malaria in Cameroon.** *Tropical Medicine and Health* 2017, **45**(1):14.
6. Ponsford MJ, Medana IM, Prapansilp P, Hien TT, Lee SJ, Dondorp AM, Esiri MM, Day NP, White NJ, Turner GD: **Sequestration and microvascular congestion are associated with coma in human cerebral malaria.** *The Journal of infectious diseases* 2012, **205**(4):663-671.
7. Storm J, Craig AG: **Pathogenesis of cerebral malaria—inflammation and cytoadherence.** *Frontiers in Cellular and Infection Microbiology* 2014, **4**:100.
8. Bhatt S, Weiss DJ, Cameron E, Bisanzio D, Mappin B, Dalrymple U: **The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015.** *Nature* 2015, **526**.
9. Organization WH: **World malaria report 2019.** 2019.
10. WHO: **WHO | World Malaria Report 2017.** In: *WHO* World Health Organization; 2017.
11. Cheah PY, Parker M, Dondorp AM: **Development of drugs for severe malaria in children.** *Int Health* 2016, **8**.
12. Singer VL, Jones LJ, Yue ST, Haugland RP: **Characterization of PicoGreen reagent and development of a fluorescence-based solution assay for double-stranded DNA quantitation.** *Anal Biochem* 1997, **249**.
13. Who: **WHO | World Malaria Report 2016.** In: *WHO.* World Health Organization; 2016.
14. Schantz-Dunn J, Nour NM: **Malaria and Pregnancy: A Global Health Perspective.** *Reviews in Obstetrics and Gynecology* 2009, **2**(3):186-192.
15. Murphy SC, Breman JG: **Gaps in the childhood malaria burden in Africa: cerebral malaria, neurological sequelae, anemia, respiratory distress, hypoglycemia, and complications of pregnancy.** *The American journal of tropical medicine and hygiene* 2001, **64**(1-2 Suppl):57-67.
16. Tarning J: **Treatment of Malaria in Pregnancy.** *New England Journal of Medicine* 2016, **374**(10):981-982.
17. Cox F: **History of the discovery of the malaria parasites and their vectors.** *Parasit Vectors* 2010, **3**.
18. Josling GA, Llinas M: **Sexual development in Plasmodium parasites: knowing when it's time to commit.** *Nat Rev Micro* 2015, **13**(9):573-587.
19. Arnot DE, Gull K: **The Plasmodium cell-cycle: facts and questions.** *Annals of tropical medicine and parasitology* 1998, **92**(4):361-365.
20. Read M, Sherwin T, Holloway SP, Gull K, Hyde JE: **Microtubular organization visualized by immunofluorescence microscopy during erythrocytic schizogony in Plasmodium falciparum and investigation of post-translational modifications of parasite tubulin.** *Parasitology* 1993, **106** (Pt 3):223-232.
21. Gerald N, Mahajan B, Kumar S: **Mitosis in the Human Malaria Parasite Plasmodium falciparum.** *Eukaryotic Cell* 2011, **10**(4):474-482.

22. Striepen B, Jordan CN, Reiff S, van Dooren GG: **Building the perfect parasite: cell division in apicomplexa.** *PLoS pathogens* 2007, **3**(6):e78.
23. Bannister LH, Hopkins JM, Fowler RE, Krishna S, Mitchell GH: **A brief illustrated guide to the ultrastructure of *Plasmodium falciparum* asexual blood stages.** *Parasitology today (Personal ed)* 2000, **16**(10):427-433.
24. Margos G, Bannister LH, Dluzewski AR, Hopkins J, Williams IT, Mitchell GH: **Correlation of structural development and differential expression of invasion-related molecules in schizonts of *Plasmodium falciparum*.** *Parasitology* 2004, **129**(Pt 3):273-287.
25. Idro R, Marsh K, John CC, Newton CRJ: **Cerebral malaria: mechanisms of brain injury and strategies for improved neurocognitive outcome.** *Pediatric research* 2010, **68**(4):267-274.
26. Guttery DS, Holder AA, Tewari R: **Sexual Development in *Plasmodium*: Lessons from Functional Analyses.** *PLoS pathogens* 2012, **8**(1):e1002404.
27. Billker O, Dechamps S, Tewari R, Wenig G, Franke-Fayard B, Brinkmann V: **Calcium and a calcium-dependent protein kinase regulate gamete formation and mosquito transmission in a malaria parasite.** *Cell* 2004, **117**(4):503-514.
28. Billker O, Shaw MK, Margos G, Sinden RE: **The roles of temperature, pH and mosquito factors as triggers of male and female gametogenesis of *Plasmodium berghei* in vitro.** *Parasitology* 1997, **115**.
29. Eisele TL, D.; Steketee, RW.: **Protective efficacy of interventions for preventing malaria mortality in children in *Plasmodium falciparum* endemic areas.** . *International Journal of Epidemiology* 2010, **39**:88-101.
30. Moiroux N, Gomez MB, Pennetier C, Elanga E, Djenontin A, Chandre F, Djegbe I, Guis H, Corbel V: **Changes in anopheles funestus biting behavior following universal coverage of long-lasting insecticidal nets in benin.** *The Journal of infectious diseases* 2012, **206**.
31. Sokhna C, Ndiath MO, Rogier C: **The changes in mosquito vector behaviour and the emerging resistance to insecticides will challenge the decline of malaria.** *Clinical Microbiology and Infection* 2013, **19**(10):902-907.
32. Barreaux P, Barreaux AMG, Sternberg ED, Suh E, Waite JL, Whitehead SA, Thomas MB: **Priorities for Broadening the Malaria Vector Control Tool Kit.** *Trends in Parasitology* 2017, **33**(10):763-774.
33. Penny MA, Verity R, Bever CA, Sauboin C, Galactionova K, Flasche S: **Public health impact and cost-effectiveness of the RTS, S/AS01 malaria vaccine: a systematic comparison of predictions from four mathematical models.** *Lancet* 2015, **387**.
34. Mahmoudi S, Keshavarz H: **Efficacy of Phase 3 Trial of RTS, S/AS01 Malaria Vaccine: The Need for an Alternative Development Plan.** *Human vaccines & immunotherapeutics* 2017:0.
35. Agnandji ST, Vansadia P: **First results of phase 3 trial of RTS,S/AS01 malaria vaccine in African children.** *The New England journal of medicine* 2011, **365**(20):1863-1875.
36. Antony HA, Parija SC: **Antimalarial drug resistance: An overview.** *Tropical Parasitology* 2016, **6**(1):30-41.
37. University PFMHMSPHC, Council NR, Education DBSS, Population C, Migration RDF, Williams HA, Bloland PB: **Malaria Control During Mass Population Movements and Natural Disasters:** National Academies Press; 2002.
38. Meunier B, Robert A: **Heme as trigger and target for trioxane-containing antimalarial drugs.** *Accounts of chemical research* 2010, **43**(11):1444-1451.
39. Varela JN, Lammoglia Cobo MF, Pawar SV, Yadav VG: **Cheminformatic Analysis of Antimalarial Chemical Space Illuminates Therapeutic Mechanisms and Offers**

- Strategies for Therapy Development.** *Journal of Chemical Information and Modeling* 2017, **57**(9):2119-2131.
40. Bonnington CA, Phyto AP, Ashley EA, Imwong M, Sriprawat K, Parker DM, Proux S, White NJ, Nosten F: ***Plasmodium falciparum* Kelch 13 mutations and treatment response in patients in Hpa-Pun District, Northern Kayin State, Myanmar.** *Malaria Journal* 2017, **16**(1):480.
 41. Gaillard T, Madamet M, Pradines B: **Tetracyclines in malaria.** *Malaria Journal* 2015, **14**:445.
 42. de Pécoulas PE, Tahar R, Ouatas T, Mazabraud A, Basco LK: **Sequence variations in the *Plasmodium vivax* dihydrofolate reductase-thymidylate synthase gene and their relationship with pyrimethamine resistance.** *Molecular and biochemical parasitology* 1998, **92**(2):265-273.
 43. Cowman AF, Morry MJ, Biggs BA, Cross G, Foote SJ: **Amino acid changes linked to pyrimethamine resistance in the dihydrofolate reductase-thymidylate synthase gene of *Plasmodium falciparum*.** *Proceedings of the National Academy of Sciences* 1988, **85**(23):9109-9113.
 44. Brooks DR, Wang P, Read M, Watkins WM, Sims PF, Hyde JE: **Sequence variation of the hydroxymethyldihydropterin pyrophosphokinase: dihydropteroate synthase gene in lines of the human malaria parasite, *Plasmodium falciparum*, with differing resistance to sulfadoxine.** *European Journal of Biochemistry* 1994, **224**(2):397-405.
 45. Yaro A: **Mechanisms of sulfadoxine pyrimethamine resistance and health implication in *Plasmodium falciparum* malaria: A mini review.** *Annals of Tropical Medicine and Public Health* 2009, **2**(1):20-23.
 46. Wilson DW, Langer C, Goodman CD, McFadden GI, Beeson JG: **Defining the timing of action of antimalarial drugs against *Plasmodium falciparum*.** *Antimicrob Agents Chemother* 2013, **57**(3):1455-1467.
 47. Le Bras J, Durand R: **The mechanisms of resistance to antimalarial drugs in *Plasmodium falciparum*.** *Fundamental & Clinical Pharmacology* 2003, **17**(2):147-153.
 48. Korsinczky M, Chen N, Kotecka B, Saul A, Rieckmann K, Cheng Q: **Mutations in *Plasmodium falciparum* Cytochrome b That Are Associated with Atovaquone Resistance Are Located at a Putative Drug-Binding Site.** *Antimicrobial agents and chemotherapy* 2000, **44**(8):2100-2108.
 49. Price RN, Uhlemann A-C, Brockman A, McGready R, Ashley E, Phaipun L, Patel R, Laing K, Looareesuwan S, White NJ: **Mefloquine resistance in *Plasmodium falciparum* and increased pfmdr1 gene copy number.** *The Lancet* 2004, **364**(9432):438-447.
 50. Radfar A, Diez A, Bautista JM: **Chloroquine mediates specific proteome oxidative damage across the erythrocytic cycle of resistant *Plasmodium falciparum*.** *Free Radic Biol Med* 2008, **44**:2034-2042.
 51. Chinappi M, Via A, Marcatili P, Tramontano A: **On the Mechanism of Chloroquine Resistance in *Plasmodium falciparum*.** *PLoS ONE* 2010, **5**(11):e14064.
 52. Collins WE, Jeffery GM: **Primaquine resistance in *Plasmodium vivax*.** *The American journal of tropical medicine and hygiene* 1996, **55**(3):243-249.
 53. Mutabingwa TK: **Artemisinin-based combination therapies (ACTs): best hope for malaria treatment but inaccessible to the needy!** *Acta tropica* 2005, **95**(3):305-315.
 54. Verlinden BK, Louw AI, Birkholtz L-M: **Resisting resistance : is there a solution for malaria?** 2016.
 55. Menard D, Dondorp A: **Antimalarial Drug Resistance: A Threat to Malaria Elimination.** *Cold Spring Harbor perspectives in medicine* 2017, **7**(7).
 56. Burrows JN, Hooft van Huijsduijnen R, Möhrle JJ, Oeuvray C, Wells TNC: **Designing the next generation of medicines for malaria control and eradication.** *Malar J* 2013, **12**.

