ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Machine Learning Approaches Identify Chemical Features for Stage-Specific Antimalarial Compounds

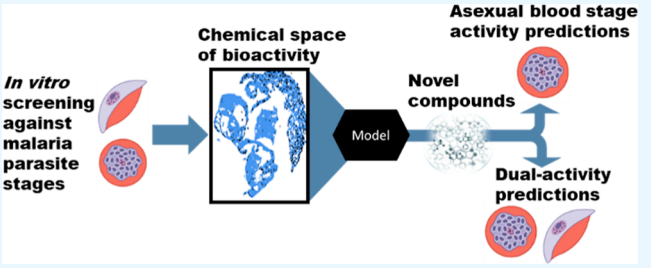
Ashleigh van Heerden, Gemma Turon, Miquel Duran-Frigola, Nelishia Pillay, and Lyn-Marié Birkholtz*

Cite This: *ACS Omega* 2023, 8, 43813−43826

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Efficacy data from diverse chemical libraries, screened against the various stages of the malaria parasite *Plasmodium falciparum*, including asexual blood stage (ABS) parasites and transmissible gametocytes, serve as a valuable reservoir of information on the chemical space of compounds that are either active (or not) against the parasite. We postulated that this data can be mined to define chemical features associated with the sole ABS activity and/or those that provide additional life cycle activity profiles like gametocytocidal activity. Additionally, this information could provide chemical features associated with inactive compounds, which could eliminate any future unnecessary screening of similar chemical analogs. Therefore, we aimed to use machine learning to identify the chemical space associated with stage-specific antimalarial activity. We collected data from various chemical libraries that were screened against the asexual (126 374 compounds) and sexual (gametocyte) stages of the parasite (93 941 compounds), calculated the compounds' molecular fingerprints, and trained machine learning models to recognize stage-specific active and inactive compounds. We were able to build several models that predict compound activity against ABS and dual activity against ABS and gametocytes, with Support Vector Machines (SVM) showing superior abilities with high recall (90 and 66%) and low false-positive predictions (15 and 1%). This allowed the identification of chemical features enriched in active and inactive populations, an important outcome that could be mined for essential chemical features to streamline hit-to-lead optimization strategies of antimalarial candidates. The predictive capabilities of the models held true in diverse chemical spaces, indicating that the ML models are therefore robust and can serve as a prioritization tool to drive and guide phenotypic screening and medicinal chemistry programs.



1. INTRODUCTION

From 2000 to 2019, malaria-associated deaths showed a steady decline, but these gains were stalled in 2020, with a 12% increase in malaria mortality reported globally.[1] Compounding factors include the COVID-19 pandemic, which hindered control efforts and the continued emergence of drug-resistant malaria parasites.[2] Therefore, efforts toward discovering and developing potent antimalarials with novel modes of action must be sustained, and such compounds should target multiple life cycle stages of *Plasmodium* parasites.[3,4] Importantly, compounds with the ability to target the transmission of the parasite are sought after as they could be employed to limit the spread of the parasite, and hence disease, and support malaria-elimination strategies.[5] Aside from the required need for compounds with prophylactic activity (blocking exoerythrocytic development of parasites during liver schizogony), transmission-blocking (TrB) activity can also be ascribed to compounds able to block sexual gametocyte development and subsequent human-to-mosquito transmission of the parasite.

The discovery of compounds with TrB activity through phenotypic whole-cell screening is fraught with challenges associated with the unique biology inherent to the sexual gametocyte stages of the human malaria parasite, *Plasmodium falciparum*. In this species, only a small portion (∼10%) of the asexual blood stage (ABS) parasites commit to gametocyte development, thereby switching from a proliferative cycle to cellular differentiation. Subsequently, sexually committed parasites differentiate from immature gametocytes (stage I−III) to transmissible, mature stage V gametocytes.[6] This process is uniquely prolonged among the *laverinia* species (*falciparum* and *reichenowi*) and can take ∼10−12 days until mature (stage V) male and female gametocytes are produced as the only forms to support transmission. Identifying gametocytocidal activity is, therefore, more complicated compared to screening for ABS-inhibiting compounds[5,7] due to the limited and time-consuming production of gametocyte populations and the typical need for complementary,

Received: August 2, 2023
Revised: October 18, 2023
Accepted: October 20, 2023
Published: November 7, 2023

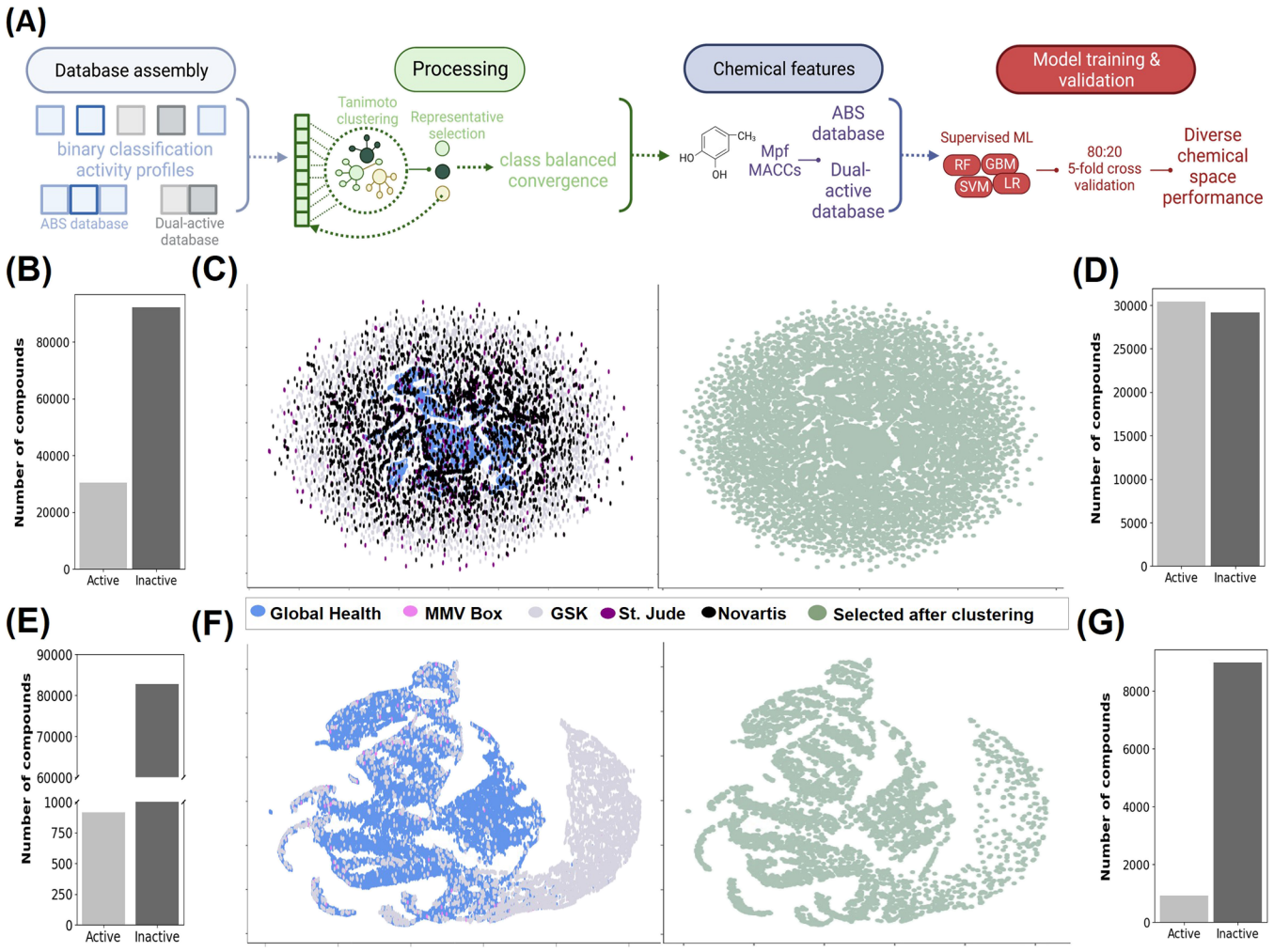


Figure 1. ABS and dual-active database assembly and preprocessing for model building. (A) The pipeline used for data assembly, curation, processing, chemical featurization and model building. Data from phenotypic screening of chemical libraries against ABS and/or gametocytes were used for binary definition of active and inactive compounds. Class imbalance was addressed via cluster-based undersampling and the compounds converted into molecular descriptors ECFP or MACCS to allow model building for compound activity prediction (created with http://biorender.com/). (B,E) Class imbalance in the ABS (B) and dual-active (C) data sets after binary classification of activity vs inactivity based on criteria as specified in the original screens. (C,F) UMAP projection of the chemical space in the databases before (left-hand image) and after (right-hand image) cluster-based undersampling on inactive compounds for each database. (D,G) Distribution of active vs inactive compounds for the ABS (D) and dual-active (G) data sets after cluster-based undersampling.

orthogonal assays to interrogate the effect of a compound on the differentiated biology associated with transmissible stages.[8]

Despite these constraints, large chemical libraries have been screened against both the ABS and gametocyte stages of the parasite.[9−14] This identified compounds with stage-specific activity (e.g., only ABS active) as well as compounds with dual activity (able to target both ABS and mature gametocytes) or multistage activity (compounds with ABS, TrB, and liver-stage activity). These phenotypic screens provide valuable information about the chemical space associated with the differential activity of antiplasmodial compounds. We proposed that these data can be mined to define chemical features associated with compounds displaying activity against only ABS or features in compounds that additionally are active against other life cycle stages including gametocytes. Such information would be very useful in the design of derivatives of hits during a hit-to-lead medicinal chemistry campaign to enable multistage activity. Additionally, this could exclude compounds that are least likely to show gametocytocidal activity and therefore allow the redistribution of time and resources toward more promising compounds.

