Supplementary information

Machine learning approaches identify chemical features for stage-specific antimalarial compounds

Ashleigh van Heerden[1], Gemma Turon[2], Miquel Duran-Frigola[2], Nelishia Pillay[3], Lyn-Marié Birkholtz[1]*

[1] Department of Biochemistry, Genetics and Microbiology, Institute for Sustainable Malaria Control, University of Pretoria, Private Bag X20, Hatfield, 0028, South Africa

[2] Ersilia Open Source Initiative, 28 Belgrave Road, CB1 3DE, Cambridge, United Kingdom

[3] Department of Computer Science, University of Pretoria, Private Bag X20, Hatfield, 0028, South Africa

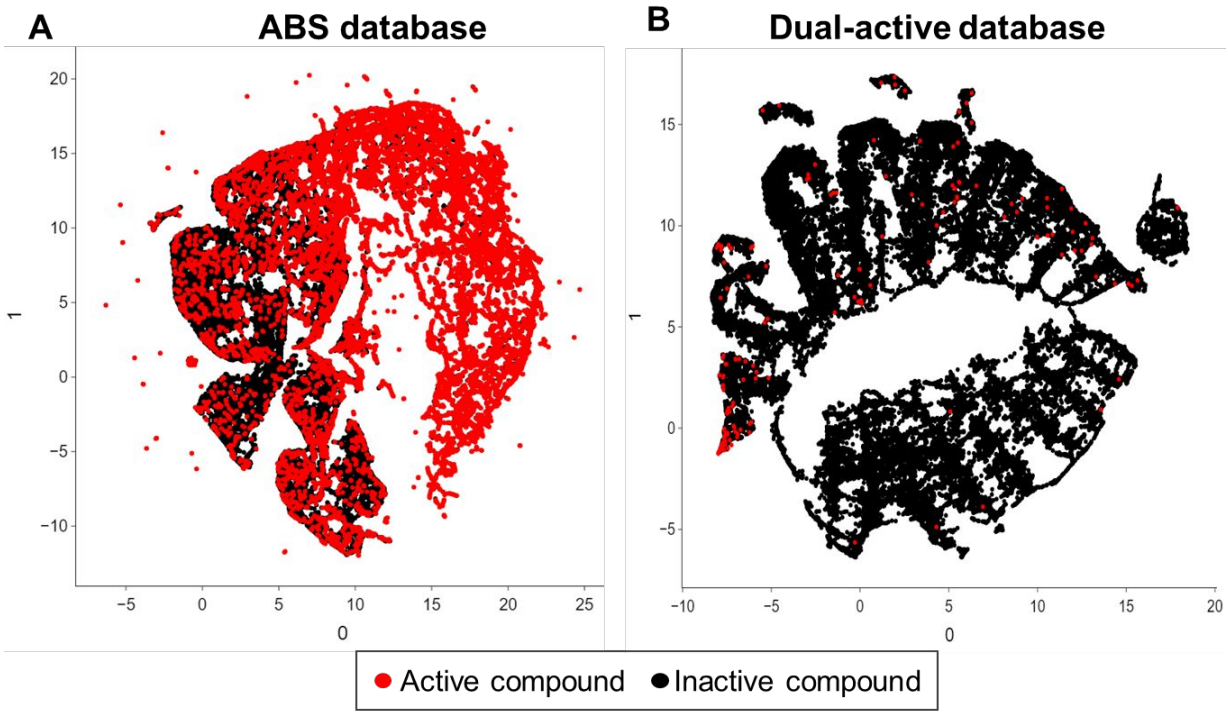

Figure S1: Outlier detection in ABS and dual-active database via UMAP.
(A) Chemical space of active (red) and inactive (black) compounds from the ABS database after pre-processing was visualized using UMAPs using Morgan fingerprints (500-bit length) and physiochemical descriptors predicted via RDKit. (B) Similarly, the chemical space of active (red) and inactive (black) compounds from the dual-active database after pre-processing was also visualized using UMAPs using Morgan fingerprints and the same physiochemical descriptors predicted via RDKit. From the chemical space projection of the databases, no outliers were detected. Physiochemical descriptors included molecular weight; log P; number of H donors/acceptors; number of rotatable bonds; TPSA; ring count; number of heteroatoms; aromatic bonds; acidic groups and basic groups.