57. Burrows JN, Duparc S, Gutteridge WE, van Huijsduijnen RH, Kaszubska W, Macintyre F, Mazzuri S, Möhrle JJ, Wells TNC: **New developments in anti-malarial target candidate and product profiles.** *Malaria journal* 2017, **16**(1):26.
58. Wells TN, Alonso PL, Gutteridge WE: **New medicines to improve control and contribute to the eradication of malaria.** *Nat Rev Drug Discov* 2009, **8**.
59. Mott BT, Eastman RT, Guha R, Sherlach KS, Siriwardana A, Shinn P, McKnight C, Michael S, Lacerda-Queiroz N, Patel PR *et al*: **High-throughput matrix screening identifies synergistic and antagonistic antimalarial drug combinations.** *Scientific Reports* 2015, **5**:13891.
60. Bréhélin L, Florent I, Gascuel O, Maréchal É: **Assessing functional annotation transfers with inter-species conserved coexpression: application to *Plasmodium falciparum*.** *BMC genomics* 2010, **11**(1):35.
61. Gowthaman R, Sekhar D, Kalita MK, Gupta D: **A database for *Plasmodium falciparum* protein models.** *Bioinformatics* 2005, **1**(2):50-51.
62. Guy AJ, Irani V, Beeson JG, Webb B, Sali A, Richards JS, Ramsland PA: **Proteome-wide mapping of immune features onto *Plasmodium* protein three-dimensional structures.** *Scientific Reports* 2018, **8**:4355.
63. van Brummelen AC, Olszewski KL, Wilinski D, Llinas M, Louw AI, Birkholtz LM: **Co-inhibition of *Plasmodium falciparum* S-adenosylmethionine decarboxylase/ornithine decarboxylase reveals perturbation-specific compensatory mechanisms by transcriptome, proteome, and metabolome analyses.** *The Journal of biological chemistry* 2008, **284**(7):4635-4646.
64. Plouffe D, Brinker A, McNamara C, Henson K, Kato N, Kuhlen K, Nagle A, Adrián F, Matzen JT, Anderson P *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**.
65. Gamo F-J, Sanz LM, Vidal J, de Cozar C, Alvarez E, Lavandera J-L, Vanderwall DE, Green DVS, Kumar V, Hasan S *et al*: **Thousands of chemical starting points for antimalarial lead identification.** *Nature* 2010, **465**:305.
66. Chibale K: **Novel antimalarial targets and the antimalarial pipeline.** *International Journal of Infectious Diseases* 2014, **21**:17.
67. Murithi JM, Owen ES, Istvan ES, Lee MCS, Otilie S, Chibale K, Goldberg DE, Winzeler EA, Llinás M, Fidock DA *et al*: **Combining Stage Specificity and Metabolomic Profiling to Advance Antimalarial Drug Discovery.** *Cell Chemical Biology* 2019.
68. Lomenick B, Jung G, Wohlschlegel JA, Huang J: **Target identification using drug affinity responsive target stability (DARTS).** *Current protocols in chemical biology* 2011, **3**(4):163-180.
69. Tulloch LB, Menzies SK, Coron RP, Roberts MD, Florence GJ, Smith TK: **Direct and indirect approaches to identify drug modes of action.** *IUBMB life* 2018, **70**(1):9-22.
70. Le Roch KG, Johnson JR, Florens L, Zhou Y, Santrosyan A, Grainger M, Yan SF, Williamson KC, Holder AA, Carucci DJ *et al*: **Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle.** *Genome Res* 2004, **14**.
71. Foth BJ, Zhang N, Chaal BK, Sze SK, Preiser PR, Bozdech Z: **Quantitative Time-course Profiling of Parasite and Host Cell Proteins in the Human Malaria Parasite *Plasmodium falciparum*.** *Molecular & Cellular Proteomics : MCP* 2011, **10**(8):M110.006411.
72. Flannery EL, Fidock DA, Winzeler EA: **Using genetic methods to define the targets of compounds with antimalarial activity.** *Journal of medicinal chemistry* 2013, **56**(20):7761-7771.
73. Ioerger TR, O'Malley T, Liao R, Guinn KM, Hickey MJ, Mohaideen N, Murphy KC, Boshoff HI, Mizrahi V, Rubin EJ *et al*: **Identification of new drug targets and resistance mechanisms in *Mycobacterium tuberculosis*.** *PLoS One* 2013, **8**(9):e75245.

74. Slater AF: **Chloroquine: mechanism of drug action and resistance in *Plasmodium falciparum***. *Pharmacology & therapeutics* 1993, **57**(2-3):203-235.
75. Napolitano F, Sirci F, Carrella D, di Bernardo D: **Drug-set enrichment analysis: a novel tool to investigate drug mode of action**. *Bioinformatics* 2016, **32**(2):235-241.
76. Sams-Dodd F: **Target-based drug discovery: is something wrong?** *Drug discovery today* 2005, **10**(2):139-147.
77. Dellarco VL, Baetcke K: **A Risk Assessment Perspective: Application of Mode of Action and Human Relevance Frameworks to the Analysis of Rodent Tumor Data**. *Toxicological Sciences* 2005, **86**(1):1-3.
78. Swinney DC: **Chapter 18 - Molecular Mechanism of Action (MMoA) in Drug Discovery**. In: *Annual Reports in Medicinal Chemistry*. Edited by Macor JE, vol. 46: Academic Press; 2011: 301-317.
79. Vincent IM, Ehmann DE, Mills SD, Perros M, Barrett MP: **Untargeted Metabolomics To Ascertain Antibiotic Modes of Action**. *Antimicrobial Agents and Chemotherapy* 2016, **60**(4):2281-2291.
80. Creek DJ, Chua HH, Cobbold SA, Nijagal B, MacRae JI, Dickerman BK, Gilson PR, Ralph SA, McConville MJ: **Metabolomics-Based Screening of the Malaria Box Reveals both Novel and Established Mechanisms of Action**. *Antimicrob Agents Chemother* 2016, **60**(11):6650-6663.
81. Van Voorhis WC, Adams JH, Adelfio R, Ahyong V, Akabas MH, Alano P, Alday A, Resto YA, Alsibae A, Alzualde A: **Open source drug discovery with the malaria box compound collection for neglected diseases and beyond**. *PLoS pathogens* 2016, **12**(7).
82. Allman EL, Painter HJ, Samra J, Carrasquilla M, Llinás M: **Metabolomic profiling of the malaria box reveals antimalarial target pathways**. *Antimicrobial agents and chemotherapy* 2016, **60**(11):6635-6649.
83. Gulati S, Ekland Eric H, Ruggles Kelly V, Chan Robin B, Jayabalasingham B, Zhou B, Mantel P-Y, Lee Marcus CS, Spottiswoode N, Coburn-Flynn O *et al*: **Profiling the Essential Nature of Lipid Metabolism in Asexual Blood and Gametocyte Stages of *Plasmodium falciparum***. *Cell Host & Microbe* 2015, **18**(3):371-381.
84. Isik Z, Baldow C, Cannistraci CV, Schroeder M: **Drug target prioritization by perturbed gene expression and network information**. *Scientific Reports* 2015, **5**:17417.
85. Iwata M, Sawada R, Iwata H, Kotera M, Yamanishi Y: **Elucidating the modes of action for bioactive compounds in a cell-specific manner by large-scale chemically-induced transcriptomics**. *Scientific Reports* 2017, **7**:40164.
86. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *Proceedings of the National Academy of Sciences* 2005, **102**(43):15545.
87. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, Murino L, Tagliaferri R, Brunetti-Pierri N, Isacchi A *et al*: **Discovery of drug mode of action and drug repositioning from transcriptional responses**. *Proc Natl Acad Sci U S A* 2010, **107**(33):14621-14626.
88. Murima P, de Sessions PF, Lim V, Naim ANM, Bifani P, Boshoff HIM, Sambandamurthy VK, Dick T, Hibberd ML, Schreiber M *et al*: **Exploring the Mode of Action of Bioactive Compounds by Microfluidic Transcriptional Profiling in Mycobacteria**. *PLoS ONE* 2013, **8**(7):e69191.
89. Augen J: **Bioinformatics in the Post-genomic Era: Genome, Transcriptome, Proteome, and Information-based Medicine**: Addison-Wesley; 2005.
90. Simon R, Radmacher MD, Dobbin K, McShane LM: **Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification**. *JNCI: Journal of the National Cancer Institute* 2003, **95**(1):14-18.

91. Woo JH, Shimoni Y, Yang WS, Subramaniam P, Iyer A, Nicoletti P, Rodríguez Martínez M, López G, Mattioli M, Realubit R *et al*: **Elucidating Compound Mechanism of Action by Network Perturbation Analysis**. *Cell* 2015, **162**(2):441-451.
92. Birkholtz L, van Brummelen AC, Clark K, Niemand J, Maréchal E, Llinás M, Louw AI: **Exploring functional genomics for drug target and therapeutics discovery in Plasmodia**. *Acta tropica* 2008, **105**(2):113-123.
93. Hu G, Cabrera A, Kono M, Mok S, Chaal BK, Haase S, Engelberg K, Cheemadan S, Spielmann T, Preiser PR *et al*: **Transcriptional profiling of growth perturbations of the human malaria parasite *Plasmodium falciparum***. *Nature Biotechnology* 2009, **28**:91.
94. N. MI, Kosmas K, G. AL: **Leveraging systems biology approaches in clinical pharmacology**. *Biopharmaceutics & Drug Disposition* 2013, **34**(9):477-488.
95. Chicco D: **Ten quick tips for machine learning in computational biology**. *BioData Mining* 2017, **10**:35.
96. Libbrecht MW, Noble WS: **Machine learning applications in genetics and genomics**. *Nature reviews Genetics* 2015, **16**(6):321-332.
97. Pirooznia M, Yang JY, Yang MQ, Deng Y: **A comparative study of different machine learning methods on microarray gene expression data**. *BMC Genomics* 2008, **9 Suppl 1**:S13.
98. Campbell JB, Wynne RH: **Introduction to Remote Sensing**: Guilford Publications; 2011.
99. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M *et al*: **Applications of machine learning in drug discovery and development**. *Nature Reviews Drug Discovery* 2019.
100. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T: **The rise of deep learning in drug discovery**. *Drug discovery today* 2018, **23**(6):1241-1250.
101. Tan M, Özgül OF, Bardak B, Ekşioğlu I, Sabuncuoğlu S: **Drug response prediction by ensemble learning and drug-induced gene expression signatures**. *Genomics* 2018.
102. Scheeder C, Heigwer F, Boutros M: **Machine learning and image-based profiling in drug discovery**. *Current Opinion in Systems Biology* 2018, **10**:43-52.
103. KalantarMotamedi Y, Eastman RT, Guha R, Bender A: **A systematic and prospectively validated approach for identifying synergistic drug combinations against malaria**. *Malaria journal* 2018, **17**(1):160.
104. Siwo GH, Smith RS, Tan A, Button-Simons KA, Checkley LA, Ferdig MT: **An integrative analysis of small molecule transcriptional responses in the human malaria parasite *Plasmodium falciparum***. *BMC Genomics* 2015, **16**:1030.
105. Tarr SJ, Nisbet RER, Howe CJ: **Transcript-level responses of *Plasmodium falciparum* to thioestrepton**; 2011.
106. Gupta DK, Patra AT, Zhu L, Gupta AP, Bozdech Z: **DNA damage regulation and its role in drug-related phenotypes in the malaria parasites**. *Sci Rep* 2016, **6**:23603-23603.
107. Shaw PJ, Chaotheing S, Kaewprommal P, Piriyaopongsa J, Wongsombat C, Suwannakitti N, Koonyosying P, Uthaipibull C, Yuthavong Y, Kamchonwongpaisan S: ***Plasmodium* parasites mount an arrest response to dihydroartemisinin, as revealed by whole transcriptome shotgun sequencing (RNA-seq) and microarray study**. In: *BMC Genomics*. vol. 16; 2015: 830.
108. Mohd Abd Razak MR, Rain Abdullah N, Chomel R, Muhamad R, Ismail Z: **Effect of choline kinase inhibitor hexadecyltrimethylammonium bromide on *Plasmodium falciparum* gene expression**, vol. 45; 2014.
109. Guler JL, Freeman DL, Ahyong V, Patrapuvich R, White J, Gujjar R, Phillips MA, DeRisi J, Rathod PK: **Asexual populations of the human malaria parasite, *Plasmodium falciparum*, use a two-step genomic strategy to acquire accurate, beneficial DNA amplifications**. *PLoS Pathog* 2013, **9**(5):e1003375-e1003375.