Analyzing compound activity profiles and relating chemical structural information to bioactivity at scale are immense and data-heavy tasks well aligned for machine learning (ML) approaches. The latter have identified important patterns with predictive power in data-dense settings[15] and are being increasingly applied to drug discovery.[16,17] While proof-of-concept ML models have been generated with fair ability to predict hit compounds with activity against ABS and liver stages independently,[18−20] this has not been extended to gametocytes. In this study, we expand on and improve these approaches to create an accurate and robust ML classification model able to predict ABS activity or gametocytocidal activity, thereby identifying stage-specific as well as dual-active compounds. Our goal was to use more simple models, such as SVM and RF rather than complex models, such as neural networks. Emphasis was also placed on capturing the chemical space from phenotypic screening data using classical and

https://doi.org/10.1021/acsomega.3c05664
ACS Omega 2023, 8, 43813−43826

Table 1. Chemical Libraries Were Phenotypically Screened against *P. falciparum*[a]

chemical library	no. compounds in library	stage screened	no. compounds obtained	no. compounds (balanced database)	refs
GSK library	∼2 million	ABS	40 510	ABS: 40 398	9
TCAMS library	∼14 000	dual	13 533	dual: 409	24
St. Jude's library	∼310 000	ABS	5456	ABS: 3754	11
Novartis-GNF Malaria Box	1.7 million	dual	11 394	ABS: 9689	13
Global Health Chemical library	∼70 000	dual	68 614	ABS: 14 584	12
				dual: 7901	
MMV Box	400	dual	400	ABS: 304	14, 23
				dual: 277	
PRB Box	400	dual	400	NA	3
Pathogen Box	400	dual	367	NA	26
total screened (ABS)			126 374	68 729	
total screened (dual)			93 941	8614	

[a]ABS = asexual blood stage. Dual = ABS and sexual stages.

interpretable molecular fingerprints that allow ML modeling over a broad range of bioactivity prediction tasks.[20−22] Since phenotypic screening data are severely imbalanced with few active compounds compared to inactive compounds, we additionally used a hybrid approach to allow the training of models on imbalanced data by combining cluster-based undersampling and algorithms that can implement stronger penalties when misclassifying minority classes. This allowed us to retain relevant chemical space information while decreasing the class-imbalance severity to enable model building. We propose that the resultant models will be highly beneficial to accelerate antimalarial drug discovery and development of multistage active antiplasmodial compounds by prioritizing compounds that show the highest probability of the desired activity toward a specific stage of the parasite. Not only can this reduce the wasteful expenditure of time and resources on compounds with a low probability of showing activity against different stages of the parasite but these models can also be mined to identify stage-specific chemical features important for activity against the parasite.

2. METHODS

2.1. Data Acquisition, Quality Control Filtering, and Preprocessing of Chemical Library Data Sets. Inhibition data of chemical libraries screened against either the asexual and/or gametocyte stages of the parasites were acquired and preprocessed (including removal of inorganics and organometallic compounds) (Figure 1A). These chemical libraries contain multiple chemical spaces, each with several chemically similar analogs. Ultimately, two databases were formed: (1) an ABS database that contained SMILES and inhibition data from 5 chemical libraries screened against *P. falciparum* ABS parasites;[3,9,11−13,23] and similarly (2) a database (referred to as the dual-active database) that contained SMILES and inhibition data from chemical libraries screened against ABS and any of the gametocyte stages (stage I−V) of *P. falciparum*, with the majority of screening data focused against stage IV/V gametocytes[12,23,24] (Table 1). As the chemical libraries were screened by different research groups, using different assay platforms, with different thresholds set to define parasite inhibition and/or gametocytocidal activity, the thresholds specified within the respective screens were used as is to define the binary definition of active/inactive compounds for parasite viability inhibition (File S1). Inactive compounds were retained for both databases, as these are informative to define the relevant chemical space for bioactivity. Compound SMILES was extracted for all compounds included in the respective databases. Uniform manifold approximation and projection (UMAP) analysis was performed for outlier detection and to obtain a projection of the chemical composition of the databases generated with the help of the umap-learn python package version 0.5.3.[25]

2.2. Cluster-Based Undersampling of Inactive Compounds in the Databases. The databases created above contained information about both active and inactive compounds on ABS and gametocyte stages of *P. falciparum* parasites. However, the data are inherently skewed toward inactive compounds, and this necessitated implementation of class balancing to address the class bias (Figure 1A). We therefore performed a cluster-based undersampling using Tanimoto dissimilarity (0.4 distance threshold) with RDKit version 2022.9.5[27] to allow chemical substructure searching and clustering of inactive compounds with similar substructures. The aim was not to completely cluster the chemical databases but to use clustering to aid in the undersampling of inactive compounds. This was independently performed for the "ABS" and "dual-active" databases.

To perform clustering using Tanimoto dissimilarity, compound SMILES of the respective databases were converted by RDKit molecular fingerprints (parameter max path = 5). Compounds were divided into seven subsets of 15−20 000 compounds each for the ABS database but only three subsets of similar size for the dual-active database to reduce the data density and allow the computation of clustering. Parallel clustering was then applied to each of these subsets individually. From this, a representative compound was randomly selected from each inactive cluster within each of the individual subsets. All these identified representatives for the clusters were merged into a single set before chemical clustering was repeated. This process was continued until the database was considered as balanced once the number of representatives (inactive compounds) were equal to or >95% of the number of active compounds. For the dual-active database, complete balancing could not be achieved with cluster-based undersampling and the iterative clustering process was halted before reaching the minimum number of clusters to prevent severely limiting the chemical space and causing difficulty in pattern detection in ML.

To determine if a cluster-based undersampling allowed better model performance than imbalanced data that had not been preprocessed (undersampled or oversampled), models trained on cluster-based, undersampled data were compared to

https://doi.org/10.1021/acsomega.3c05664
ACS Omega 2023, 8, 43813−43826

models trained on imbalanced data. Similarly, to ensure cluster-based undersampling was a better class imbalance correction technique compared to oversampling, the performance of models trained on cluster-based undersampled data was compared to models trained on oversampled data preprocessed via SMOTE version 0.11.0.[28]

2.3. Conversion of SMILES to Molecular Fingerprints for ML. For each of the undersampled databases, canonical compound SMILES was extracted and converted into molecular fingerprints representing the respective compound to allow ML (Figure 1A). Two molecular fingerprints, Morgan fingerprint (ECFP) and MACCS, were generated for each compound to successfully relate structural information for subsequent ML applications.[16,21] Ultimately, an ECFP was generated with RDKit[27] that produced a 500-bit feature matrix (5 atoms distance per bit) containing the chemical group and substructure information on the compound. Second, MACCS keys were generated for each compound in DeepChem version 2.7.1[29] to generate a 166-feature matrix containing certain chemical groups that have been identified as important within drug discovery and virtual screening.[30]

2.4. Training Supervised ML Models on the ABS and Dual-Active Database. For each chemical database (ABS and dual active), two models were built for each of the algorithms used, one using ECFP as molecular descriptors and the other using MACCS. This was to allow comparative evaluation of the best molecular fingerprint to predict bioactivity in either ABS and/or gametocytes stages in representative and novel chemical spaces.

For model building, the respective databases were randomly split into training and testing sets at a ratio of 80:20, whereby 80% of the compounds were used to train models and the resultant 20% were merged with compounds excluded during undersampling to generate an imbalanced test set for model evaluation on untrained data. Models underwent a grid search cross-validation hyperparameter tuning to identify the optimal hyperparameters (Table S1). Each hyperparameter-tuned model then underwent fivefold cross-validation to assess the average accuracy and variability within model predictions.

Since class imbalance was still present (particularly within the dual-active database) and to prevent class bias within models, ML algorithms that applied weight-based mechanisms/penalties on the misclassification of minority classes (dual-active compounds) or other mechanisms for training on imbalanced data were subsequently selected for model building from the scikit-learn python package version 0.20.[31] This included ensemble methods such as random forest (RF) and gradient boosting machines (GBM) that have been shown to perform well on imbalanced data.[32] Additionally, single classifiers such as support vector machines (SVM) and logistic regression (LR) were also applied as these algorithms can attribute weights to a minority class (active compounds), thereby penalizing the model more heavily for misclassifying active compounds.

For the balanced ABS database, the ABS activity prediction models were trained on 80% (47 530) of the compounds, whereby the models had to identify patterns within molecular fingerprints of compounds (ECFP or MACCS keys) for the correct prediction of compound bioactivity against ABS. For SVM, RF, GBM, and LR models, the scikit-learn python package was used to build and train the model to the training set. During the training, ABS activity models were built using the optimal hyperparameters identified (Table S1) and underwent subsequent fivefold cross-validation. Thereafter, the imbalanced test set (61 029) excluded from training was used to assess the model bioactivity prediction accuracy and overfitting on imbalanced untrained data.

Dual-activity prediction models were similarly trained on 80% (7913 compounds) of the dual-active database. Like the models for the ABS data, scikit-learn was used to build the RF and GBM models. For LR models, however, the class weight was additionally set as 1 for inactive compounds and 10 for active compounds to compensate for class imbalance in dual-activity prediction models when training on the training set. Similarly, for SVM the class weight was set to "balanced" to adjust class weights inversely proportionally to the frequency of the class. Each dual-activity prediction model was built using the optimal hyperparameters identified (Table S1) and subsequently underwent 5-fold cross-validation before evaluation on the imbalanced test set (62 375 compounds) to assess model bioactivity prediction accuracy and overfitting.