The optimal bit-length for ABS activity prediction models was determined to be 500-bit as there appeared to be no change in ROC-AUC values as bit length was increased, whereas at a drastic drop in FPR was observed at 500-bit which then again increased as the bit-length was increased. In contrast to FPR, the opposite trend was observed for precision. For Dual activity prediction models, ROC-AUC values slightly increased as bit length increased with GBM decreasing in ROC-AUC once larger than 500-bit length was used. FPR values decreased with increasing bit length, however models plateaued at 300-bit length (excluding LR). Precision also tended to increase with increasing bit-length however, at 500-bits only slight increases was observed at 500-bits. Based on ROC-AUC, FPR and precision metrics as well as the training time required to build models 500-bits were determined to be the best bit length to use for ECFP whilst training most models. Using 500-bits, the optimal atom radius for ABS and dual-activity prediction models was determined to a radius of 5 was optimal as this increased the precision as well as G-mean scores of the models, whilst not impacting ROC-AUC values. Such an atom radius would also aid in the identification of substructures during feature analysis, hence a bit length of 500 with an atom radius of 5 was selected when generating Morgan fingerprints of compounds.

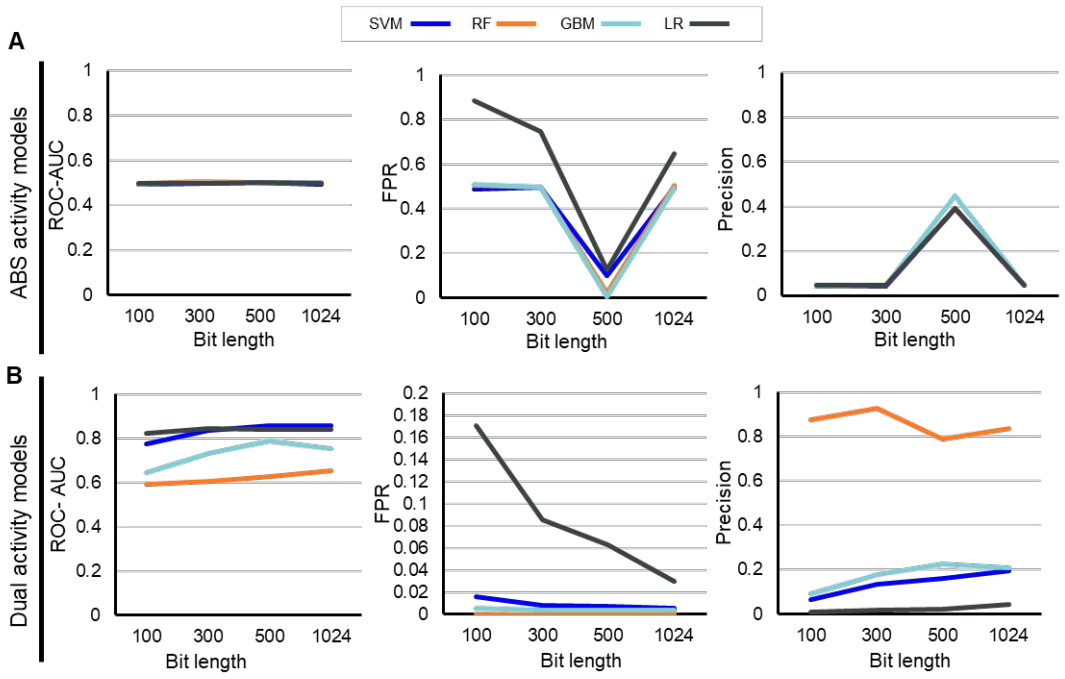

Figure S2: Optimal Morgan fingerprints (ECFP) bit length that enabled better model performance.
(A) Performance of ABS activity models using different ML algorithms (blue = Support vector machine, orange = Random forest, light blue= gradient boosting machine, dark grey= logistic regression) in identifying compounds with ABS inhibition activity within test set using differing bit lengths of Morgan fingerprints (ECFP) during training. ROC AUC scores indicate the classifier's ability in distinguishing active and inactive compounds against ABS. FPR scores indicate false positive rate of test set predictions. (B) Performance of dual active models (trained on data with class imbalance) using different ML algorithm's (blue = Support vector machine, orange = Random forest, light blue= gradient boosting machine, dark grey= logistic regression) ability in identifying compounds with dual activity within test set using differing bit lengths of Morgan fingerprints (ECFP) during training.