110. Brunner R, Aissaoui H, Boss C, Bozdech Z, Brun R, Corminboeuf O, Delahaye S, Fischli C, Heidmann B, Kaiser M: **Identification of a new chemical class of antimalarials.** *The Journal of infectious diseases* 2012, **206**(5):735-743.
111. Andrews KT, Gupta AP, Tran TN, Fairlie DP, Gobert GN, Bozdech Z: **Comparative Gene Expression Profiling of *P. falciparum* Malaria Parasites Exposed to Three Different Histone Deacetylase Inhibitors.** *PLOS ONE* 2012, **7**(2):e31847.
112. Kritsiriwuthinan K, Chaotheing S, Shaw PJ, Wongsombat C, Chavalitshewinkoon-Petmitr P, Kamchonwongpaisan S: **Global gene expression profiling of *Plasmodium falciparum* in response to the anti-malarial drug pyronaridine.** *Malar J* 2011, **10**(1):242.
113. Becker JW, Mtwisha L, Crampton BG, Stoychev S, van Brummelen AC, Reeksting S, Louw AI, Birkholtz L-M, Mancama DT: ***Plasmodium falciparum* spermidine synthase inhibition results in unique perturbation-specific effects observed on transcript, protein and metabolite levels.** *BMC Genomics* 2010, **11**:235-235.
114. Becker JW, van der Merwe MM, van Brummelen AC, Pillay P, Crampton BG, Mmutlane EM, Parkinson C, van Heerden FR, Crouch NR, Smith PJ *et al*: ***In vitro* anti-plasmodial activity of *Dicoma anomala* subsp. *gerrardii* (Asteraceae): identification of its main active constituent, structure-activity relationship studies and gene expression profiling.** *Malar J* 2011, **10**:295-295.
115. Cheemadan S, Ramadoss R, Bozdech Z: **Role of calcium signaling in the transcriptional regulation of the apicoplast genome of *Plasmodium falciparum*.** *Biomed Res Int* 2014, **2014**:869401-869401.
116. Park FD, Sasik R, Reya T: **Chapter 4 - Microarrays: An Introduction and Guide to Their Use.** In: *Basic Science Methods for Clinical Researchers*. Edited by Jalali M, Saldanha FYL, Jalali M. Boston: Academic Press; 2017: 57-76.
117. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Research* 2015, **43**(7):e47.
118. Schratz P, Muenchow J, Iturritxa E, Richter J, Brenning A: **Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data.** *Ecological Modelling* 2019, **406**:109-120.
119. Cawley GC, Talbot NL, Girolami M: **Sparse multinomial logistic regression via bayesian l1 regularisation.** In: *Advances in neural information processing systems: 2007*. 209-216.
120. Erin LeDell, Gill N, Aiello S, Fu A, Candel A, Click C, Kraljevic T, Nykodym T, Aboyou P, Kurka M *et al*: **h2o: R Interface for 'H2O'.** 2019.
121. Moreira J, Carvalho A, Horvath T: **A General Introduction to Data Analytics:** Wiley; 2018.
122. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F: **e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien.** 2019.
123. Liaw A, Wiener M: **Classification and Regression by randomForest.** *R News* 2002, **2**(3):18-22.
124. Brownlee J: **Machine Learning Mastery With R: Get Started, Build Accurate Models and Work Through Projects Step-by-Step:** Machine Learning Mastery; 2016.
125. Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T *et al*: **xgboost: Extreme Gradient Boosting.** 2019.
126. Kuhn M, Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, Benesty M *et al*: **caret: Classification and Regression Training.** 2019.
127. Motoda H, Liu H: **Feature selection, extraction and construction.** *Communication of IICM (Institute of Information and Computing Machinery, Taiwan) Vol* 2002, **5**:67-72.
128. Coronado LM, Montealegre S, Chaverra Z, Mojica L, Espinosa C, Almanza A, Correa R, Stoute JA, Gittens RA, Spadafora C: **Blood Stage *Plasmodium falciparum***

- Exhibits Biological Responses to Direct Current Electric Fields.** *PloS one* 2016, **11(8):**e0161207-e0161207.
129. Karaman MW, Herrgard S, Treiber DK, Gallant P, Atteridge CE, Campbell BT, Chan KW, Ciceri P, Davis MI, Edeen PT *et al*: **A quantitative analysis of kinase inhibitor selectivity.** *Nat Biotechnol* 2008, **26(1):**127-132.
 130. Dluzewski AR, Garcia CR: **Inhibition of invasion and intraerythrocytic development of *Plasmodium falciparum* by kinase inhibitors.** *Experientia* 1996, **52(6):**621-623.
 131. Dynarowicz-Łątka P, Wnętrzak A, Makyla-Juzak K: **Cyclosporin A in Membrane Lipids Environment: Implications for Antimalarial Activity of the Drug--The Langmuir Monolayer Studies.** *The Journal of membrane biology* 2015, **248(6):**1021-1032.
 132. Fowler RE, Fookes RE, Lavin F, Bannister LH, Mitchell GH: **Microtubules in *Plasmodium falciparum* merozoites and their importance for invasion of erythrocytes.** *Parasitology* 1998, **117 (Pt 5):**425-433.
 133. Tan-No K, Shimoda M, Sugawara M, Nakagawasai O, Niijima F, Watanabe H, Furuta S, Sato T, Satoh S, Arai Y *et al*: **Cysteine protease inhibitors suppress the development of tolerance to morphine antinociception.** *Neuropeptides* 2008, **42(3):**239-244.
 134. Moura PA, Dame JB, Fidock DA: **Role of *Plasmodium falciparum* digestive vacuole plasmepsins in the specificity and antimalarial mode of action of cysteine and aspartic protease inhibitors.** *Antimicrob Agents Chemother* 2009, **53(12):**4968-4978.
 135. Meshnick SR: **Artemisinin: mechanisms of action, resistance and toxicity.** *Int J Parasitol* 2002, **32(13):**1655-1660.
 136. Keller TL, Zocco D, Sundrud MS, Hendrick M, Edenius M, Yum J, Kim Y-J, Lee H-K, Cortese JF, Wirth DF *et al*: **Halofuginone and other febrifugine derivatives inhibit prolyl-tRNA synthetase.** *Nature Chemical Biology* 2012, **8:**311.
 137. Petersen I, Eastman R, Lanzer M: **Drug-resistant malaria: Molecular mechanisms and implications for public health.** *FEBS Letters* 2011, **585(11):**1551-1562.
 138. Assaraf Y, Golenser J, Spira DT, Messer G, Bachrach U: **Cytostatic effect of DL-?-difluoromethylornithine against *Plasmodium falciparum* and its reversal by diamines and spermidine,** vol. 73; 1987.
 139. Brunschwig C, Lawrence N, Taylor D, Abay E, Njoroge M, Basarab GS, Le Manach C, Paquet T, Cabrera DG, Nchinda AT *et al*: **UCT943, a Next-Generation *Plasmodium falciparum* PI4K Inhibitor Preclinical Candidate for the Treatment of Malaria.** *Antimicrob Agents Chemother* 2018, **62(9):**e00012-00018.
 140. Darkin-Rattray SJ, Gurnett AM, Myers RW, Dulski PM, Crumley TM, Allocco JJ, Cannova C, Meinke PT, Colletti SL, Bednarek MA *et al*: **Apicidin: a novel antiprotozoal agent that inhibits parasite histone deacetylase.** *Proc Natl Acad Sci U S A* 1996, **93(23):**13143-13147.
 141. Colantuoni C, Henry G, Zeger S, Pevsne J: **Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts.** *Biotechniques* 2002, **32(6):**1316-1320.
 142. Ballman KV, Grill DE, Oberg AL, Therneau TM: **Faster cyclic loess: normalizing RNA arrays via linear models.** *Bioinformatics* 2004, **20(16):**2778-2786.
 143. Wilson DL, Buckley MJ, Helliwell CA, Wilson I: **New normalization methods for cDNA microarray data.** *Bioinformatics* 2003, **19(11):**1325-1332.
 144. Dudoit S, Yang YH, Callow MJ, Speed TP: **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.** *Statistica sinica* 2002:111-139.

145. Bozdech Z, Llinás M, Pulliam BL, Wong ED, Zhu J, DeRisi JL: **The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum***. *PLoS biology* 2003, **1**(1):e5.
146. de Azevedo MF, del Portillo HA: **Control of Gene Expression in *Plasmodium***. 2007.
147. Kim B-H, Yu K, Lee PCW: **Cancer classification of single-cell gene expression data by neural network**. *Bioinformatics* 2019.
148. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C *et al*: **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks**. *Nature Medicine* 2001, **7**(6):673-679.
149. Sirci F, Napolitano F, di Bernardo D: **Computational Drug Networks: a computational approach to elucidate drug mode of action and to facilitate drug repositioning for neurodegenerative diseases**. *Drug Discovery Today: Disease Models* 2016, **19**:11-17.
150. Goodfellow I, Bengio Y, Courville A: **Deep Learning**: MIT Press; 2016.
151. Mell LK, Tran PT, James B, Yu MDMHS, Zhang Q: **Principles of Clinical Cancer Research**: Springer Publishing Company; 2018.
152. Roli F, Kittler J: **Multiple Classifier Systems: Third International Workshop, MCS 2002, Cagliari, Italy, June 24-26, 2002. Proceedings**: Springer Berlin Heidelberg; 2003.
153. Bhagwat R, Abdollahnejad M, Moocarme M: **Applied Deep Learning with Keras: Solve complex real-life problems with the simplicity of Keras**: Packt Publishing; 2019.
154. Li Z, Xie W, Liu T: **Efficient feature selection and classification for microarray data**. *PLOS ONE* 2018, **13**(8):e0202167.
155. Bolón-Canedo V, Sánchez-Marroño N, Alonso-Betanzos A: **Feature Selection for High-Dimensional Data**: Springer International Publishing; 2015.
156. Sanders H, Saxe J: **Garbage in, garbage out: how purportedly great ML models can be screwed up by bad data**. *Technical report* 2017.
157. Beam AL, Kohane IS: **Big Data and Machine Learning in Health Care** *Big Data and Machine Learning in Health Care*. *JAMA* 2018, **319**(13):1317-1318.
158. Pu Y, Apel DB, Liu V, Mitri H: **Machine learning methods for rockburst prediction-state-of-the-art review**. *International Journal of Mining Science and Technology* 2019, **29**(4):565-570.
159. Gholami R, Fakhari N: **Chapter 27 - Support Vector Machine: Principles, Parameters, and Applications**. In: *Handbook of Neural Computation*. Edited by Samui P, Sekhar S, Balas VE: Academic Press; 2017: 515-535.
160. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet C, Ares M, Haussler D: **Support vector machine classification of microarray gene expression data**. *University of California, Santa Cruz, Technical Report UCSC-CRL-99-09* 1999.
161. Hepworth PJ, Nefedov AV, Muchnik IB, Morgan KL: **Broiler chickens can benefit from machine learning: support vector machine analysis of observational epidemiological data**. *Journal of The Royal Society Interface* 2012, **9**(73):1934.
162. Yahyaoui's A, Yahyaoui I, Yumuşak N: **13 - Machine Learning Techniques for Data Classification**. In: *Advances in Renewable Energies and Power Technologies*. Edited by Yahyaoui I: Elsevier; 2018: 441-450.
163. Frunza M-C: **Chapter 2I - Support Vector Machines**. In: *Solving Modern Crime in Financial Markets*. Edited by Frunza M-C: Academic Press; 2016: 205-215.
164. Wittek P: **7 - Supervised Learning and Support Vector Machines**. In: *Quantum Machine Learning*. Edited by Wittek P. Boston: Academic Press; 2014: 73-84.

165. Knerr S, Personnaz L, Dreyfus G: **Single-layer learning revisited: a stepwise procedure for building and training a neural network**. In: *Neurocomputing: 1990//1990; Berlin, Heidelberg*. Springer Berlin Heidelberg: 41-50.
166. Jakkula V: **Tutorial on support vector machine (svm)**. *School of EECS, Washington State University* 2006, **37**.
167. Nylen EL, Wallisch P: **Chapter 7 - Regression**. In: *Neural Data Science*. Edited by Nylen EL, Wallisch P: Academic Press; 2017: 189-221.
168. Hoffman JIE: **Chapter 33 - Logistic Regression**. In: *Basic Biostatistics for Medical and Biomedical Practitioners (Second Edition)*. Edited by Hoffman JIE: Academic Press; 2019: 581-589.
169. Fávero LP, Belfiore P: **Chapter 14 - Binary and Multinomial Logistic Regression Models**. In: *Data Science for Business and Decision Making*. Edited by Fávero LP, Belfiore P: Academic Press; 2019: 539-615.
170. Ouyed O, Allili MS: **Feature weighting for multinomial kernel logistic regression and application to action recognition**. *Neurocomputing* 2018, **275**:1752-1768.
171. **How Multinomial Logistic Regression Model Works In Machine Learning** [<https://dataaspirant.com/2017/03/14/multinomial-logistic-regression-model-works-machine-learning/>]
172. Genuer R, Poggi J-M, Tuleau-Malot C, Villa-Vialaneix N: **Random Forests for Big Data**. *Big Data Research* 2017, **9**:28-46.
173. Fratello M, Tagliaferri R: **Decision Trees and Random Forests**. In: *Encyclopedia of Bioinformatics and Computational Biology*. Edited by Ranganathan S, Gribskov M, Nakai K, Schönbach C. Oxford: Academic Press; 2019: 374-383.
174. Breiman L: **Random Forests**. *Machine Learning* 2001, **45**(1):5-32.
175. Cao D-S, Huang J-H, Liang Y-Z, Xu Q-S, Zhang L-X: **Tree-based ensemble methods and their applications in analytical chemistry**. *TrAC Trends in Analytical Chemistry* 2012, **40**:158-167.
176. Machado G, Recamonde-Mendoza M, Corbellini L: **What variables are important in predicting bovine viral diarrhea virus? A random forest approach**. *Veterinary Research* 2015, **46**:85.
177. Ferreira AJ, Figueiredo MAT: **Boosting Algorithms: A Review of Methods, Theory, and Applications**. In: *Ensemble Machine Learning: Methods and Applications*. Edited by Zhang C, Ma Y. Boston, MA: Springer US; 2012: 35-85.
178. Golden CE, Rothrock MJ, Mishra A: **Comparison between random forest and gradient boosting machine methods for predicting Listeria spp. prevalence in the environment of pastured poultry farms**. *Food Research International* 2019, **122**:47-55.
179. Touzani S, Granderson J, Fernandes S: **Gradient boosting machine for modeling the energy consumption of commercial buildings**. *Energy and Buildings* 2018, **158**:1533-1543.
180. Basheer IA, Hajmeer M: **Artificial neural networks: fundamentals, computing, design, and application**. *Journal of Microbiological Methods* 2000, **43**(1):3-31.
181. LeCun Y, Bengio Y, Hinton G: **Deep learning**. *Nature* 2015, **521**:436.
182. Shi G: **Chapter 3 - Artificial Neural Networks**. In: *Data Mining and Knowledge Discovery for Geoscientists*. Edited by Shi G. Oxford: Elsevier; 2014: 54-86.
183. Elbayoumi M, Ramli NA, Fitri Md Yusof NF: **Development and comparison of regression models and feedforward backpropagation neural network models to predict seasonal indoor PM_{2.5-10} and PM_{2.5} concentrations in naturally ventilated schools**. *Atmospheric Pollution Research* 2015, **6**(6):1013-1023.
184. Nyakeriga AM, Perlmann H, Hagstedt M, Berzins K, Troye-Blomberg M, Zhivotovsky B, Perlmann P, Grandien A: **Drug-induced death of the asexual blood stages of Plasmodium falciparum occurs without typical signs of apoptosis**. *Microbes and Infection* 2006, **8**(6):1560-1568.