2.5. Metrics for Evaluating Different ML Algorithms in Predicting Asexual and Dual Activity. Tuned models were assessed on their cross-validation results as well as their test set results to determine the model performance on untrained imbalanced chemical data and to highlight any overfitting within the models. Metrics used for model predictions evaluation were recall (eq 1), Precision (eq 2), false-positive rate (eq 3), receiver operator characteristic curve (ROC−AUC), and the *F*1-score (eq 4).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{3}$$

$$F1 - \text{score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

Recall and precision of models were calculated to determine whether models were able to correctly predict active and inactive compounds, whereas FPR assessed false-positive predictions. The ROC−AUC score was used to evaluate how well the model could distinguish between two classes. A score of 0.7 indicates that the model has a 70% chance to correctly distinguish active and inactive compounds. In addition, we also used the *F*1-score, which combines recall and precision.[33] Besides precision and recall, models were additionally evaluated for optimized sensitivity (eq 5 below) and specificity (eq 6), used to define the geometric mean (eq 7), when predicting active and inactive compounds within novel chemical spaces.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{6}$$

$$G - \text{mean} = \sqrt{\text{sensitivity} \times \text{specificity}} \tag{7}$$

Considering the class imbalance present, especially within the dual-active database, the GHOST python package version 0.6.1[34] was used to adjust the probability thresholds for model

https://doi.org/10.1021/acsomega.3c05664
ACS Omega 2023, 8, 43813−43826

decision and to examine if this enabled better model performance upon the test set to the metrics mentioned above. The influence of adjusting the probability discrimination threshold on model performance was visualized using the yellowbrick python package version 1.5.[35] To determine whether such simplistic models is on par or even better than complex models such as neural networks, the top performing model based on these metrics would then be compared to more complex models generated via the autogluon python package version 0.8.2[36] using the same training and test set data.

2.6. Chemical Features Important for Predicting ABS Inhibition and Dual Activity. These compound activity prediction models hold important chemical space information for activity prediction. The aim was to use the models to identify chemical features that are both statistically significant and important for the model in compound activity prediction. Some models have the ability to rank features (i.e., chemical features) according to the weight/importance such features have in predicting activity or inactivity of a compound. Unfortunately, highlighting the ECFP-enriched features important for predicting dual activity was complicated due to our best performing model, SVM, employed a radial basis function (RBF) kernel transformation technique which complicated how one can calculate the true weight a feature holds within model predictions. To allow interpretability, RF models trained on ECFP with good precision and ability to distinguish active and inactive compounds within representative chemical spaces against ABS and/or gametocytes, were selected to perform feature importance analysis. To highlight such features important for activity prediction, ECFP features significantly enriched within active compounds (ABS inhibiting or dual active) were identified using the *Z*-test on two sample proportions. From the *Z*-scores calculated using *Z*-test, ECFP features significantly enriched (*Z*-score > 2.5, *p*-value < 0.01) within active compounds compared to inactive compounds were identified and then ranked according to the feature importance score obtained from the best performing activity prediction model. From this the top 100 ECFP features were identified and compared to the top 100 ECFP features of other activity prediction models to highlight any overlap between the top ECFP features of different models. This assumes that if features are predictive of compound activity, such features should overlap as the top features for different models that employ different algorithms in pattern recognition.

2.7. External Validation of ML Performance on Chemically Diverse PRB and Pathogen Box. To evaluate the limits of activity prediction for the models in novel chemical spaces and to externally validate the models, they were additionally tested against new chemical libraries that included chemically diverse compounds. This included the data from the open-source Medicines for Malaria Venture (MMV) Pandemic Response Box (PRB) and the Pathogen Box, with potent activity against various stages of the parasite (http://www.mmv.org/).[3,26] Compounds within the PRB and Pathogen Box that were not present within either the training or test sets were extracted and used for external validation. The hit rate of the best performing model was then compared with that of the chemical library and random selection. Here the hit rate of the top 100 compounds (ordered according to model probability) was compared with the randomly selected 100 compounds. To assess whether the models aided in limiting the number of active compounds in the bottom 100, that is, least likely to have activity or last to be screened via random selection, the hit rate for the bottom 100 was also calculated with the goal of having a hit rate lower than that of randomly screening. The idea being the top 100 would be the first compounds that are randomly selected to start with during phenotypic screening whereas the bottom would be the last compounds chosen to screen. Additionally, to better evaluate compound prioritization, the enrichment factor (EF) was calculated for the top 10 and top 50 compounds based on the predicted probability of compound activity.

3. RESULTS

3.1. Database Generation and Class Imbalance Correction. A data analysis pipeline was devised to allow database generation, class imbalance correction and processing, chemical featurization, and model training and validation (Figure 1A). This pipeline was applied in parallel to create two different modeling environments: (1) an environment for compounds with predicted stage-specific activity against ABS parasites alone, and (2) for compounds with dual activity against both ABS parasites and gametocytes. No outliers were detected within either database (Figure S1). Therefore, the two databases were generated from data sets of chemical libraries screened against either the asexual and/or gametocyte stages of *P. falciparum* parasites (Figure 1A and Table 1).[3,9,11−13,23,24] The first database included compounds that were screened against ABS and the second database, where compounds were screened against both ABS and gametocyte stages. Of all active compounds, 96% were those inhibiting ABS and 3% showed dual activity, with only 0.3% displaying sole activity against gametocyte stages. Binary classifications of activity were retained per criteria defined within each screen.

Within these data sets, the ratio of inactive to active compounds is inherently skewed toward the inactive compounds, creating imbalanced data sets with inactive compounds comprising 75% of the ABS database (30 393 active vs 92 178 inactive compounds) compared to a more severe situation of 99% in the dual-active database (916 active vs 68 801 inactive compounds) (Figure 1B,E). To prevent loss of relevant chemical space information present in inactive samples,[37] cluster-based undersampling was performed on inactive compounds for each database (Figure 1A) to generate balanced data sets more amenable for conventional ML modeling.[38] This clustering approach could be justified as it was observed that these chemical libraries contained multiple structurally related compounds within specific chemical spaces that have similar inactivity, which could function as representative compounds (Figure 1C). UMAP analysis and spatial projection of the ABS database indicated retention of the chemical composition and diversity after two rounds of subset clustering of inactive compounds (Figure 1C), with class balancing attained (30939 active compounds vs 29 143 inactive compounds, Figure 1D). However, after multiple rounds of clustering for the dual-active database, only a maximal of 3745 clusters representing the inactive compounds were obtained, which did not correct the class imbalance. Therefore, additional parallel chemical clustering was performed, and this process halted before reaching the maximal number of chemical clusters. This resulted in more than one chemical representative of clusters present (8975 clusters) to define the chemical space for ML for the inactive compounds and corrected the class imbalance to less than eightfold (916 active compounds vs 8975 inactive compounds, Figure 1E,G).

https://doi.org/10.1021/acsomega.3c05664
ACS Omega 2023, 8, 43813−43826

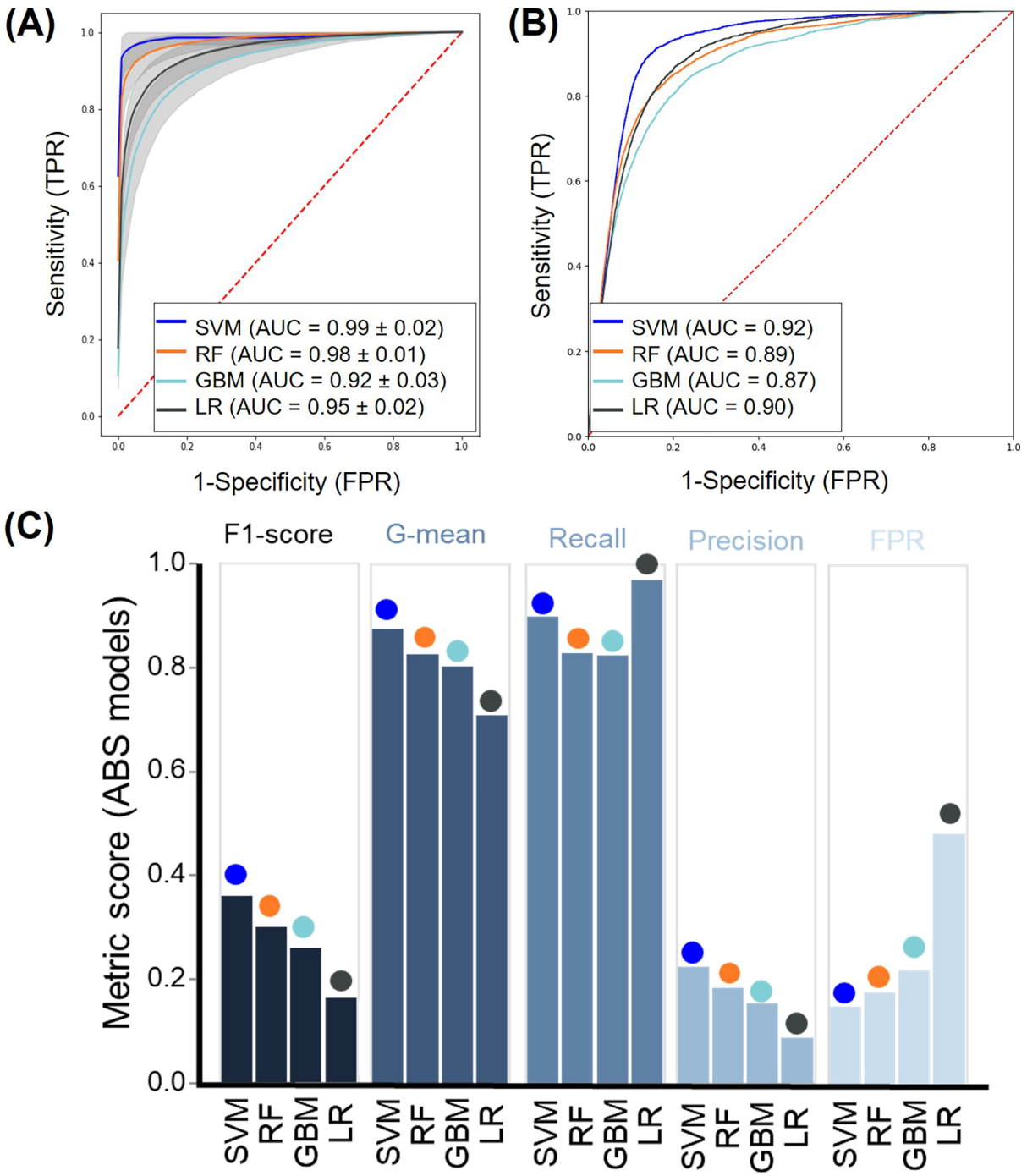


Figure 2. Performance of different ML algorithms in identifying compounds with ABS activity. (A) ROC−AUC curves showing performance of different ML algorithms in predicting compounds with ABS activity when trained on the ECFP of compounds after fivefold cross-validation. Insert indicates AUC mean values ± standard deviation. (B) ROC−AUC curves showing the performance of different ML algorithms on the imbalanced test set. (C) Model performance metrics associated with the performance of the different models in predicting the imbalanced test set data. The *F*1-score evaluated model performance on imbalanced data, whereas *G*-mean scores determined how well models could optimize sensitivity and specificity. Recall and precision indicated accuracy of activity predictions, whereas false-positive rate (FPR) indicated error within predictions.