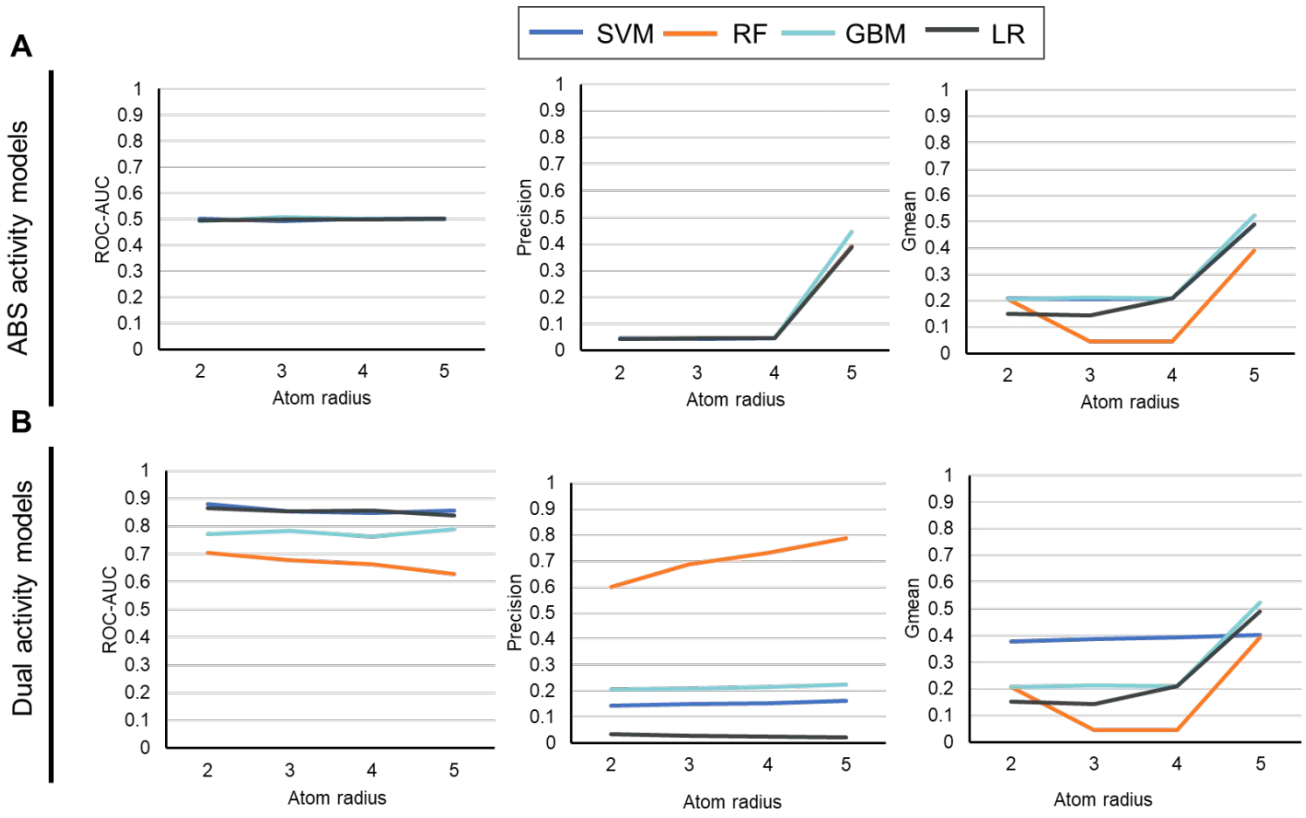

Figure S3: Optimal Morgan fingerprints (ECFP) atom radius that enabled better model performance.
(A) Performance of ABS activity models using different ML algorithms (blue = Support vector machine, orange = Random forest, light blue= gradient boosting machine, dark grey= logistic regression) in identifying compounds with ABS inhibition activity within test set using differing atom radius of Morgan fingerprints (ECFP) at 500-bit length during training. ROC AUC scores indicate the classifier's ability in distinguishing active and inactive compounds against ABS. FPR scores indicate false positive rate of test set predictions. (B) Performance of dual active models (trained on data with class imbalance) using different ML algorithm's (blue = Support vector machine, orange = Random forest, light blue= gradient boosting machine, dark grey= logistic regression) ability in identifying compounds with dual activity within test set using differing atom radius of ECFP at 500-bit length during training.