185. Sandefur CI, Wooden JM, Quaye IK, Sirawaraporn W, Sibley CH: **Pyrimethamine-resistant dihydrofolate reductase enzymes of *Plasmodium falciparum* are not enzymatically compromised *in vitro*.** *Mol Biochem Parasitol* 2007, **154**(1):1-5.
186. Naaktgeboren N, Roobol K, Gubbens J, Voorma HO: **The mode of action of thiostrepton in the initiation of protein synthesis.** *European journal of biochemistry* 1976, **70**(1):39-47.

Appendix A

A.1 Machine learning theory

Hyperparameter testing

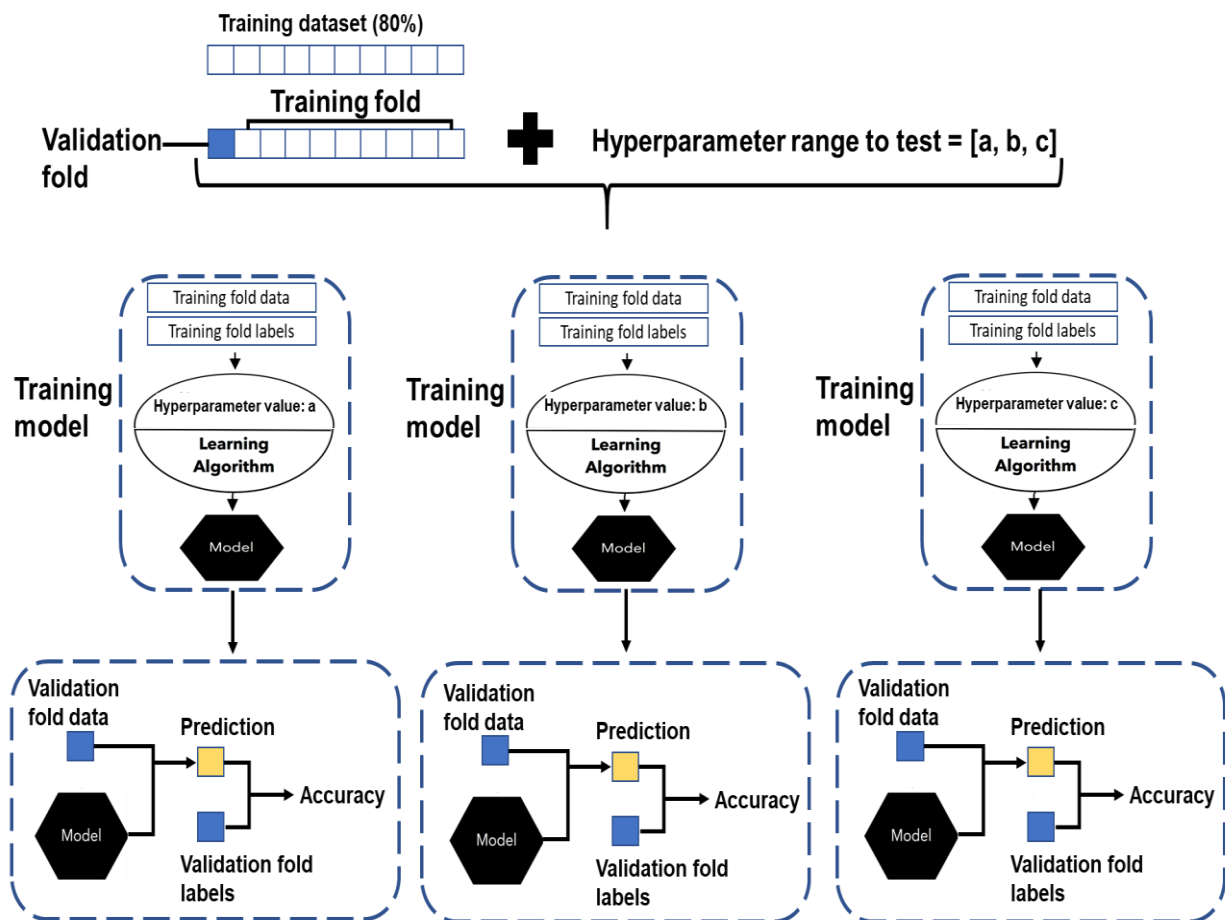


Figure 24: Principle of hyperparameter tuning.

A hyperparameter range or grid is given to the ML algorithm, whereby the ML algorithm trains on the training fold and builds a model/classifier using a hyperparameter value within the range or grid to define the model's architecture. For each hyperparameter value given a model is built and the performance of the model assessed. This can also be done to assess different combinations of different hyperparameter values. The hyperparameter values which gives the model the best accuracy is then identified.

A.1.1 Principle of multiclassification support vector machines

In machine learning, SVM is a supervised algorithm and can be separated into two categories, namely Support Vector Regression (SVR) and Support Vector Classification (SVC) [159]. For the purpose of this study which addresses a classification problem, only SVCs will be considered. SVMs were originally developed to solve binary problems by identifying the optimal separating linear hyperplane that can separate and differentiate between members and non-members of a given class in an abstract space as shown in Figure 20 [160].

As seen in Figure 25, there can be multiple hyperplanes that can separate the two classes, but not all hyperplanes will perform as well in classifying members of the circle class that are situated close to members of the square class. The SVM algorithm thus selects the optimal hyperplane which has the maximum margin i.e. distance from observations of each class [159].

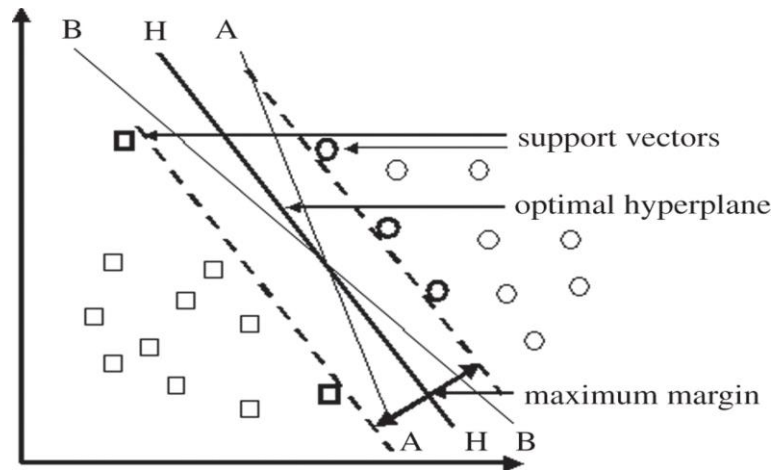


Figure 25: Principle of SVM classification

The squares and dots are spread onto a 2D feature space (not restricted to 2D) based on their respective properties. The SVM algorithm then produces multiple hyperplanes (A, H, B) to help separate the two classes (dots and squares). SVM then assesses each hyperplane in their ability to separate the two classes with the maximum distance between the two classes. Source: [161]

Not all observations, however, are linearly separable, e.g. Figure 26, and thus one solution SVM uses is to create a nonlinear feature space by applying a “kernel trick,” whereby the observations of the two classes can then be separated by the hyperplane [162, 163]. This kernel is a statistical mapping function which allows nonlinear data to be transformed into a higher dimension that will allow separation of different classes by a hyperplane [164].

Although SVC had been developed to address binary classification problems, real-life classification problems are multi-class and thus the algorithm has been adapted to address these problems as well [163].

The ‘one-against-one’ approach is an example such of a method developed by Knerr *et al.* which is used to implement a multi-class SVM, where several classifiers are combined [165]. Each classifier is binary and is built based on its’ training on two of the n classes, thereby resulting in $n(n-1)/2$ classifiers [163, 165]. For new data, each of these classifiers is applied and classification is made for each classifier resulting in a vector of individual classifications being created, e.g. AAB. From these individual classifications, the final class is identified by

majority vote, which in this case is class A. SVCs models using the polynomial and linear kernel performed the best compared to alternative kernels (see Appendix B).

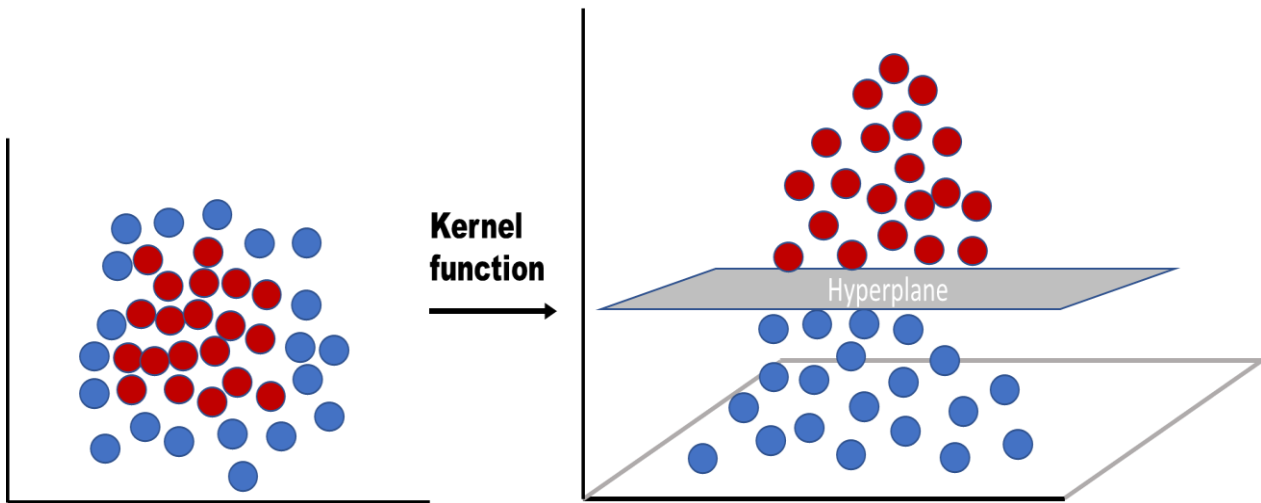


Figure 26: Support vector machine kernel function to separate nonlinear data

Support vector machines apply a kernel function to transform the data into a higher-dimensional space whereby the nonlinear data of two groups (red and blue) can be separated with a hyperplane whereas this would not have been accomplished linearly [166].

A.1.2 Principle of multinomial logistic regression

Machine learning extensively uses statistics and mathematic tools to help build a model from the training data it is given so it can predict or classify new data. Logistic regression (LR) is an example of such statistical tools used in ML and is similar to linear regression. With linear regression a linear relationship is assumed between the input variables and the output variable and a generalized linear model (GLM) is built that describes this linear relation [167]. However, in cases where the data is not linearly correlated and/or the output variable is discontinuous or categorical in nature, it is more beneficial to use logistic regression than linear regression [168]. Since our problem is categorical classification and we cannot assume that the GEP data are linearly correlated, LR is more useful.