The chemical distribution and composition for inactive compounds (Figure 1F) were also retained for the dual-active database. From these more balanced databases, the chemical clusters were shuffled and randomly split where 80% of compounds were used for model training and the remaining 20% was merged with inactive compounds excluded during cluster-based undersampling to create an imbalanced test set for model evaluation.

To determine if the training on cluster-based undersampled data as mentioned above skewed model performance on the imbalanced test set, we compared this strategy to oversampling or training on imbalanced data without preprocessing. The models trained on undersampled data for both ABS and dual activity resulted in higher sensitivity but similar specificity compared to models trained on oversampled or severely imbalanced data, with higher *G*-mean and ROC−AUC scores

https://doi.org/10.1021/acsomega.3c05664
ACS Omega 2023, 8, 43813−43826

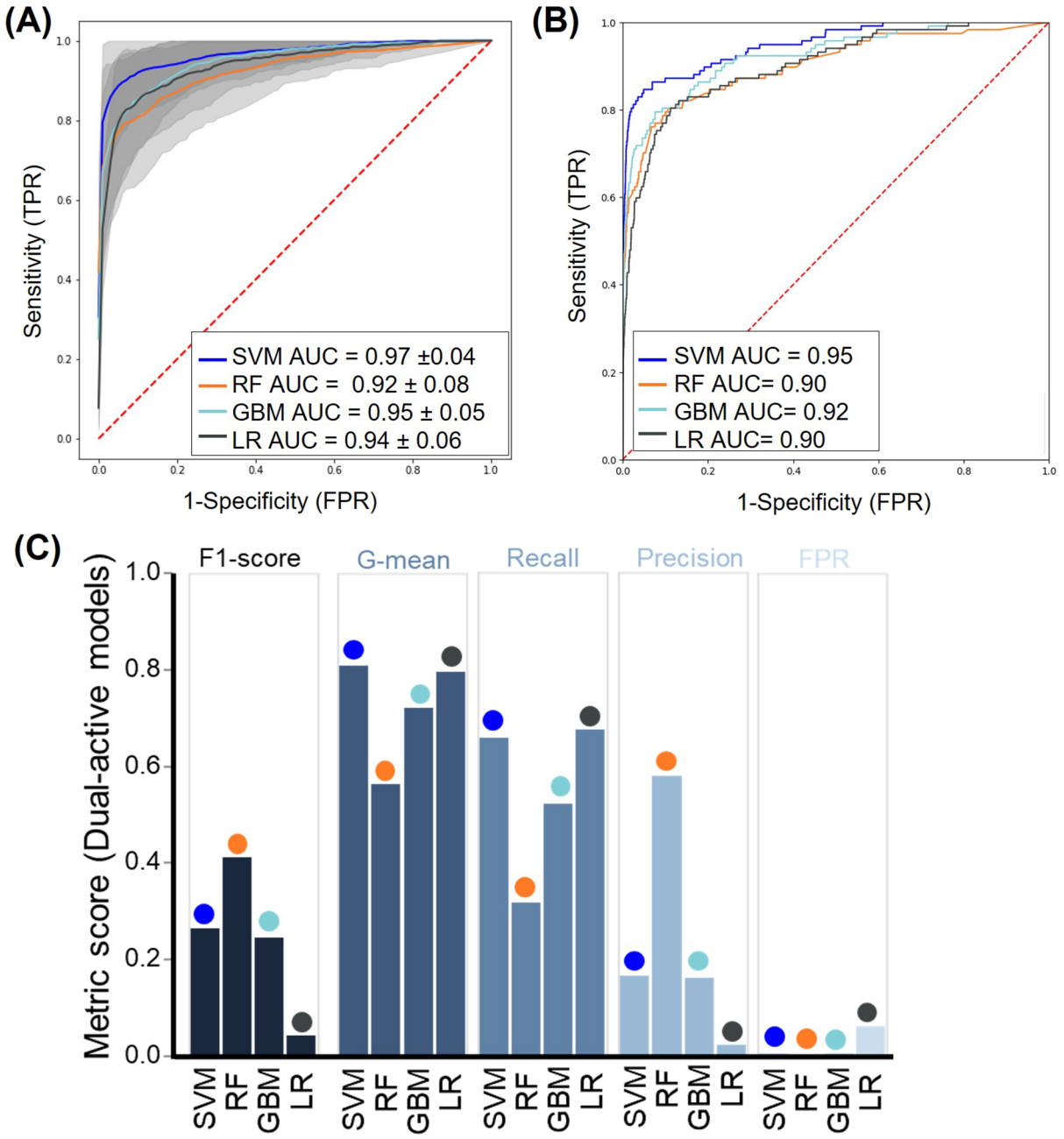


Figure 3. Performance of different ML algorithms in identifying compounds with dual activity. (A) ROC−AUC curves showing fivefold cross-validation performance of different ML algorithms in predicting compounds with dual activity when trained on the ECFP of compounds. Insert indicates AUC mean values ± standard deviation. (B) ROC−AUC curves showing performance of different ML algorithms on the imbalanced test set. (C) Model performance metrics associated with the performance of the different models in predicting imbalanced test set data. The *F*1-score evaluated model performance on imbalanced data, whereas *G*-mean scores determined how well models were able to optimize sensitivity and specificity. Recall and precision indicated accuracy of activity predictions whereas false-positive rate (FPR) indicated error within predictions.

for the undersampled data (File S1). This indicates that the models trained on oversampled data fails to identify active compounds due to the model possibly fixating on patterns associated with the oversampled active compounds in the training data.[39] Cluster-based undersampling, therefore, provided a better class imbalance correction technique and improved model sensitivity.

3.2. ABS Activity Prediction Model Selection and Performance. To associate chemical features of compounds with activity against ABS (or lack thereof) during model training, a two-pronged approach was used: the SMILES for compounds from the ABS database was converted into either ECFP or MACCS molecular fingerprints. Subsequently, data from both featurization methods were used, and four different models each were trained on 80% of the data in each instance (Figure 1A). The best bit-length for ECFP was determined to be 500 bits by model evaluation on different bit lengths (Figure S2), with an atom radius of 5 providing better precision within models (Figure S3). For the models trained on ECFPs of compounds from the ABS database, the SVM model achieved the highest ROC−AUC score with the lowest variability (0.99 ± 0.02) during fivefold cross-validation