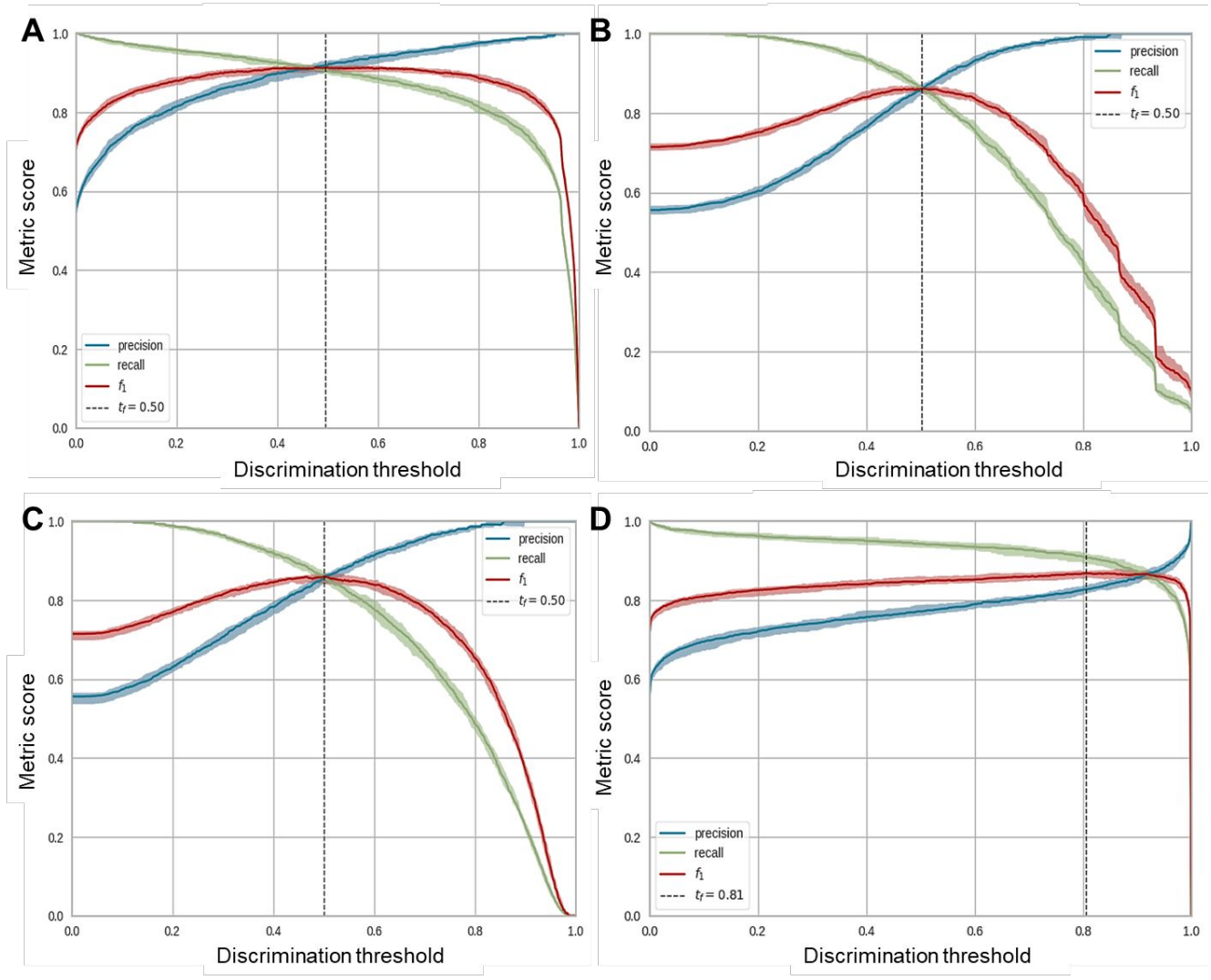

Figure S4: Influence of discrimination threshold shift on ABS model performance within the untrained test set. (A) SVM, (B) RF, (C) GBM and (D) LR model performance regarding precision, recall and f1-score was calculated and plotted for each threshold defining active and inactive compounds from the predicted probability, i.e., discrimination threshold. Discrimination threshold adjustment was conducted on test set data with ABS activity models. T_f indicated the threshold at which both recall and precision was the highest.