The principle behind the LR algorithm is that it uses a sigmoid function to calculate the probability of whether an object belongs to a class or not [167, 168]. It does this by estimating the coefficients (parameter/beta weights) that link the input variables to the outcome variable using a maximum likelihood estimation approach [167]. Yet few real-world classification problems are binary but rather multi-class, such as ours.

The multinomial logistic regression approach was developed to address such multiclass problems, in which log odds of outcomes (logit values as shown in Figure 27) are modelled as

a linear combination of the input variables [169]. A logit value is the natural logarithmic probability of an event, such as belonging to a class. However, these values as seen in Figure 27, do not add up to one. Hence a softmax function is used to transform these values into probability distributions of a list of potential classes/ outcomes [170]. To help identify the predicted class the cross-entropy function is applied, which measures the distance of these probabilities for each class and the list of classes within the model and selects the one which has the shortest distance.

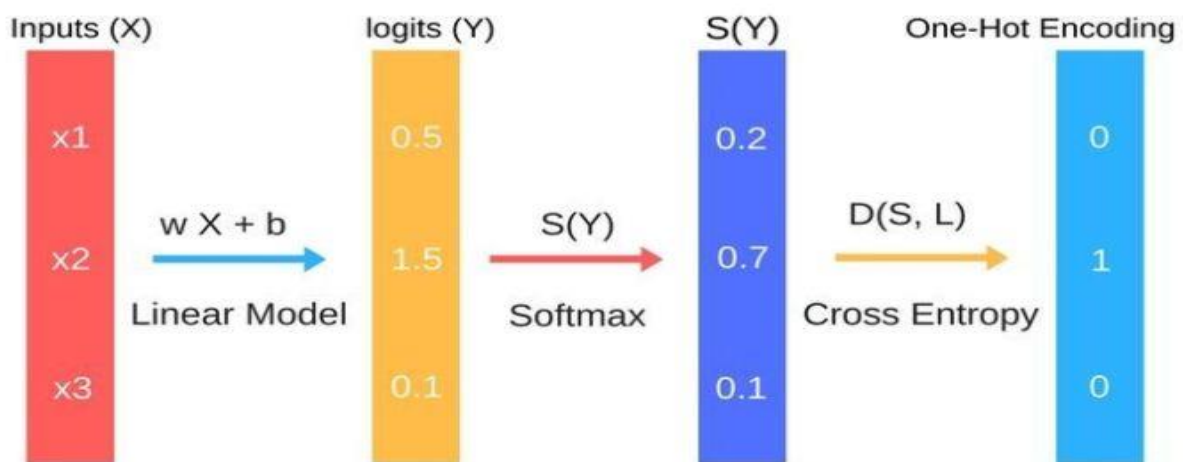


Figure 27: Multinomial logistic regression algorithm

The multinomial logistic algorithm analyses each input i.e. feature and builds a linear model for each input so that each input has its own weight (w) which is applied to a feature during the training phase of the algorithm. Each model will produce a logit score that with the help of a softmax function can convert the score into the probability of belonging to a class. Cross entropy calculates the distance between the probabilities for each class and selects the class with the shortest distance as the output. Source:[171]

A.1.3 Principle of random forest

Random forest is an ensemble classifier that employs decision trees and bootstrap aggregating [172]. Ensemble classifiers is a machine learning technique that combines several base models to build an optimal model with better performance [173]. Random forests (RF), for example, create multiple decision trees and the output from these decision trees helps it make a classification as shown in Figure 28 and is much more powerful and accurate than a single decision tree [174]. In principle of decision trees, the training dataset is repeatedly partitioned until the data can no longer be split. At the root of the decision tree, which contains the whole dataset, a feature is identified and a decision rule made that will employ a splitting criterion [173].

At this node the data will be partitioned into subsets, wherewith each subset a feature is again selected and a split criterion implemented until the data is no longer able to be split [174]. With

RF, multiple trees are made, but the algorithm does not select the data points or variables in each of the decision trees. Rather it randomly samples the data points and variables from each of these trees that it creates and combines the output and makes a vote on the class [175].

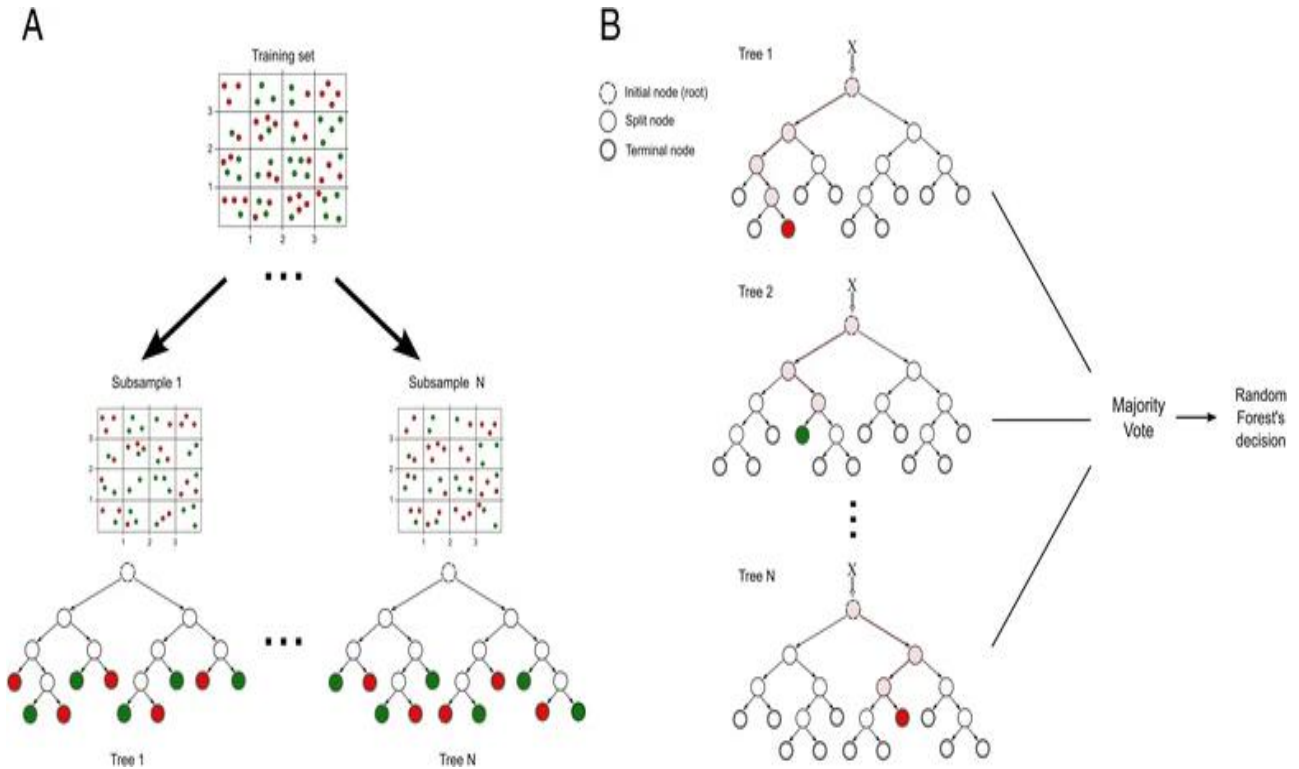


Figure 28: Random forest employing bootstrap aggregation and multiple decision trees.

A) From the training data, the algorithm applies bootstrap aggregating whereby subsets of the training data are used to build a decision tree. B) To predict the class of new input data, the algorithm takes the decision of all decision trees into account and uses a majority vote to identify the class (green and red), which in this case is the red class. For each new input data (X), the algorithm starts at the root of the tree and based on intrinsic properties of the data selects a branch to transverse down the tree until a leaf is reached whereby the class decision is made. This is done simultaneously for several decision trees. Source:[176]

A.1.4 Principle of gradient boosting machines

Another ensemble classifier, called gradient boosting machines (GBM), has gained wide interest in recent years in their ability to efficiently identify patterns for multiclassification problems. GBMs have been successfully applied in face detection, iris recognition, speech and multiclass text categorization [177].

Gradient boosting machines are similar random forest trees in that it also combines several simple base models to obtain a model with better accuracy, but how this is done differs. GBM builds an initial tree-based model and the next consecutive tree-model is built in such a way as to mitigate the faults of the previous tree-model [178]. This self-correction will continue until

an additive model which minimizes the error is found, or the number of trees specified is reached [179].

A.1.5 Principle of artificial neural networks

Artificial neural networks (ANN) has gained a lot of popularity in recent years as it has shown a remarkable ability to process information of biological systems that are prone to nonlinearity, noise, high parallelism and their ability to generalize [180].

ANN is a deep machine learning approach that is more advanced than the previously stated algorithms, in that it can gradually extract higher-level features from raw data using multi-layered processing units [181]. An ANN in its' simplest form contains an input layer, hidden layer, and output layer as illustrated in Figure 29. Within these layers are nodes that can be fully or partially connected to nodes in other layers [100].

The input layer contains nodes that represent the input variables of the model and these input variables are transformed using an activation function as they pass through to the hidden nodes. As these transformed variables are fed into the output nodes, output values are calculated that help in making a classification or prediction [182]. The number of output nodes corresponds to the number of classes or prediction variables.

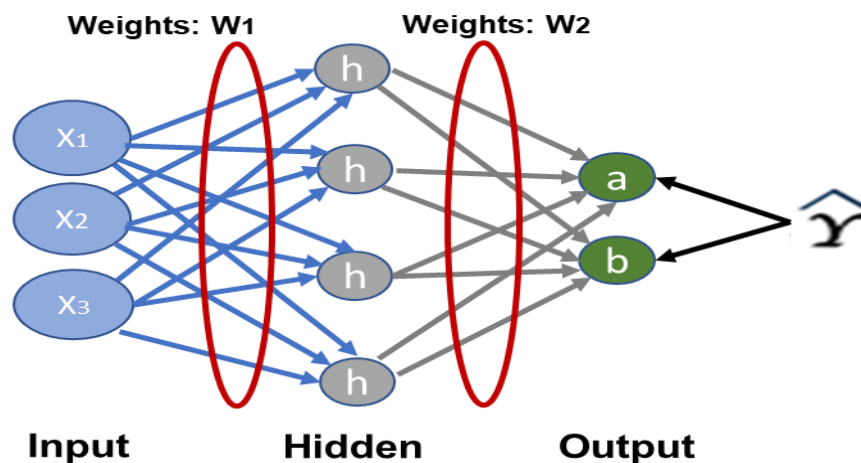


Figure 29: Simple artificial neural network

Neural networks have input nodes where data (X) are fed into a hidden layer where hidden nodes can assess information from the input nodes. This hidden layer can be extended to multiple layers and the hidden nodes (processing units) can also be increased. This hidden layer then connects to output nodes which can be increased to the number of classes or events. The hidden nodes give to each output node/class a probability of being true based on the input information fed into the input layer.

ANNs are powerful in that each node in the hidden layer functions as a processing unit that can consider all the variables or only a subset and analyse the relationships between these variables [100]. Not only this, but ANNs can also add weights to links connecting nodes as well as self-correct themselves during their training phase by using backpropagation. The ANNs do this self-correction by comparing the output values to the actual values and then adjust the weights on connecting links of nodes accordingly and reassesses the error between the output to actual values [183]. This is done repeatedly until the ANNs predictive and/or classification performance is optimized.