https://doi.org/10.1021/acsomega.3c05664
ACS Omega 2023, 8, 43813−43826

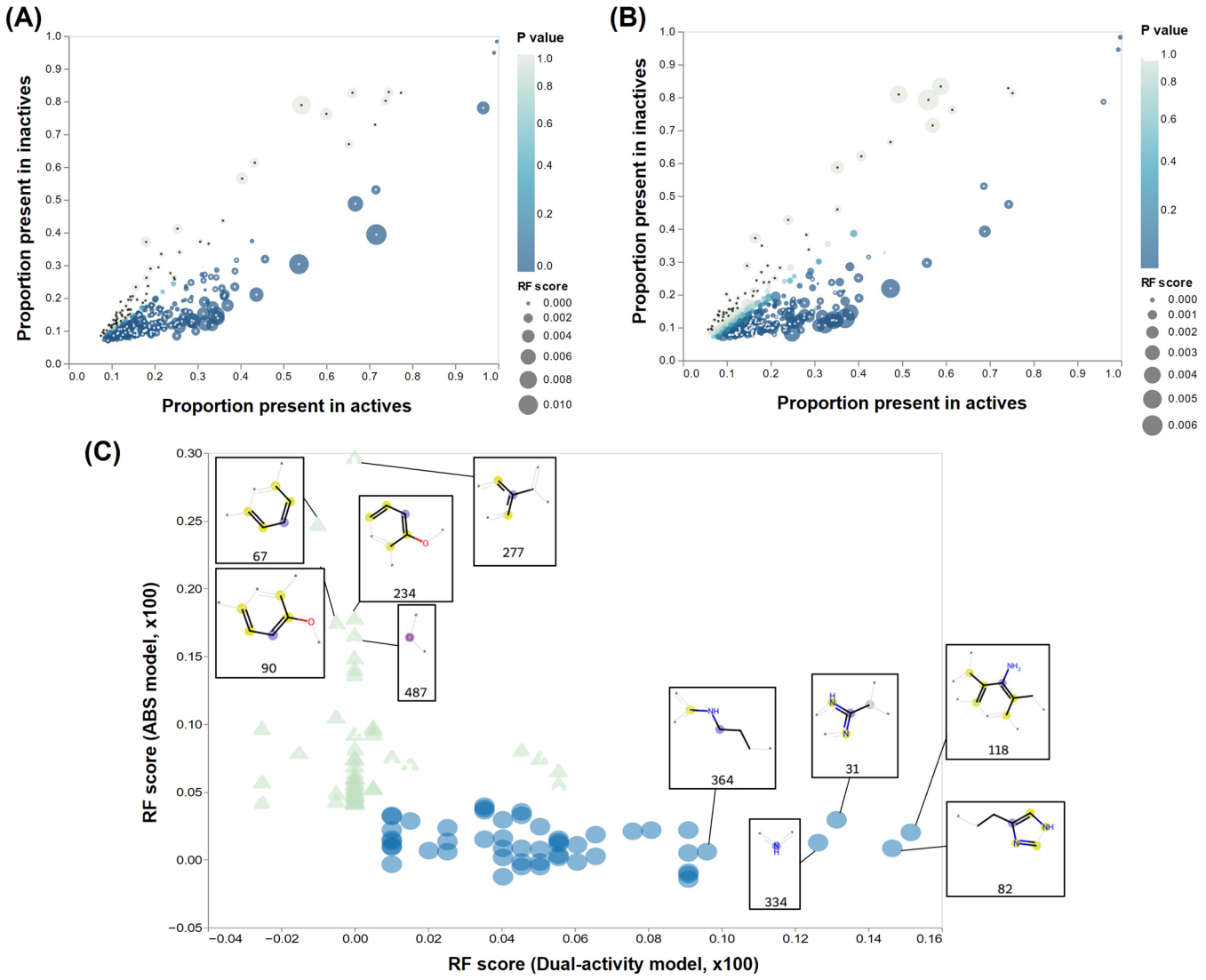


Figure 4. Enriched ECFP features within inactive and active compounds for stage-specific antiplasmodial action. (A) The proportion of active/inactive compounds against ABS containing a specific ECFP feature is plotted as circles, with the size of the circles corresponding to the RF permutation score of the ECFP feature. Enrichment of a feature toward active compounds compared to inactive compounds is indicated by the *p*-value color obtained from the *Z*-test on two proportions. The top 100 enriched ECFP features within active (white) and the top 67 enriched ECFP features within inactive (black) compounds were selected according to the RF score and *p*-value. (B) The proportion of dual-active/inactive compounds containing a specific ECFP feature is plotted as circles with the size corresponding to the RF permutation score of the ECFP feature. Enrichment of a feature toward dual-active compounds compared to inactive compounds is indicated by the *p*-value color. The top 100 enriched ECFP features within dual-active (white) and the top 52 enriched ECFP features within inactive (black) compounds were selected according to RF score and *p*-value. (C) Comparison of the unique ECFP features associated with activity against ABS (52) or dual stages (52). For the top unique ECFP features, structural elements are indicated, with all features summarized in File S1.

(Figure 2A). The SVM model maintained similar accuracy when predicting compound ABS inhibition activity on untrained imbalanced test data (ROC−AUC score of 0.92, Figure 2B), indicating no overfitting of the model to training data.

Additionally, the SVM models' recall ability (at 0.90) and precision (at 0.22) were comparable to those of the other ensemble models (RF, GMB; Figure 2B), with only LR models obtaining a higher recall (0.97) than SVM. However, it does this at the expense of the LR model's precision (0.08), resulting in low specificity with a higher false-positive rate (FPR: 0.43 vs <0.22 for the other models) (Figure 2B,C). This indicated that the model derived from SVM for ABS activity prediction is more precise and is better able to identify compounds with ABS activity while limiting its false-positive rate (FPR: 0.15) (Figure 2C), with losses in FPR associated with gains in precision to as high as 0.675 at higher probability thresholds (Table S2). While increasing the discrimination probability threshold does improve the models' precision (Figure S4), a threshold > 0.90 results in a drastic loss in recall (from 0.9 to <0.5) of the SVM model. Interestingly, SVM also showed similar, if not better, performance than that of more complex models such as NeuralNetFastAI (Table S3) and did much better at reducing its FPR compared to those of such complex models.

Like the data on models trained using ECFPs, SVM models trained with MACCS keys of compounds also achieved the highest ROC−AUC scores on both fivefold cross-validation (0.99 ± 0.02) (Figure S5) and untrained test data (0.92, Figure S5). No major differences could be detected between

https://doi.org/10.1021/acsomega.3c05664
ACS Omega 2023, 8, 43813−43826

the performance of models trained on ECFPs or MACCS keys beyond a slight improvement in performance metrics for LR and RF models (File S1).

3.3. Model Selection and Performance of Dual-Activity Prediction Models. To evaluate the performance of the models to predict compounds with dual activity, different metrics were used to ensure accurate evaluation of models within the constraints of the class-imbalanced dual-active database. Models were evaluated on their recall and precision in identifying dual-active compounds. SVM outperformed other models within fivefold cross-validation by obtaining ROC−AUC means > 0.96 (Figures 3A and S5). This extended to the performance of the models against imbalanced test data, where SVM reached a ROC−AUC score of 0.95 for models trained on ECFP (Figures 3B and S5) indicating no overfitting of models. Considering the large disparity in the number of actives vs inactive compounds for dual activity, the *F*1, *G*-mean and recall scores were also considered, with the SVM models identified as the best performing model trained on ECFP for dual-activity prediction with optimal *F*1 (0.26), *G*-mean (0.81), and recall (0.66) scores and low false-positive predictions (0.006) (Figure 3C). Higher probability thresholds (>0.9) again resulted in increased precision to 0.562 (Table S2, Figure S6) at the cost of drastically decreasing recall (<0.3) in identifying dual-active compound, still considerably less than that of RF models. Interestingly, MACCS generally resulted in a higher FPR than models trained on ECFP, indicating that MACCS is not as good at generating molecular descriptors of the chemical space for dual activity, whereas ECFP in comparison is more extensive and descriptive (File S1).

The high ROC−AUC and recall scores of the SVM models were on par with those obtained when more complex models were generated on the same data sets (Table S3). However, we observed that our simple models tended to outperform these more complex models with the SVM models having the lowest FPR, indicating that complex models tend to overfit the data and have poor generalization on these rather small data sets. In conclusion, SVM was the best model for predicting compounds with dual activity with low false-positive predictions on test data within representative chemical spaces.

3.4. Identification of Top Chemical Features for Activity Prediction. Considering the performance of models trained on ECFP as a molecular descriptor of chemical compounds' structural features, it was of interest to determine if there were any features that are particularly associated with antiplasmodial activity, or lack thereof. However, SVM models could not be used for feature importance analysis due to these models using an RBF data transformation technique which complicates attributing importance scores to chemical features. Therefore, RF was used because it was the second-best model in each instance, and due to the predicted probability scores on test sets for SVM and RF models having good correlation (Figure S7), hence these models could provide a feature importance score for particular chemical features. This indicated that a larger number of ECFP features were enriched within compounds with ABS activity (383) compared with inactive compounds (67) (Figure 4A). Recursive feature elimination tended to select for ECFP features enriched within inactive compounds (data not shown), and therefore an arbitrary selection of the top 100 features based on model importance scores were applied to provide a more comprehensive, inclusive data set (File S1). More than half of these features identified with the RF models were additionally confirmed by at least one other model (Table S5 and Figure S8). Importantly, there was no overlap between the top 100 features enriched for ABS activity and the features enriched in inactive compounds (Figure 4A and File S1), validating that a high level of specificity was obtained when ECFP features were associated with ABS activity. For the compounds with dual activity (Figure 4B), fewer ECFP features were enriched within active compounds compared to those observed in the ABS space (266 features compared to 383 for ABS activity). However, as with the features describing ABS activity, the features enriched in compounds with dual activity were specific, with no overlap present between the top 100 ECFP features for dual activity compared with the 52 features enriched in compounds that are inactive against gametocytes (Figure 4B).

As expected, around 50% of the top 100 features associated with activity were shared between compounds having ABS inhibition activity or dual activity. However, importantly, of the top 100 features enriched in the respective active fractions, 52 features each uniquely described features associated with either sole ABS activity or dual activity (Figure 4C and File S1). This provides clear distinction between chemical compositions of compounds required to kill multiple stages of the parasite compared to that required to only kill ABS parasites. Indeed, the top 5 unique features enriched in compounds able to target ABS parasites contain heterocyclic structures that are activated via oxygenation or alkylation, making the compound more reactive to electrophilic attack. Oxygenation may relate to a compound's bioavailability by increasing the compound's solubility and uptake.[40]

Comparatively, the top five unique features that are enriched in compounds with dual−stage activity are enriched for amine groups whether it is nitrogen-containing four-membered heterocyclic structures or the nitration of benzene and/or carbon groups. Such amine groups typically create localized electron deficient sites in the compound that allows interaction with cellular components such as amino acids and nucleic acids to be more favorable.[41] Hence, such amine groups may be more involved in the killing/drug effect as well as aid in the solubility of compounds.

Together with this, chemical features enriched in compounds that are inactive against ABS parasites (67) and gametocytes (52) tended to overlap with one another. Generally, these features indicate that features where nitrogen is unable to bind to hydrogen or structures containing amides and branching carbon chains may be associated with a lack in bioactivity (File S1).