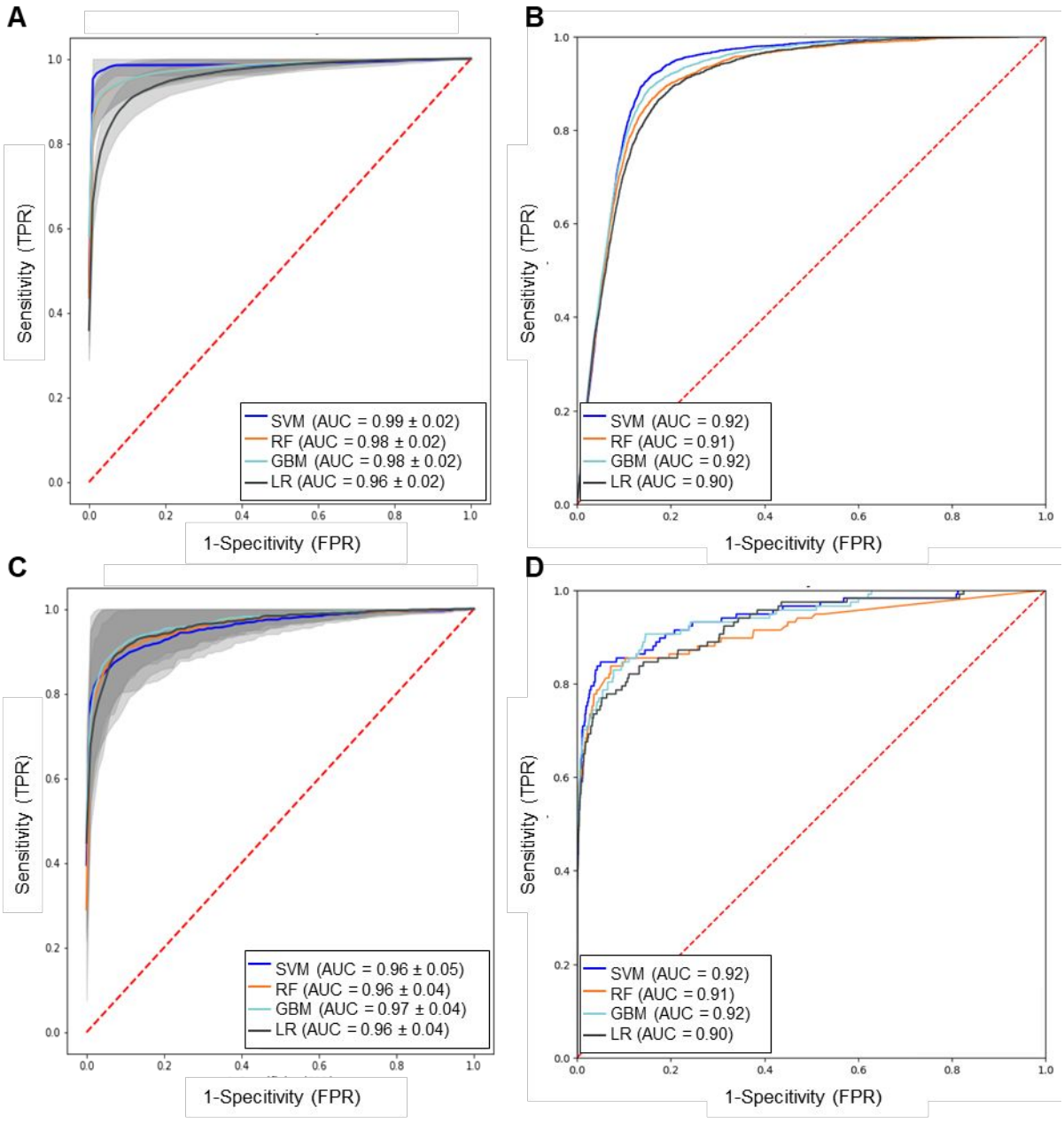

Figure S5: ROC-AUC curves of ABS and dual-activity MACCS models on 5-fold cross-validation and untrained test set.

ROC-AUC curves showing performance of different ML algorithms in predicting compounds with ABS (A) or (C) dual-activity when trained on MACCS keys of compounds after 5-fold cross-validation. Insert indicates AUC mean values ± standard deviation. The ROC-AUC curves of the different models trained on MACCS descriptors on untrained test set in predicting ABS (B) or dual-activity (D).