Table A.1: Optimal Hyperparameter tuning ranges for algorithms

Algorithm	Hyperparameter	Tuning range	Interval	Category	R tuning package
Support vector machine <ul style="list-style-type: none"> Polynomial kernel (P) Sigmoid kernel (S) Linear kernel (L) Radial kernel (R) 	Gamma		P= (0, 0.1, 0.3, 0.5, 1, 2, 4, 8, 10) S= (0, 0.1, 0.3, 0.5, 1, 2, 4, 8, 10) L= (0, 0.1, 0.3, 0.5, 1, 2, 4, 8, 10) R= (0.5,1,2)		e1071
	Degrees		P= (1, 2, 3, 4, 5, 6) L= (1, 2, 3, 4, 5, 6)		
	Cost		P= $10^{-3}:10^{10}$ S= $10^{-3}:10^{10}$ L= $10^{-3}:10^{10}$ R= $10^{-1}:10^2$		
Multinomial logistic regression	N/A	-	-	-	-
Random Forest (RandomForest)	Number of trees		1,10,100, 500,1000, 5000		e1071
	Mtries		6, 10, 20		
Random Forest (h2o package)	ntrees		100,250, 500,1000, 5000		h2o
	Mtry		1,5,10,15,20		
	Max depth		2,3,4,5,6		
Gradient boosting machine (h2o)	Number of trees	100-4000	100,200,300, 400,500, 1000,4000	-	h2o,
	col_sample_rate	0.3-1	0.3, 0.7, 1.0	-	
	max_depth	4-20	4,6,8,12, 16, 20	-	
Gradient boosting machine (Xgboost)	Col sample rate	0.1:1			caret
	Max depth		2, 3, 4, 5, 6		
	Subsample	0.1:1			
	nrounds		50, 100, 150		
	Eta		0.025, 0.05, 0.1, 0.3		
Artificial neural network	Activation function	-	-	Rectifier, RectifierWithDropout Maxout, MaxoutWithDropout	h2o
	Hidden drop out ratio	0-0.3	(0,0), (0.15,0.15), (0.3,0.3)	-	
	Input drop out ratio	0-0.3	0, 0.15, 0.3	-	
	L1 and L2 regularization	0-0.1	0,0.00001, 0.0001, 0.001, 0.01, 0.1	-	
	Adaptive rate	0.005-0.02	0.005, 0.01, 0.015, 0.02	-	
	Loss function	-	-	Automatic, CrossEntropy, Quadratic, Huber, Absolute, Quantile	

Table A.2: Accepted and rejected datasets

compound	Mode of action	Reference	Dataset	GEO accession	Time points per treatment	Gene coverage after pre-processing	Accepted
W7	Calcium/calmodulin-dependent protein kinase inhibitor	[93, 110]	Hu <i>et al.</i> , 2009	GSE19468	4-5 time points	3705/5400 (69%)	√
ML-7		[93, 128]					√
Staurosporine	Inhibits serine/threonine kinases, reduces merozoite invasion	[129, 130]					√
Cyclosporin A	It has a strong affinity to sphingomyelin in membrane environment like parasitized erythrocytes membranes, thus aids in inhibiting merozoite invasion. Also believed to be a calcineurin pathway inhibitor.	[93, 131]					√
Colchicine	Microtubule is the target, inhibits merozoite invasion	[132]					√
PMSF	Serine protease inhibitor	[133]					√
Leupeptin	A cysteine, serine, and threonine peptidase inhibitor which affects haemoglobin degradation	[134]					√
Artemisinin	Partially understood but hypothesized to be involved in producing carbon-centered free radicals that in turn alkylate heme and proteins	[135]					√
Chloroquine	Inhibits the heme polymerase enzyme	[74]					√
Febrifugine	Targets <i>P. falciparum</i> prolyl-tRNA synthetase activity	[136]					√
Quinine	Partially understood but accumulate in the parasite's digestive vacuole (DV) and may inhibit the detoxification of heme	[137]	√				
DFMO	Inhibits ornithine decarboxylase causing parasite arrest	[138]	van Brummelen <i>et al.</i> , 2008	GSE13578	3 time points with replicates	4050/5400 (75%)	√
MMV 048 and MMV 943	Inhibits <i>Plasmodium</i> phosphatidylinositol 4-kinase (PI4K)	[139]	Connacher <i>et al.</i> , 2016	GSE100692	2 time points each	4971/5400 (92%)	√
ACT-213615	Artemisinin derivative that has an unknown MoA which is different from other antimalarials based different transcriptional responses to that of the Hu <i>et al.</i> dataset	[110]	Brunner <i>et al.</i> , 2012	GSE39485	5 time points	4857/5400 (90%)	√
Ionomycin	Increases cytoplasmic calcium concentrations	[115]	Cheemadan <i>et al.</i> , 2014	GSE33869	5 time points	4495 /5400 (83%)	√
Trichostatin A (TSA), Suberoylanilide hydroxamic acid, 2-aminosuberic acid derivative, Apicidin	Histone deacetylase (HDAC) inhibitors that perturb the transcriptome	[111] [93] [93, 140]	Hu <i>et al.</i> , 2009	GSE19468	4-5 time points	3705/5400 (69%)	√
			Andrews <i>et al.</i> , 2012	GSE25642	1 time point	4364/5400 (80%)	
Cyclohexamine *	Inhibition of <i>P. falciparum</i> spermidine synthase causing perturbations in transcript, protein and metabolite levels	[113]	Becker <i>et al.</i> , 2010	GSE18075	3 time points	2595 (48%)	X
Methyl methanesulphonate*	Alkylates DNA bases and causes strand breaks	[106]	Gupta <i>et al.</i> 2016	GSE72580	1 time point	Not further investigated due to lack of multiple time points	X
Etoposide *	A topoisomerase II inhibitor	[184]					X
Pyrimethamine *	Competitive inhibitor of dihydrofolate reductase	[185]					X
Thiostrepton	Inhibits plastid protein synthesis through binding to apicoplast ribosomes	[186]	Tarr <i>et al.</i> 2011	GSE28701	1 time point		X

*Note: Compounds not shown from this table were filtered out due to the criteria in Table 4: Methods section regarding controls, concentration and parasite strains used or lack of time points and unknown MoA

Appendix B

Table B.1: Optimal hyperparameters identified from hyperparameter tuning

Algorithm	Hyperparameter	Optimal Hyperparameter for biomarkers	Optimal Hyperparameter for database	Logloss	Classification error	Out-of-bag error	Accuracy	R tuning package
Support vector machine <ul style="list-style-type: none"> Polynomial kernel (P) Sigmoid kernel (S) Linear kernel (L) Radial kernel (R) 	Gamma	P=0.1 S=0.1 L=0 R=0	P=0.1 S=0.1 L=0 R=0	N/A	B: P= 0.14 S= 0.5 L= 0.15 R= 0.8 D: P= 0.25 S= 0.7 L= 0.17 R= 0.8	N/A	N/A	e1071
	Degrees	P=1 L=1	P=1 L=1					
	Cost	P=1 S=1000 L=0.1 R=0.001	P=0.1 S=0.1 L=0.01 R=0.001					
Multinomial logistic regression	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
RandomForest	Number of trees	460	4000	N/A	B = 0.26 D = 0.29	N/A	N/A	e1071, RandomForest
	Mtries	6	6	N/A	N/A	B = 29.27% D = 23.17%	N/A	
Random Forest (h2o)	ntrees	500	1000	B = 0.957 D = 0.995	N/A	N/A	N/A	h2o
	Mtry	6	6					
	Max depth	20	20					
Xgboost	Col sample rate	0.6	0.6	N/A	N/A	N/A	B = 78.87% D = 77.24%	caret
	Max depth	1	2					
	Subsample	0.75	0.75					
	Nrounds	50	50					
	Min child weight	1	1					
Eta	0.4	0.4						
Gradient Boosting Machine	col_sample_rate	0.3	0.3	B = 2.30x 10 ⁻⁸ D = 1.15x 10 ⁻⁶	N/A	N/A	N/A	h2o
	max_depth	6	4					
	Ntrees	500	100					
Artificial neural network	Activation function	MaxoutWithDropout	MaxoutWithDropout	B = 0.001 D = 0.001	N/A	N/A	N/A	h2o
	Hidden drop out ratio	0.15	0.3					
	Input drop out ratio	0.3	0.3					
	L1 regularization	1.0x 10 ⁻⁵	0.01					
	L2 regularization	0	0.001					
	Adaptive rate	false	false					
	Loss function	Automatic	Automatic					

Note: D= database model, B= biomarker model

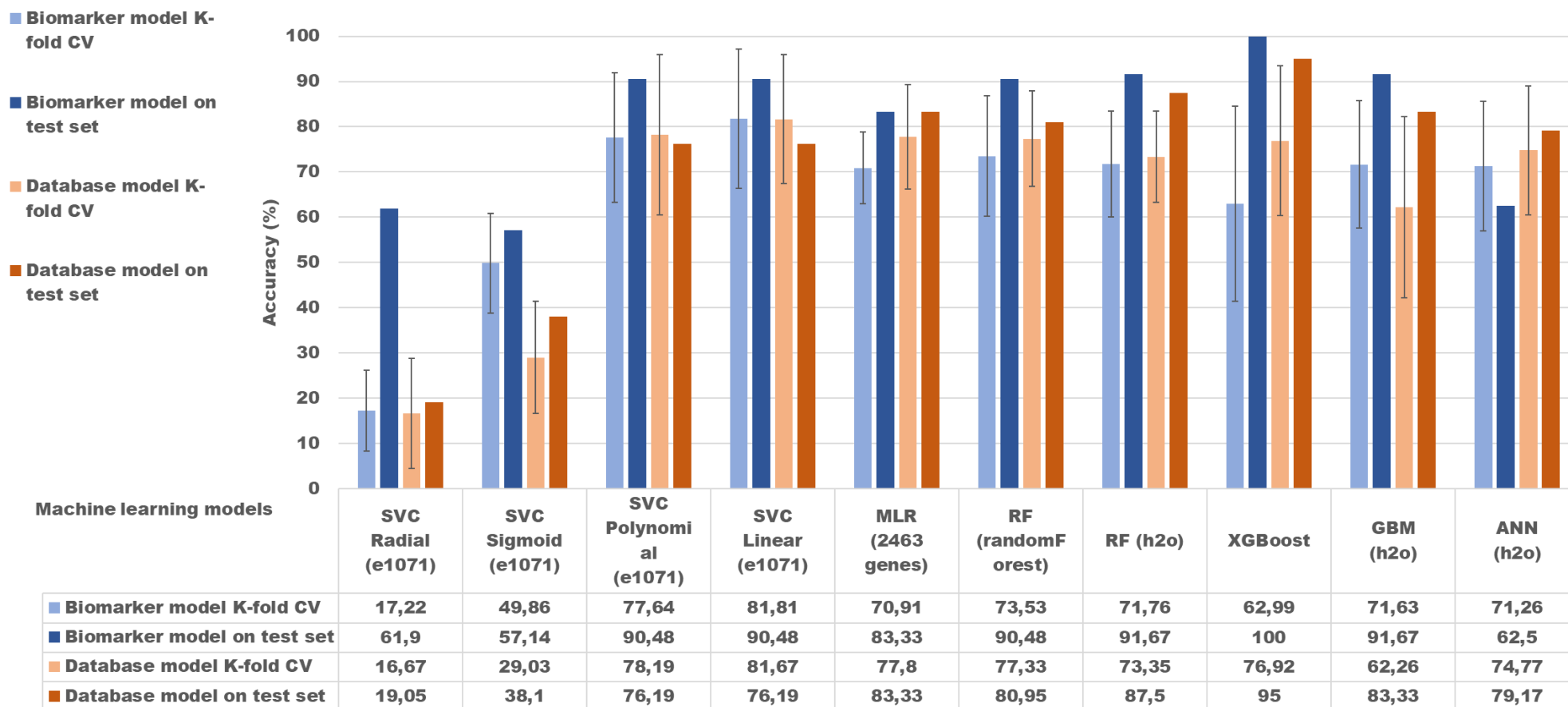


Figure 30: Performance of algorithms investigated for MoA classification built using either biomarker genes or all genes in the database.

The different algorithms investigated are shown on the x-axis and each consists of a classifier built on either the 175 biomarker genes (blue) or the 2463 genes within the database (orange). K-fold cross-validation is shown in a lighter colour whereas performance on the test set is shown in a darker colour.