3.5. Model Validation in Novel Chemical Spaces. To further validate and interrogate the extent to which the abovementioned models can predict stage-specific activity of antimalarial compounds, the SVM and RF models were exposed to previously unseen chemical matter and data from curated data sets. These data sets serve as an added level of interrogation to determine how well our models would perform when exposed to very diverse and novel chemical spaces and can give an indication of what to expect of these models in real-world application. We used data from the MMV PRB and Pathogen Box, which contained unique compounds not included in the data from the chemical libraries used to train the models, and with activity data available against both *P. falciparum* ABS parasites and gametocytes.[3,26] The compounds included in the boxes were also individually distinct and chemically diverse, providing more extreme data

https://doi.org/10.1021/acsomega.3c05664
ACS Omega 2023, 8, 43813−43826

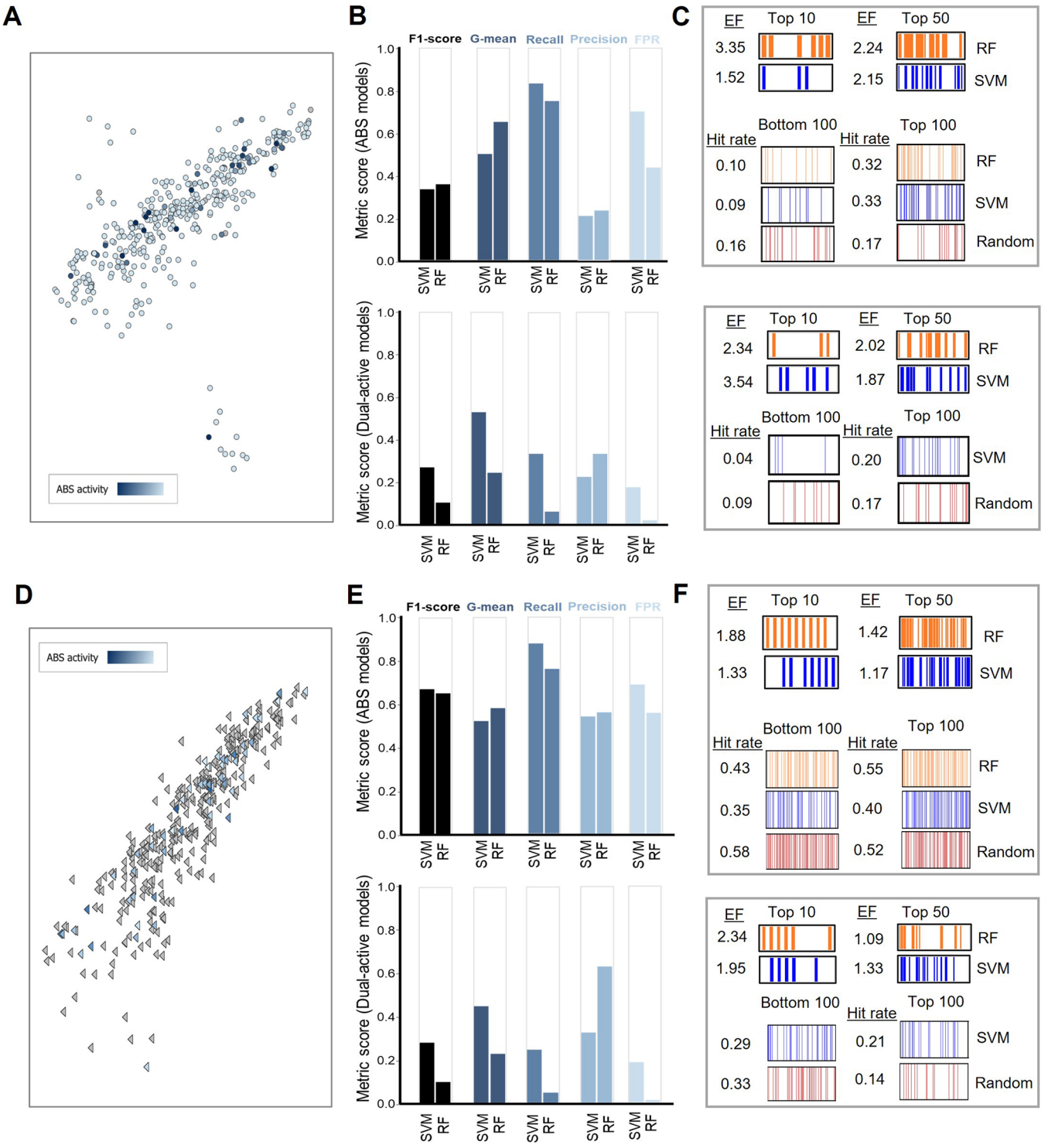


Figure 5. Performance of the top models against unseen chemical matter. To evaluate model robustness, models were exposed to extreme data sets from the PRB box (A) and Pathogen Box (D) that were individually distinct, chemically diverse (displayed within context of the launched drug chemical space (available on StarDrop v 7.3.0), with heatbars indicating potency) and had differential activity against ABS and gametocytes. ABS and dual-activity models trained on ECFP descriptors were evaluated for their activity predictions within the PRB box (B) and the Pathogen Box (E) for *F*1-scores (model performance exposed to imbalanced data) and *G*-mean scores (ability to optimize sensitivity and specificity), recall, precision, and false-positive rate (FPR). The hit rate of the best performing model for predicting ABS and/or dual activity within these chemical spaces (C,F) was compared to random selection. The enrichment factor (EF) of models was also calculated for the top 10 and top 50 compounds to determine how effective models were in prioritizing active compounds.

sets to evaluate the robust nature of the models, compared to the larger databases used to train the models, where structurally related compounds were present within a chemical space (Figure 5A,D). The best model must thus maintain fair accuracy and recall under these conditions, while optimizing sensitivity and specificity in predictions to limit the models'

https://doi.org/10.1021/acsomega.3c05664
ACS Omega 2023, 8, 43813−43826

FPR for these novel and diverse chemical matter. Class imbalance was not corrected for these data sets to evaluate the performance of the models, with hit rates against ABS at 18% for the PRB box[3] and 31% for the Pathogen Box[26] (Figure 5A,D). For compounds with gametocytocidal activity, the hit rate for the PRB box (∼13%) and Pathogen Box (∼24%) was even lower.

For the two top-performing ABS activity models (SVM and RF), both models obtained similar scores for most metrics in the PRB box (Figure 5B). However, the enrichment factor of the top 10 (3.35 vs 1.52) and top 50 (2.24 vs 2.15) compounds, RF clearly outperforms the SVM (Figure 5C). Both models were able to show a significant enrichment of hits (hit rates at >30%) in the top 100 compounds, compared to random selection (17%) and the hit rate of the PRB box itself (18%) (Figure 5C). The specificity is also confirmed, as these models enable selection of compounds that would be inactive in the bottom 100. Within this chemical space, the models differentiated from each other when predicting activity for dual-active compounds, with the SVM model outperforming on all metrics, excluding precision (Figure 5B). Although precision of the models tends to be low, closer inspection of compounds within the PRB box compared to the training set (Figure S9) revealed that the chemical similarity distribution is lower than that of test set compounds, which may then fall outside of the applicability domain of the model, hence lowering the precision in correctly identifying active compounds (Figure S9, Table S6). Although SVM tended to have a higher FPR (0.17 vs 0.02), RF failed to identify dual-active compounds with a recall below <0.20, and SVM tends to shift and prioritize dual-active compounds to the top of the list more than RF for both the top 10 (3.54 vs 2.34) and top 50 (1.87 vs 2.02) compounds (Figure 5C). Together with this, the hit rate of the top 100 hits from the SVM model (20%) exceeded the hit rate of random selection (17%) and the PRB box (13%) (Figure 5C).

Within a different novel chemical space (Pathogen Box, Figure 5D), both SVM and RF ABS activity models obtained higher *F*1-scores (Figure 5E), compared to the PRB box (Figure 5B) and this can be attributed to the higher hit rate within the Pathogen Box resulting in a less severe class imbalance compared to the PRB box. Model performance was maintained (Figure 5E) as also evident in enrichment of hits (EF > 1.3) for predicting ABS activity as well as dual-active compounds (Figure 5F). With the Pathogen Box, the dual-activity models maintained similar *F*1-scores when predicting dual-active compounds, indicating that despite the difference in hit rates between the PRB and Pathogen Box, the dual-active models are not influenced by class imbalance, which is highly beneficial. Although the dual-activity model's performance is lower compared to the test set, this was expected considering the diverse data the models were exposed to and the limited chemical space data available for training compared to the ABS database.

4. DISCUSSION

The identification of gametocytocidal compounds remains challenging compared to the identification of compounds with activity against ABS parasites, requiring alternative approaches to streamline screening and compound optimization campaigns. To our knowledge, no study has fully explored the chemical space information on phenotypic screens conducted against gametocytes to help in predicting gametocytocidal compounds, although this has been done for asexual phenotypic screens via DeepMalaria and MAIB[18,20] as well as liver-stage phenotypic screens[19] (Table S7). Here, we were able to successfully train models capable of predicting compounds with activity against both ABS parasites and gametocytes with good accuracy, precision, and recall.

Our best performing SVM and RF models for ABS inhibition activity prediction were on par with DeepMalaria[18] that used neural−network-based models, with SVM obtaining higher recall abilities (95 vs 87.75%) but has slightly less precision (2.40 vs 3.54%) on untrained data. This could be because the chemical space information we used for training was larger as we incorporated more chemical libraries screened against ABS. Additionally, the cluster-based undersampling of inactive compounds as used here seems to be a more important strategy compared to the oversampling of active compounds used before. One disadvantage of oversampling can be that the model fails to generalize and recognize patterns in novel active compounds due to the model being fixated on patterns associated with the oversampled active compounds as a result of such features being over represented in the training data.[39] Our data suggest that more simplified algorithms, such as those used here, together with strict attention to class bias, could be more informative in building ML models for the antimalarial chemical space, negating the use of more complex approaches using neural network analysis. Similar observations have been seen in different fields involving image classification when comparing deep learning to simpler machine learning methods such as SVM.[42] Single classifiers such as SVM outperformed ensemble models within representative chemical spaces in our data, which could be considered smaller than the typical big data on which ensemble methods such as GBM and RF generally performs better. Alternatively, the fact that SVM algorithms balances the contributions of chemical features in correct predictions could be more important in the antimalarial drug space compared to the way ensemble methods rely mostly only on the presence/absence of such features.[43]

Although the dual-active models within this study had low precision, the high recall is desirable despite the identification of false positives and is driven by the need to identify dual-active compounds. This is especially true since these models are capable of sorting dual-active compounds to the forefront of potential candidate lists, as observed by the enrichment factor of the models.