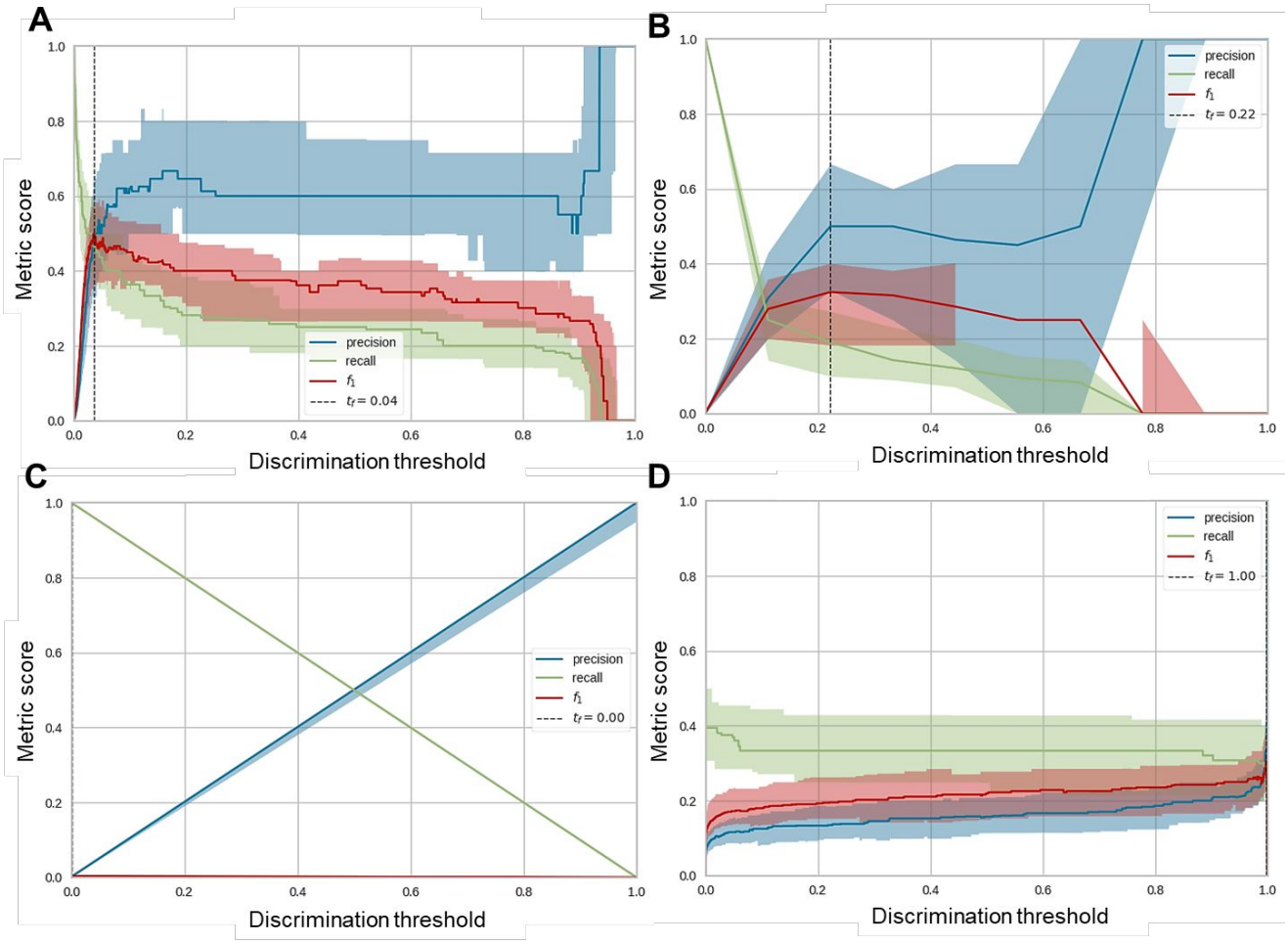

Figure S6: Discrimination threshold shift impact on the dual-activity model performance within the untrained test set.

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

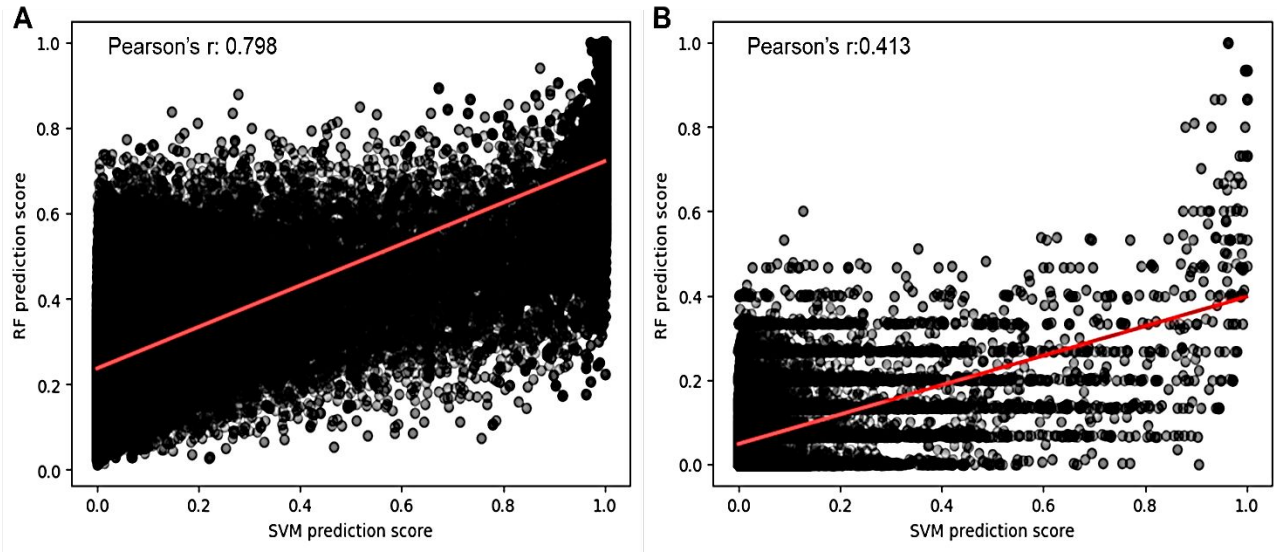

Figure S7: Correlation of predicted probability scores between RF and SVM on test set.
(A) Pearson correlation of predicted probability scores for ABS activity prediction SVM and RF models (trained on ECFP) on the test set. (B) Pearson correlation of predicted probability scores for dual-activity prediction SVM and RF models (trained on ECFP) on the test set.