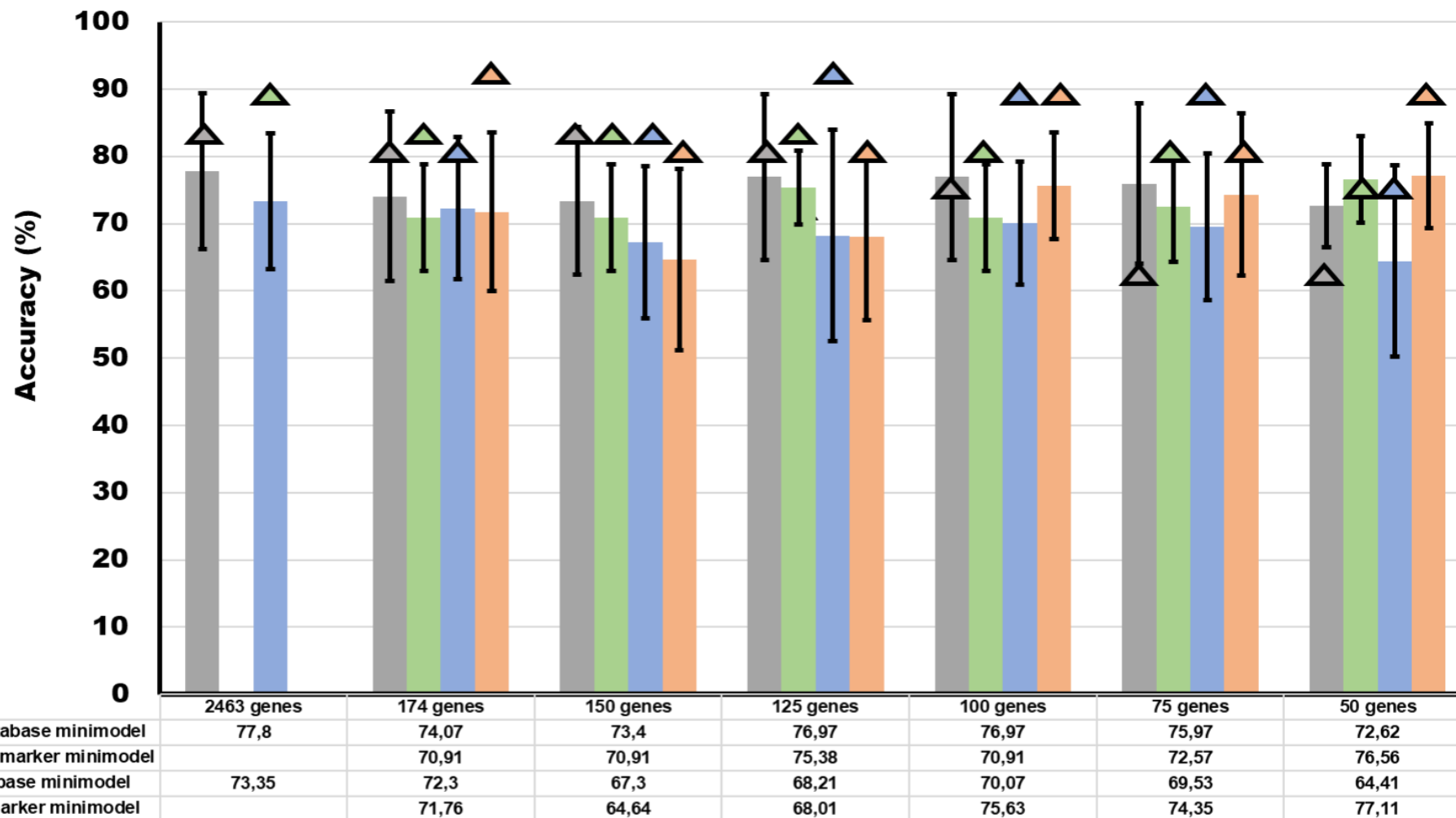


Figure 31: Performance in MoA accuracy of minimodels made from biomarker or database genes using random forest and multinomial logistic regression algorithms (h2o R package).

Bars indicate the accuracy obtained from 10-fold cross-validation with error bars indicating the standard deviation/ variability of the classifier. Triangles indicate the classifier's accuracy in classifying the test set. On the x-axis, the number of top genes used to build the classifier is indicated for each minimodel. These top genes are identified by the algorithm as genes that had importance in stratification decisions.

Table B. 2: Top 200 genes ranked according to importance in MoA stratification from the 2463-gene database

Rank	Feature	Relative importance	Scaled importance	Percentage
1	PF10_0374	0,509002	1	0,011241
2	PF10_0215	0,422073	0,829218	0,009321
3	PFF0875w	0,390202	0,766603	0,008617
4	PF14_0512	0,349927	0,687478	0,007728
5	PFL1240c	0,341779	0,67147	0,007548
6	PF14_0642	0,332423	0,653088	0,007341
7	PFF1580c	0,319261	0,627229	0,007051
8	PF10_0321	0,318752	0,626229	0,007039
9	PFL1125w	0,301269	0,591882	0,006653
10	PFC0090w	0,299878	0,589149	0,006623
11	PF13_0081	0,29466	0,578897	0,006507
12	PFD0625c	0,264675	0,519988	0,005845
13	PFB0530c	0,254792	0,500572	0,005627
14	PF07_0049	0,251232	0,493577	0,005548
15	PFI1770w	0,227414	0,446784	0,005022
16	PFB0655c	0,224483	0,441026	0,004958
17	PF14_0598	0,217328	0,426968	0,0048
18	PFC0960c	0,211083	0,4147	0,004662
19	PFB0161c	0,20664	0,405971	0,004564
20	PFL2235w	0,205782	0,404286	0,004545
21	PFC0885c	0,200706	0,394312	0,004432
22	PF10_0142	0,199621	0,392182	0,004409
23	MAL7P1.137	0,194386	0,381896	0,004293
24	PFF0610c	0,193942	0,381025	0,004283
25	MAL13P1.301	0,193498	0,380151	0,004273
26	PF11_0276	0,190392	0,374049	0,004205
27	PFF0940c	0,189277	0,371859	0,00418
28	PF11_0527	0,18066	0,35493	0,00399
29	PFF1315w	0,176455	0,346668	0,003897
30	PF14_0507	0,174924	0,343661	0,003863
31	PF14_0139	0,174244	0,342325	0,003848
32	MAL8P1.55	0,172921	0,339725	0,003819
33	PFL1815c	0,170604	0,335174	0,003768
34	PFL0175c	0,169753	0,333502	0,003749
35	PF14_0062	0,15999	0,314321	0,003533
36	PFF1150w	0,158806	0,311995	0,003507
37	PF11_0471	0,158472	0,311339	0,0035
38	PFF0200c	0,158002	0,310416	0,003489
39	PF14_0238	0,154881	0,304284	0,00342
40	PF13_0107	0,154192	0,30293	0,003405
41	PF11_0354	0,153966	0,302487	0,0034
42	PFL0585w	0,150614	0,295901	0,003326
43	PF11_0396	0,149923	0,294543	0,003311

44	PFE1320w	0,145711	0,286268	0,003218
45	PF14_0736	0,144421	0,283733	0,003189
46	PFE0830c	0,138169	0,271452	0,003051
47	PFI1020c	0,13735	0,269842	0,003033
48	MAL7P1.18	0,137081	0,269314	0,003027
49	PFB0745w	0,134038	0,263336	0,00296
50	PFI1580c	0,133592	0,262459	0,00295
51	PF14_0719	0,133422	0,262125	0,002947
52	PF14_0013	0,130965	0,257298	0,002892
53	PF13_0120	0,129367	0,254158	0,002857
54	PF11_0146	0,128553	0,252559	0,002839
55	PF08_0054	0,126241	0,248017	0,002788
56	MAL7P1.122	0,12552	0,2466	0,002772
57	PF11_0141	0,125036	0,24565	0,002761
58	PF07_0125	0,12424	0,244085	0,002744
59	PF11_0184	0,123747	0,243118	0,002733
60	PFI0915w	0,123046	0,24174	0,002717
61	PFE1520c	0,122804	0,241264	0,002712
62	PF13_0131	0,122659	0,24098	0,002709
63	PF08_0057	0,12179	0,239272	0,00269
64	PFF1370w	0,119804	0,235371	0,002646
65	PF14_0519	0,118918	0,23363	0,002626
66	PF14_0041	0,11816	0,23214	0,002609
67	PF11_0170	0,117714	0,231264	0,0026
68	PF14_0133	0,11478	0,2255	0,002535
69	PF10_0060	0,113209	0,222415	0,0025
70	MAL13P1.82	0,113191	0,222379	0,0025
71	PFI1175c	0,111241	0,218547	0,002457
72	PFI1410c	0,106307	0,208854	0,002348
73	PFD1050w	0,10289	0,202141	0,002272
74	PFE0285c	0,100142	0,196742	0,002212
75	PF11_0108	0,099931	0,196327	0,002207
76	PF14_0021	0,099	0,194499	0,002186
77	PF11_0210	0,098982	0,194463	0,002186
78	MAL8P1.4	0,097342	0,191241	0,00215
79	PF10_0309	0,09614	0,18888	0,002123
80	PF11_0129	0,09503	0,186699	0,002099
81	PFD0400w	0,094979	0,186598	0,002098
82	MAL13P1.127	0,091082	0,178943	0,002011
83	PF14_0644	0,090486	0,177771	0,001998
84	PF14_0454	0,085098	0,167186	0,001879
85	PF10_0182	0,083011	0,163085	0,001833
86	PFF0625w	0,081541	0,160198	0,001801
87	PF10_0203	0,081296	0,159717	0,001795
88	MAL8P1.65	0,07922	0,155638	0,00175
89	PF14_0117	0,079134	0,15547	0,001748
90	PF13_0166	0,078842	0,154895	0,001741

91	PFL2230c	0,078315	0,15386	0,00173
92	PFC0925w	0,07694	0,151159	0,001699
93	PFC0455w	0,073966	0,145316	0,001633
94	PF14_0249	0,073801	0,144991	0,00163
95	PFE1000c	0,073082	0,143579	0,001614
96	PF14_0461	0,072939	0,143298	0,001611
97	PFE0585c	0,07234	0,142121	0,001598
98	PF13_0073	0,0723	0,142044	0,001597
99	PFE0810c	0,072288	0,142019	0,001596
100	PFC0390w	0,071707	0,140878	0,001584
101	PF14_0123	0,07057	0,138645	0,001559
102	PF13_0036	0,07004	0,137602	0,001547
103	MAL13P1.183	0,069778	0,137089	0,001541
104	PF14_0368	0,068717	0,135004	0,001518
105	PFI0170w	0,067976	0,133547	0,001501
106	PFE0400w	0,067281	0,132183	0,001486
107	PFD0830w	0,067029	0,131687	0,00148
108	PF14_0380	0,066332	0,130318	0,001465
109	PF08_0068	0,065564	0,128809	0,001448
110	PF11_0321	0,065539	0,12876	0,001447
111	PF13_0137	0,065437	0,12856	0,001445
112	MAL13P1.229	0,065102	0,1279	0,001438
113	MAL13P1.137	0,064722	0,127155	0,001429
114	PFL1335w	0,062691	0,123164	0,001384
115	PF14_0346	0,062526	0,12284	0,001381
116	PFF1335c	0,062497	0,122784	0,00138
117	PFE0715w	0,059331	0,116563	0,00131
118	PFI1105w	0,05901	0,115933	0,001303
119	PFE1510c	0,058852	0,115622	0,0013
120	MAL8P1.86	0,057822	0,113599	0,001277
121	PFB0125c	0,057652	0,113265	0,001273
122	PFI1650w	0,056791	0,111574	0,001254
123	PFC0570c	0,05628	0,110569	0,001243
124	MAL7P1.134	0,055424	0,108888	0,001224
125	PFI1280c	0,055076	0,108204	0,001216
126	PF14_0392	0,054665	0,107396	0,001207
127	PF07_0037	0,054534	0,10714	0,001204
128	PFF0950w	0,054271	0,106622	0,001199
129	PF13_0077	0,053472	0,105052	0,001181
130	PF10_0046	0,05273	0,103594	0,001165
131	PFI0420c	0,051472	0,101124	0,001137
132	PFE1445c	0,051094	0,100381	0,001128
133	PF07_0020	0,050693	0,099593	0,00112
134	PF14_0038	0,04942	0,097093	0,001091
135	PFF1080w	0,049099	0,096461	0,001084
136	PFF0690c	0,048927	0,096124	0,001081
137	PF14_0559	0,048871	0,096013	0,001079