Within this study, we also identified chemical features enriched within dual-active compounds. Interestingly, ECFP molecular featurization enabled models to perform better within novel and diverse antimalarial chemical spaces compared with MACCS keys. This could be attributed to ECFP capturing the atom environments around each atom, whereas MACCS summarizes the presence/absence of chemical groups within compounds that have been identified as important within drug discovery and virtual screening[44] with both these featurizations incorporated into tools such as Chemical Checker.[45] Hence, ECFP due to them being more descriptive may be better at capturing the chemical space and translating this over to models, whereby such models can then identify chemical features important for bioactivity as well as a lack in bioactivity. The most notable commonality among enriched features was the nitration of heterocyclic and carbon groups that may aid in interactions with cellular targets.[41] Similarly for features enriched within ABS-inhibiting compounds and important for predicting ABS inhibition, oxygen-

https://doi.org/10.1021/acsomega.3c05664
ACS Omega 2023, 8, 43813−43826

ation was more common and may relate to solubility and drug uptake. This study therefore contributes to the identification for chemical features important for stage-specific activity, and we anticipate that inclusion of chemical features (or exclusion of those features associated with inactivity) will aid medicinal chemists in guiding compound derivatization during hit-to-lead optimization of stage-specific and/or dual-active candidates. However, it must be noted that the features described here lack the context of connectivity, and it is more than likely that a combination of such features is important for stage-specific activity. Our models also depend on 2D fingerprints, which are not sensitive to the stereochemistry of compounds. The inclusion of additional biological and molecular descriptors (such as 3D fingerprints to capture stereochemistry) with physicochemical properties[45] may improve model performance in the future and allow better understand how chemical features contribute to stage-specific activity of compounds, possibly deconvoluting mode of action. Alternatively, the capabilities of generative models may be useful to artificially expand this limited chemical space by creating de novo chemical starting points with a high probability of having desirable gametocytocidal activity and drug-like properties.[46]

5. CONCLUSIONS

In summary, this study has come up with a new ML tool to guide and optimize both phenotypic screens as well as compound derivatization in hit-to-lead and lead optimization campaigns, to reduce time and cost allocated toward compounds with low probability of having dual activity against both ABS parasites and gametocytes. Considering such models have good understanding of the chemical space for bioactivity, we were successful in mining for features related to bioactivity and highlight chemical features predictive toward ABS and/or dual activity. The tools provided by this study can accelerate the identification and hit-to-lead optimization of dual-active compounds to aid in malaria elimination strategies. These models are deployed within the Ersilia Model Hub repository[47] for open-source access, particularly aiding infectious disease drug discovery.

■ ASSOCIATED CONTENT

Data Availability Statement

Code Availability Statement: All python scripts for clustering, undersampling and model building as well as evaluation can be obtained from github: http://github.com/M2PL/Machines-Against-Malaria. To facilitate model usage, we have also incorporated the models in the Ersilia Model Hub (https://www.ersilia.io/model-hub; identifier eos80ch).

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c05664.

Information on hyperparameters used for model building and additional performance metrics of model predictions in representative and novel chemical spaces (PDF)

Chemical information on compounds within the databases used for ML as well as performance metrics of models trained on imbalanced/oversampled/undersampled data using either ECFP or MACCS molecular fingerprints as well as information on enriched ECFP features for activity/inactivity against ABS and/or gametocytes. SMILES contains Simplified Molecular Input Line Entry System (SMILES) of compounds used for machine learning (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Lyn-Marié Birkholtz − *Department of Biochemistry, Genetics and Microbiology, Institute for Sustainable Malaria Control, University of Pretoria, Hatfield 0028, South Africa;* orcid.org/0000-0001-5888-2905; Email: lbirkholtz@up.ac.za

Authors

Ashleigh van Heerden − *Department of Biochemistry, Genetics and Microbiology, Institute for Sustainable Malaria Control, University of Pretoria, Hatfield 0028, South Africa*

Gemma Turon − *Ersilia Open Source Initiative, Cambridge CB1 3DE, U.K.*

Miquel Duran-Frigola − *Ersilia Open Source Initiative, Cambridge CB1 3DE, U.K.;* orcid.org/0000-0002-9906-6936

Nelishia Pillay − *Department of Computer Science, University of Pretoria, Hatfield 0028, South Africa*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c05664

Author Contributions

AvH performed the research with inputs from MDF, GT, and NP. LMB conceptualized the study and wrote the paper with AvH. All coauthors approved the final version of the paper.

Funding

This work was supported by the South African Department of Science and Innovation and National Research Foundation South African Research Chairs Initiative Grant (LMB UID: 84627).

Funding

The data sets used in this study can be found at ChEMBL using ChEMBL ID provided in the File S1.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Jason Hlozeck from the University of Cape Town for the useful discussions and proofreading of the paper. Additionally, the authors acknowledge Google Colab for their cloud-based service that allowed us to build machine learning models.

■ ABBREVIATIONS

ABS, asexual blood stages; ACT, artemisinin combination therapy; FPR, false-positivity rate; GBM, gradient boosting machines; *G*-mean, geometric mean; LR, logistic regression; MACCS, molecular access keys; ECFP, Morgan fingerprint; ML, machine learning; PRB, pandemic response box; SVM, support vector machines; RBF, radial basis function; RF, random forests; ROC−AUC, receiver operator characteristic curve; TrB, transmission blocking

■ REFERENCES

(1) World Health Organization. *World Malaria Report 2020:20 Years of Global Progress and Challenges*; World Health Organization Geneva, 2020; pp 1−151.