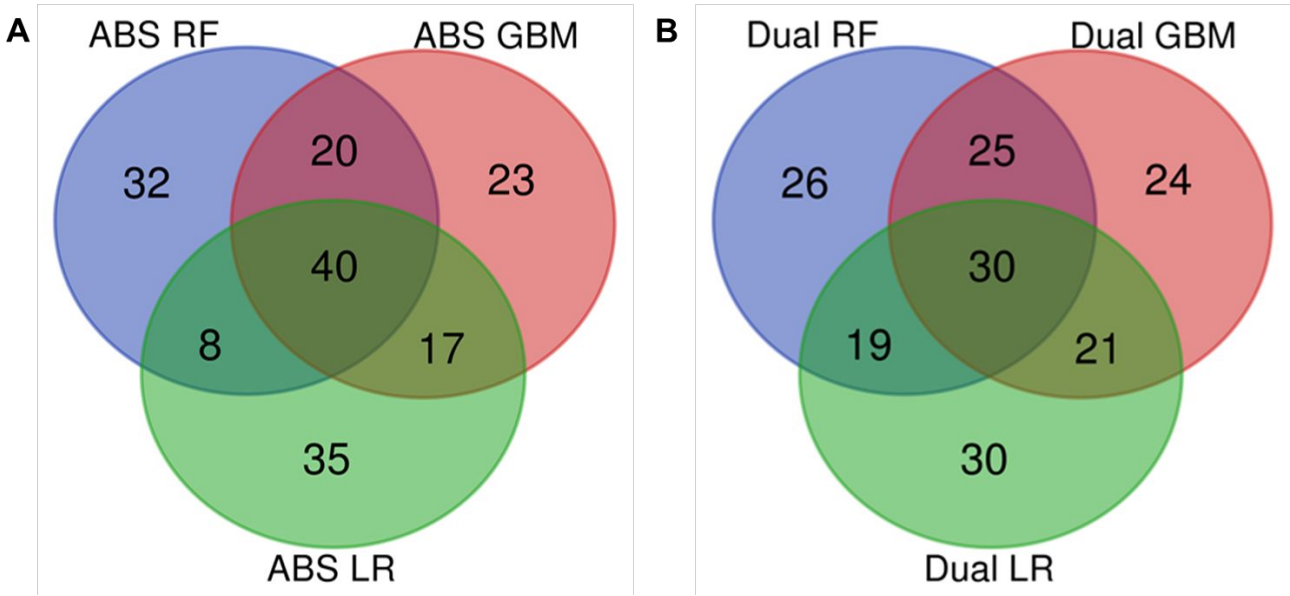

Figure S8: Shared features between models
The top 100 ECFP features were identified for (A) ABS activity prediction RF, GBM and LR models (trained on ECFP). Similarly, the top 100 ECFP features were identified for (B) dual-activity prediction RF, GBM and LR models (trained on ECFP). Intersections indicate the number of features shared between models and are detailed in Table S5.

Comparing the chemical similarity of the PRB box and the test set to that of the training set of models (Figure S8) one can clearly see that both for the ABS and dual-active compounds from the test set, though not very similar, is closer to the training data than that of the PRB box in comparison. Model precision in identifying such ABS active compounds with low similarity to the training data tended to be low (Table S5). This was also similarly seen for PRB compounds with gametocytocidal activity. Considering the low chemical similarity between such compounds and the training data such compounds may fall outside of the chemical space on which chemical models have been trained and result erroneous classification of compound activity.