138	PFB0410c	0,048343	0,094976	0,001068
139	PF08_0012	0,047897	0,0941	0,001058
140	PFL1700c	0,047391	0,093106	0,001047
141	PF14_0565	0,047374	0,093073	0,001046
142	PFL1245w	0,046832	0,092007	0,001034
143	PF14_0231	0,046253	0,09087	0,001021
144	PFE0515w	0,045991	0,090355	0,001016
145	MAL13P1.165	0,045811	0,090002	0,001012
146	PFA0130c	0,04491	0,088232	0,000992
147	PF14_0248	0,04428	0,086994	0,000978
148	PF14_0564	0,044276	0,086986	0,000978
149	PF14_0433	0,04368	0,085814	0,000965
150	PFC0070c	0,043057	0,08459	0,000951
151	PFC0970w	0,043018	0,084514	0,00095
152	PF10_0030	0,042535	0,083565	0,000939
153	PFF0720w	0,042472	0,083443	0,000938
154	PFL0695c	0,042165	0,082839	0,000931
155	PF10_0280	0,04131	0,08116	0,000912
156	PFI0230c	0,039485	0,077574	0,000872
157	PFE0335w	0,039448	0,077501	0,000871
158	PFC0980c	0,039073	0,076764	0,000863
159	PF11_0140	0,038587	0,07581	0,000852
160	PF07_0015	0,038352	0,075348	0,000847
161	PF13_0261	0,038112	0,074877	0,000842
162	PF14_0633	0,037964	0,074586	0,000838
163	PFL2280w	0,037846	0,074354	0,000836
164	MAL13P1.124	0,037744	0,074153	0,000834
165	PF14_0063	0,037234	0,073151	0,000822
166	PFB0820c	0,037197	0,073079	0,000821
167	PF10_0293	0,037074	0,072836	0,000819
168	PFL0930w	0,035675	0,070087	0,000788
169	PF11_0083	0,035659	0,070056	0,000788
170	MAL8P1.30	0,034697	0,068166	0,000766
171	PFI0155c	0,03469	0,068153	0,000766
172	PFE0865c	0,034532	0,067842	0,000763
173	PFE1010w	0,033859	0,066521	0,000748
174	PF14_0084	0,033285	0,065393	0,000735
175	PF08_0117	0,033203	0,065232	0,000733
176	PF11_0258	0,033159	0,065145	0,000732
177	PFL1410c	0,032576	0,064	0,000719
178	PF13_0310	0,032564	0,063977	0,000719
179	PF10_0152	0,031999	0,062866	0,000707
180	PFA0590w	0,031419	0,061727	0,000694
181	PF14_0100	0,031035	0,060973	0,000685
182	PF11_0438	0,030807	0,060524	0,00068
183	PF11_0294	0,030424	0,059772	0,000672
184	PF07_0090	0,029506	0,057968	0,000652

185	PFE1275c	0,029254	0,057474	0,000646
186	PF13_0347	0,029121	0,057212	0,000643
187	PF11_0206	0,028804	0,056589	0,000636
188	PFI0590c	0,028626	0,056239	0,000632
189	PF14_0612	0,02817	0,055343	0,000622
190	PFE0310c	0,028083	0,055172	0,00062
191	PF13_0279	0,02808	0,055167	0,00062
192	PF10_0191	0,027547	0,05412	0,000608
193	PFF0805c	0,0274	0,053832	0,000605
194	PF14_0282	0,026962	0,05297	0,000595
195	PFF1440w	0,026843	0,052736	0,000593
196	PFF0105w	0,026098	0,051274	0,000576
197	PFL1930w	0,025623	0,050339	0,000566
198	MAL13P1.337	0,024924	0,048967	0,00055
199	PFB0520w	0,024522	0,048177	0,000542
200	PFF0685c	0,024404	0,047944	0,000539

Table B. 3: Top 174 genes ranked according to importance in MoA stratification from the 174-gene database

Rank	Feature	Relative importance	Scaled importance	Percentage
1	PF14_0642	1,170794385	1	0,04044429
2	PF10_0374	0,6596918	0,56345658	0,022788601
3	PFF1150w	0,640443679	0,547016357	0,022123688
4	PF14_0238	0,635361205	0,54267531	0,021948118
5	MAL7P1,18	0,632863021	0,540541558	0,02186182
6	PF13_0137	0,627244547	0,535742702	0,021667733
7	PFB0590w	0,59132845	0,505066011	0,020427036
8	PF14_0719	0,570275387	0,487084149	0,019699773
9	PFF0200c	0,564798262	0,482406022	0,019510569
10	PFF0875w	0,559871002	0,478197546	0,01934036
11	PF11_0527	0,556120806	0,474994425	0,019210812
12	PF11_0290	0,539762062	0,46102208	0,018645711
13	PFB0295w	0,52283989	0,446568498	0,018061146
14	PFF0965c	0,50905151	0,434791554	0,017584836
15	PFC0090w	0,508153347	0,434024414	0,017553809
16	PF10_0132	0,496192579	0,423808472	0,017140633
17	PF10_0191	0,487648556	0,416510843	0,016845485
18	PFE0585c	0,475323862	0,405984064	0,016419737
19	MAL13P1,124	0,473562628	0,404479757	0,016358897
20	PF07_0074	0,467524303	0,399322297	0,016150307
21	PFE0165w	0,427000834	0,364710353	0,014750451
22	PF14_0249	0,425756787	0,363647787	0,014707477
23	PFE0430w	0,402624458	0,343889981	0,013908386

24	PF14_0713	0,382754165	0,326918346	0,013221981
25	PFF0690c	0,382659815	0,32683776	0,013218721
26	PFB0270w	0,377027517	0,322027097	0,013024157
27	PF08_0032	0,376272929	0,321382587	0,012998091
28	MAL13P1,262	0,364396803	0,31123894	0,012587838
29	PFC0760c	0,35366006	0,302068462	0,012216945
30	PF14_0041	0,352275013	0,300885465	0,012169099
31	PF10_0299	0,352076068	0,300715542	0,012162227
32	MAL13P1,206	0,351600438	0,300309296	0,012145796
33	PFF0610c	0,345773791	0,295332635	0,011944519
34	PFL2055w	0,33335515	0,284725614	0,011515525
35	PF14_0562	0,31143001	0,265998892	0,010758136
36	PFL0980w	0,309747694	0,264561991	0,010700022
37	PF13_0095	0,30775753	0,262862151	0,010631273
38	PF13_0135	0,304230441	0,25984959	0,010509432
39	PFA0430c	0,293773819	0,25091837	0,010148215
40	PF10_0380	0,282256359	0,241081066	0,009750353
41	MAL13P1,40	0,275011403	0,234892997	0,009500081
42	PF10_0152	0,268429886	0,229271586	0,009272727
43	MAL8P1,132	0,26197581	0,223759025	0,009049775
44	PFC0980c	0,255234759	0,218001352	0,00881691
45	PF14_0360	0,251632286	0,214924404	0,008692465
46	PFE0450w	0,247351703	0,211268269	0,008544595
47	PF14_0384	0,246396172	0,21045213	0,008511587
48	MAL13P1,137	0,245780177	0,209925996	0,008490308
49	PF11_0162	0,243095583	0,207633028	0,00839757
50	PFB0175c	0,241068726	0,205901847	0,008327554
51	PF10_0307	0,240789033	0,205662955	0,008317892
52	PF14_0214	0,240551394	0,205459983	0,008309683
53	PFC0375c	0,238424582	0,203643428	0,008236214
54	PF13_0170	0,22892863	0,195532737	0,007908183
55	MAL8P1,105	0,226774838	0,193693137	0,007833781
56	PFL0625c	0,223433676	0,190839381	0,007718363
57	PFC0440c	0,221573178	0,189250291	0,007654094
58	PF14_0138	0,219449798	0,187436668	0,007580743
59	PFL1135c	0,212090019	0,181150526	0,007326504
60	PFA0180w	0,204200343	0,174411789	0,007053961
61	MAL13P1,135	0,195065735	0,166609729	0,006738412
62	PF10_0111	0,194943857	0,166505631	0,006734202
63	PF14_0112	0,192912587	0,16477068	0,006664033
64	PF14_0499	0,19243797	0,1643653	0,006647638
65	PFI0965w	0,191508119	0,163571095	0,006615517
66	PF11_0090	0,185638845	0,158558025	0,006412767
67	PF10_0020	0,182795059	0,156129087	0,00631453
68	PFE0815w	0,182520333	0,155894439	0,00630504
69	PF10_0136	0,170876019	0,145948786	0,005902795
70	PF10_0085	0,170702234	0,145800352	0,005896792

71	PFE1340w	0,169213042	0,144528402	0,005845349
72	MAL8P1,91	0,169047206	0,144386758	0,00583962
73	PFI0500w	0,168901816	0,144262578	0,005834598
74	MAL13P1,338	0,163052487	0,139266544	0,005632537
75	PFC0400w	0,159712871	0,136414108	0,005517172
76	PF10_0243	0,150404154	0,128463337	0,005195609
77	PFL2200w	0,130709634	0,111641835	0,004515275
78	PF13_0032	0,130209928	0,111215026	0,004498013
79	PF08_0069	0,128724875	0,109946611	0,004446713
80	PFD0470c	0,124259113	0,10613231	0,004292446
81	PFF1280w	0,114307442	0,09763238	0,003948672
82	PFI1425w	0,114014473	0,097382148	0,003938552
83	PFL1680w	0,111120469	0,094910319	0,003838581
84	PF11_0435	0,107609583	0,091911598	0,003717299
85	PFL0580w	0,105190805	0,089845669	0,003633744
86	PFL2485c	0,097515266	0,08328983	0,003368598
87	PFL0685w	0,096890279	0,082756016	0,003347008
88	PFI0905w	0,08688386	0,074209324	0,003001343
89	PFL1270w	0,086296732	0,073707846	0,002981062
90	PF11_0214	0,079144047	0,067598588	0,002733977
91	PFB0923c	0,078907322	0,067396396	0,002725799
92	PF14_0593	0,078713159	0,067230557	0,002719092
93	PFC0100c	0,076991507	0,065760059	0,002659619
94	PF13_0042	0,069084694	0,059006684	0,002386483
95	PF13_0350	0,068193155	0,058245202	0,002355686
96	PF07_0079	0,065599591	0,056029984	0,002266093
97	PF10_0163	0,058553163	0,050011482	0,002022679
98	PF11_0348	0,057670455	0,049257543	0,001992186
99	PFE1280w	0,05156611	0,044043694	0,001781316
100	MAL13P1,322	0,046987156	0,040132714	0,001623139
101	PF11_0260	0,042054613	0,035919725	0,001452748
102	PF14_0552	0,041797955	0,035700509	0,001443882
103	PF13_0023	0,041743771	0,035654229	0,00144201
104	PF11_0079	0,039624547	0,033844155	0,001368803
105	MAL7P1,94	0,036010514	0,030757334	0,001243959
106	PF13_0219	0,035827487	0,030601007	0,001237636
107	PFF1155w	0,029193031	0,024934379	0,001008453
108	MAL13P1,321	0,028800495	0,024599106	0,000994893
109	PF14_0455	0,027506026	0,023493473	0,000950177
110	PFA0490w	0,021995817	0,018787087	0,00075983
111	PFC0965w	0,020674435	0,017658468	0,000714184
112	MAL8P1,108	0,014640583	0,012504829	0,000505749
113	MAL7P1,75	0,011687204	0,009982285	0,000403726
114	PFD0285c	0,010983744	0,009381445	0,000379426
115	PFE1390w	0,010904277	0,009313571	0,000376681
116	MAL13P1,179	0,007488856	0,006396389	0,000258697
117	PF14_0730	0,005041678	0,004306203	0,000174161

118	PFE0745w	0	0	0
119	PFB0915w	0	0	0
120	PFC0230c	0	0	0
121	PF14_0540	0	0	0
122	PFL2370c	0	0	0
123	MAL8P1,96	0	0	0
124	PF14_0419	0	0	0
125	PFE0205w	0	0	0
126	MAL7P1,29	0	0	0
127	MAL13P1,293	0	0	0
128	PF08_0018	0	0	0
129	MAL8P1,92	0	0	0
130	PF14_0327	0	0	0
131	PFI1185c	0	0	0
132	PF13_0208	0	0	0
133	PFD0785c	0	0	0
134	PFL1715w	0	0	0
135	PFE1255w	0	0	0
136	MAL8P1,123	0	0	0
137	PF13_0228	0	0	0
138	PFL0925w	0	0	0
139	PFA0330w	0	0	0
140	PFL2120w	0	0	0
141	PFF0715c	0	0	0
142	PFL1490w	0	0	0
143	PFL0885w	0	0	0
144	PFF0150c	0	0	0
145	PF14_0071	0	0	0
146	PF14_0350	0	0	0
147	PFB0395w	0	0	0
148	PFI0855w	0	0	0
149	PFF1390w	0	0	0
150	PF11_0329	0	0	0
151	PF14_0140	0	0	0
152	PFF0325c	0	0	0
153	PFL0865w	0	0	0
154	PF11_0252	0	0	0
155	PF14_0372	0	0	0
156	PF07_0047	0	0	0
157	PFI0710c	0	0	0
158	PF13_0087	0	0	0
159	MAL7P1,171	0	0	0
160	PF11_0438	0	0	0
161	PF11_0087	0	0	0
162	PFB0920w	0	0	0
163	PFF0490w	0	0	0
164	MAL13P1,256	0	0	0

165	PFE0260w	0	0	0
166	PF14_0286	0	0	0
167	PFD0545w	0	0	0
168	PFL2520w	0	0	0
169	PF10_0084	0	0	0
170	MAL8P1,153	0	0	0
171	PF14_0606	0	0	0
172	PFF0595c	0	0	0
173	MAL7P1,65	0	0	0
174	PF13_0090	0	0	0