(2) Ataba, E.; Dorkenoo, A. M.; Nguepou, C. T.; Bakai, T.; Tchadjobo, T.; Kadzahlo, K. D.; Yakpa, K.; Atcha-Oubou, T. Potential Emergence of Plasmodium Resistance to Artemisinin Induced by the Use of Artemisia annua for Malaria and COVID-19 Prevention in Sub-African Region. *Acta Parasitol.* **2022**, *67* (1), 55−60.
(3) Reader, J.; van der Watt, M. E.; Taylor, D.; Le Manach, C.; Mittal, N.; Ottilie, S.; Theron, A.; Moyo, P.; Erlank, E.; Nardini, L.; et al. Multistage and transmission-blocking targeted antimalarials discovered from the open-source MMV Pandemic Response Box. *Nat. Commun.* **2021**, *12* (1), 269.
(4) Yang, T.; Ottilie, S.; Istvan, E. S.; Godinez-Macias, K. P.; Lukens, A. K.; Baragana, B.; Campo, B.; Walpole, C.; Niles, J. C.; Chibale, K.; et al. MalDA, Accelerating Malaria Drug Discovery. *Trends Parasitol.* **2021**, *37* (6), 493−507.
(5) Birkholtz, L. M.; Alano, P.; Leroy, D. Transmission-blocking drugs for malaria elimination. *Trends Parasitol.* **2022**, *38* (5), 390−403.
(6) Josling, G. A.; Llinas, M. Sexual development in Plasmodium parasites: knowing when it's time to commit. *Nat. Rev. Microbiol.* **2015**, *13* (9), 573−587.
(7) Birkholtz, L. M.; Coetzer, T. L.; Mancama, D.; Leroy, D.; Alano, P. Discovering New Transmission-Blocking Antimalarial Compounds: Challenges and Opportunities. *Trends Parasitol.* **2016**, *32* (9), 669−681.
(8) (a) Reader, J.; Botha, M.; Theron, A.; Lauterbach, S. B.; Rossouw, C.; Engelbrecht, D.; Wepener, M.; Smit, A.; Leroy, D.; Mancama, D.; et al. Nowhere to hide: interrogating different metabolic parameters of Plasmodium falciparum gametocytes in a transmission blocking drug discovery pipeline towards malaria elimination. *Malar. J.* **2015**, *14*, 213. (b) van Biljon, R.; van Wyk, R.; Painter, H. J.; Orchard, L.; Reader, J.; Niemand, J.; Llinas, M.; Birkholtz, L. M. Hierarchical transcriptional control regulates Plasmodium falciparum sexual differentiation. *BMC Genomics* **2019**, *20* (1), 920. (c) van der Watt, M. E.; Reader, J.; Birkholtz, L. M. Adapt or Die: Targeting Unique Transmission-Stage Biology for Malaria Elimination. *Front. Cell. Infect. Microbiol.* **2022**, *12*, 901971.
(9) Gamo, F.-J.; Sanz, L. M.; Vidal, J.; de Cozar, C.; Alvarez, E.; Lavandera, J.-L.; Vanderwall, D. E.; Green, D. V. S.; Kumar, V.; Hasan, S.; et al. Thousands of chemical starting points for antimalarial lead identification. *Nature* **2010**, *465*, 305−310.
(10) Almela, M. J.; Lozano, S.; Lelievre, J.; Colmenarejo, G.; Coteron, J. M.; Rodrigues, J.; Gonzalez, C.; Herreros, E. A new set of chemical starting points with Plasmodium falciparum transmission-blocking potential for antimalarial drug discovery. *PLoS One* **2015**, *10*, No. e0135139. (a) Guiguemde, W. A.; Shelat, A. A.; Garcia-Bustos, J. F.; Diagana, T. T.; Gamo, F. J.; Guy, R. K. Global phenotypic screening for antimalarials. *Chem. Biol.* **2012**, *19* (1), 116−129. (b) Lucantoni, L.; Duffy, S.; Adjalley, S. H.; Fidock, D. A.; Avery, V. M. Identification of MMV malaria box inhibitors of Plasmodium falciparum early-stage gametocytes using a luciferase-based high-throughput assay. *Antimicrob. Agents Chemother.* **2013**, *57* (12), 6050−6062.
(11) Guiguemde, W. A.; Shelat, A. A.; Bouck, D.; Duffy, S.; Crowther, G. J.; Davis, P. H.; Smithson, D. C.; Connelly, M.; Clark, J.; Zhu, F.; et al. Chemical genetics of Plasmodium falciparum. *Nature* **2010**, *465* (7296), 311−315.
(12) Abraham, M.; Gagaring, K.; Martino, M. L.; Vanaerschot, M.; Plouffe, D. M.; Calla, J.; Godinez-Macias, K. P.; Du, A. Y.; Wree, M.; Antonova-Koch, Y.; et al. Probing the Open Global Health Chemical Diversity Library for Multistage-Active Starting Points for Next-Generation Antimalarials. *ACS Infect. Dis.* **2020**, *6* (4), 613−628.
(13) Plouffe, D.; Brinker, A.; McNamara, C.; Henson, K.; Kato, N.; Kuhen, K.; Nagle, A.; Adrian, F.; Matzen, J. T.; 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* (26), 9059−9064.
(14) Duffy, S.; Avery, V. M. Identification of inhibitors of Plasmodium falciparum gametocyte development. *Malar. J.* **2013**, *12*, 408.
(15) (a) Zhou, L.; Pan, S.; Wang, J.; Vasilakos, A. V. Machine learning on big data: Opportunities and challenges. *Neurocomputing* **2017**, *237*, 350−361. (b) Silva, J. C. F.; Teixeira, R. M.; Silva, F. F.; Brommonschenkel, S. H.; Fontes, E. P. B. Machine learning approaches and their current application in plant molecular biology: A systematic review. *Plant Sci.* **2019**, *284*, 37−47.
(16) Rifaioglu, A. S.; Atas, H.; Martin, M. J.; Cetin-Atalay, R.; Atalay, V.; Doğan, T. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Briefings Bioinf.* **2019**, *20* (5), 1878−1912.
(17) (a) 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. *Nat. Rev. Drug Discovery* **2019**, *18* (6), 463−477. (b) You, Y.; Lai, X.; Pan, Y.; Zheng, H.; Vera, J.; Liu, S.; Deng, S.; Zhang, L. Artificial intelligence in cancer target identification and drug discovery. *Signal Transduction Targeted Ther.* **2022**, *7* (1), 156. (c) Arul Murugan, N.; Ruba Priya, G.; Narahari Sastry, G.; Markidis, S. Artificial intelligence in virtual screening: Models versus experiments. *Drug Discovery Today* **2022**, *27* (7), 1913−1923. (d) Ferreira, L. L.; Andricopulo, A. D. From chemoinformatics to deep learning: an open road to drug discovery. *Future Med. Chem.* **2019**, *11* (5), 371−374. (e) Tan, A. C.; Gilbert, D. Ensemble machine learning on gene expression data for cancer classification. *Appl. Bioinf.* **2003**, *2*, S75−S83.
(18) Keshavarzi Arshadi, A.; Salem, M.; Collins, J.; Yuan, J. S.; Chakrabarti, D. DeepMalaria: Artificial Intelligence Driven Discovery of Potent Antiplasmodials. *Front. Pharmacol* **2020**, *10*, 1526.
(19) Mughal, H.; Bell, E. C.; Mughal, K.; Derbyshire, E. R.; Freundlich, J. S. Random Forest Model Predictions Afford Dual-Stage Antimalarial Agents. *ACS Infect. Dis.* **2022**, *8* (8), 1553−1562.
(20) Bosc, N.; Felix, E.; Arcila, R.; Mendez, D.; Saunders, M. R.; Green, D. V. S.; Ochoada, J.; Shelat, A. A.; Martin, E. J.; Iyer, P.; et al. MAIP: a web service for predicting blood-stage malaria inhibitors. *J. Cheminf.* **2021**, *13* (1), 13.
(21) Bertoni, M.; Duran-Frigola, M.; Badia-i-Mompel, P.; Pauls, E.; Orozco-Ruiz, M.; Guitart-Pla, O.; Alcalde, V.; Diaz, V. M.; Berenguer-Llergo, A.; Brun-Heath, I.; et al. Bioactivity descriptors for uncharacterized chemical compounds. *Nat. Commun.* **2021**, *12* (1), 3932.
(22) Jasial, S.; Hu, Y.; Vogt, M.; Bajorath, J. Activity-relevant similarity values for fingerprints and implications for similarity searching. *F1000Research* **2016**, *5*, 591.
(23) Van Voorhis, W. C.; Adams, J. H.; Adelfio, R.; Ahyong, V.; Akabas, M. H.; Alano, P.; Alday, A.; Aleman Resto, Y.; Alsibaee, A.; Alzualde, A.; et al. Open Source Drug Discovery with the Malaria Box Compound Collection for Neglected Diseases and Beyond. *PLoS Pathog.* **2016**, *12* (7), No. e1005763.
(24) Miguel-Blanco, C.; Molina, I.; Bardera, A. I.; Díaz, B.; de Las Heras, L.; Lozano, S.; González, C.; Rodrigues, J.; Delves, M. J.; Ruecker, A.; et al. Hundreds of dual-stage antimalarial molecules discovered by a functional gametocyte screen. *Nat. Commun.* **2017**, *8*, 15160.
(25) McInnes, L.; Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. **2018**, arXiv:1802.03426v3.
(26) Duffy, S.; Sykes, M. L.; Jones, A. J.; Shelper, T. B.; Simpson, M.; Lang, R.; Poulsen, S. A.; Sleebs, B. E.; Avery, V. M. Screening the Medicines for Malaria Venture Pathogen Box across Multiple Pathogens Reclassifies Starting Points for Open-Source Drug Discovery. *Antimicrob. Agents Chemother.* **2017**, *61* (9), No. e00379.
(27) *RDKit: Open-Source Cheminformatics*, 2006. https://www.rdkit.org.
(28) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321−357.

https://doi.org/10.1021/acsomega.3c05664
ACS Omega 2023, 8, 43813−43826

(29) Ramsundar, B.; Eastman, P.; Walters, P.; Pande, V. *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*; O'Reilly Media, 2019.
(30) Cereto-Massague, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallve, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58−63.
(31) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(32) (a) More, A. S.; Rana, D. P. Review of random forest classification techniques to resolve data imbalance. *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)*, 2017; pp 72−78.. (b) Brown, I.; Mues, C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Syst. Appl.* **2012**, *39* (3), 3446−3453.
(33) Korkmaz, S. Deep Learning-Based Imbalanced Data Classification for Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60* (9), 4180−4190.
(34) Esposito, C.; Landrum, G. A.; Schneider, N.; Stiefl, N.; Riniker, S. GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning. *J. Chem. Inf. Model.* **2021**, *61* (6), 2623−2640.
(35) Bengfort, B.; Bilbro, R. Yellowbrick: Visualizing the scikit-learn model selection process. *J. Open Source Softw.* **2019**, *4* (35), 1075.
(36) Fakoor, R.; Mueller, J. W.; Erickson, N.; Chaudhari, P.; Smola, A. J. Fast, accurate, and simple models for tabular data via augmented distillation. *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020; pp 8671−8681.
(37) Park, S.; Park, H. Combined oversampling and undersampling method based on slow-start algorithm for imbalanced network traffic. *Computing* **2021**, *103* (3), 401−424.
(38) Krawczyk, B. Learning from imbalanced data: open challenges and future directions. *Prog. Artif. Intell.* **2016**, *5* (4), 221−232.
(39) (a) Zhang, C.; Gao, W.; Song, J.; Jiang, J. An imbalanced data classification algorithm of improved autoencoder neural network. *2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)*, 2016; pp 95−99.. (b) Haibo, H.; Garcia, E. A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21* (9), 1263−1284.
(40) Karich, A.; Kluge, M.; Ullrich, R.; Hofrichter, M. Benzene oxygenation and oxidation by the peroxygenase of Agrocybe aegerita. *AMB Express* **2013**, *3* (1), 5.
(41) Nepali, K.; Lee, H.-Y.; Liou, J.-P. Nitro-Group-Containing Drugs. *J. Med. Chem.* **2019**, *62* (6), 2851−2893.
(42) (a) Sakr, G. E.; Mokbel, M.; Darwich, A.; Khneisser, M. N.; Hadi, A. Comparing deep learning and support vector machines for autonomous waste sorting. *2016 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET)*, 2016; pp 207−212.. (b) Liu, P.; Choo, K.-K. R.; Wang, L.; Huang, F. SVM or deep learning? A comparative study on remote sensing image classification. *Soft Comput.* **2017**, *21* (23), 7053−7065.
(43) Siemers, F. M.; Bajorath, J. Differences in learning characteristics between support vector machine and random forest models for compound classification revealed by Shapley value analysis. *Sci. Rep.* **2023**, *13* (1), 5983.
(44) Capecchi, A.; Probst, D.; Reymond, J. L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminf.* **2020**, *12* (1), 43.
(45) Duran-Frigola, M.; Pauls, E.; Guitart-Pla, O.; Bertoni, M.; Alcalde, V.; Amat, D.; Juan-Blanco, T.; Aloy, P. Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker. *Nat. Biotechnol.* **2020**, *38* (9), 1087−1096.
(46) Zeng, X.; Wang, F.; Luo, Y.; Kang, S.-g.; Tang, J.; Lightstone, F. C.; Fang, E. F.; Cornell, W.; Nussinov, R.; Cheng, F. Deep generative molecular design reshapes drug discovery. *Cell Rep. Med.* **2022**, *3* (12), 100794.
(47) Turon, G.; Duran-Frigola, M. *Ersilia Model Hub: A Repository of AI/ML Models for Neglected Tropical Diseases (v0.1.16)*, 2023..

https://doi.org/10.1021/acsomega.3c05664
ACS Omega 2023, 8, 43813−43826