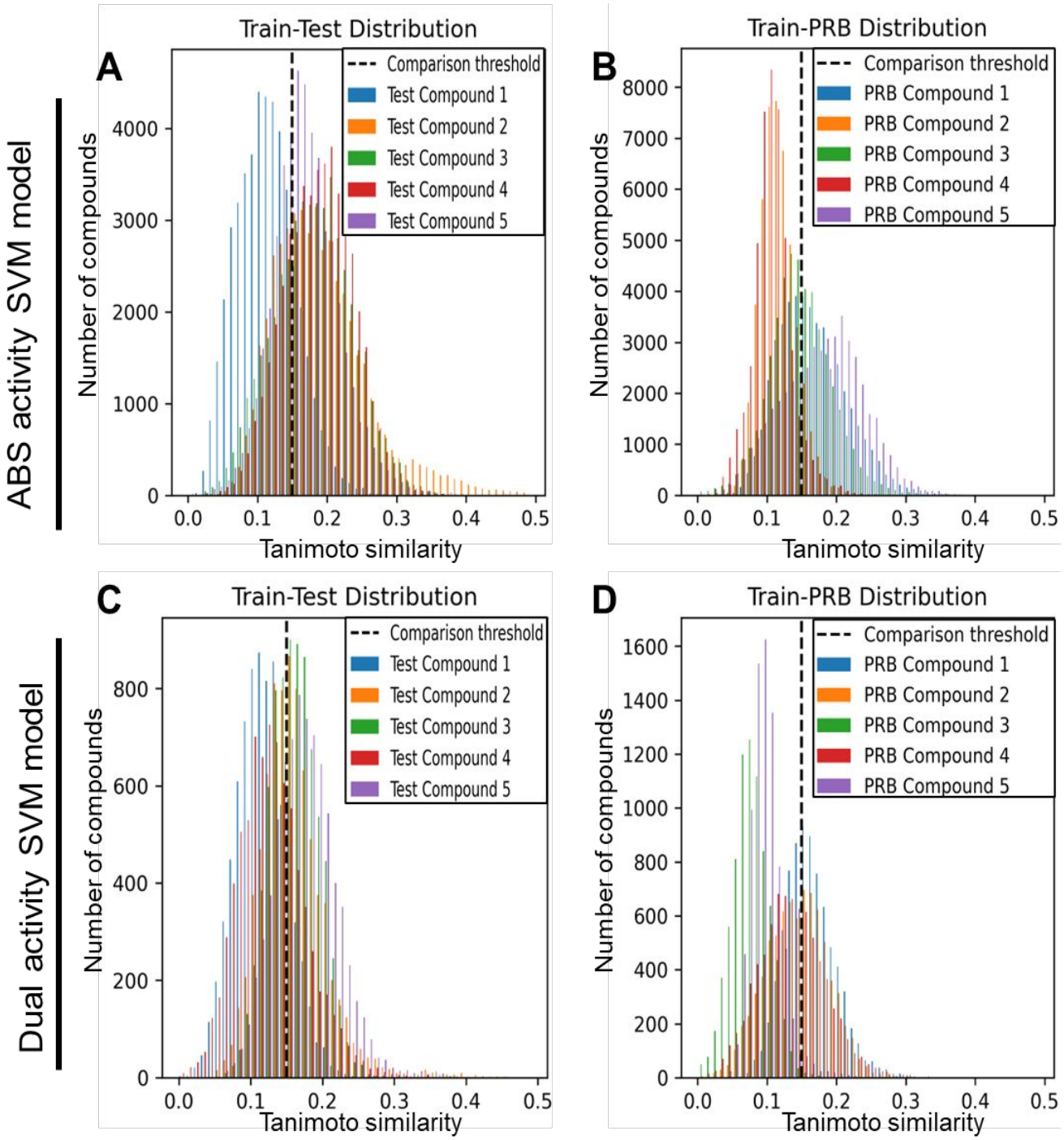

Figure S9: Tanimoto similarity distribution of PRB box or test set compounds on training set
Five compounds were randomly selected from the (A) ABS test set and PRB Box (B). Similarly, five gametocytocidal compounds were randomly selected from the (C) dual-active test set and (D) PRB box. Tanimoto similarity distribution plots were generated for each of the five compounds based on their structural similarity to the training set used for the (top) ABS activity models or (bottom) dual-activity model. The predicted activity of the randomly selected compounds is shown in Table S6.

Table S1: Hyperparameter tuning and optimal parameters identified for models

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

Table S2: Optimised probability threshold

Model	G-Mean	FPR	ROC-AUC	Recall	Precision	Probability threshold
ABS activity prediction model trained on undersampled balanced data						
SVM (ECFP)	0.875	0.149	0.875	0.899	0.224	0.5
RF (ECFP)	0.825	0.178	0.825	0.828	0.182	0.5
GBM (ECFP)	0.802	0.219	0.802	0.824	0.152	0.5
LR (ECFP)	0.708	0.483	0.743	0.969	0.088	0.5
SVM (ECFP)	0.714	0.468	0.745	0.958	0.675	0.8

RF (ECFP)	0.702	0.415	0.714	0.842	0.673	0.64
GBM (ECFP)	0.706	0.071	0.733	0.537	0.267	0.74
LR (ECFP)	0.822	0.140	0.823	0.785	0.212	0.96
Dual activity prediction model trained on undersampled balanced data						
SVM class-weighted (ECFP)	0.809	0.006	0.826	0.658	0.164	0.5
RF (ECFP)	0.562	0.000	0.658	0.316	0.578	0.5
GBM (ECFP)	0.720	0.005	0.758	0.521	0.159	0.5
LR class-weighted (ECFP)	0.796	0.061	0.807	0.675	0.020	0.5
SVM class-weighted (ECFP)	0.654	0.001	0.713	0.427	0.562	0.92
RF (ECFP)	0.562	0.000	0.658	0.316	0.617	0.52
GBM (ECFP)	0.547	0.000	0.649	0.299	0.565	0.82
LR class-weighted (ECFP)	0.768	0.041	0.787	0.615	0.027	0.96

Table S3: Complex model Autogluon comparison on test data

Model	G-Mean	FPR (FP/FP+TN)	ROC-AUC	Recall	Precision	F1-Score
ABS activity prediction model						
SVM (ECFP)	0.875	0.149	0.875	0.899	0.224	0.359
RF (ECFP)	0.825	0.178	0.825	0.828	0.182	0.298
NeuralNetFastAI	0.863	0.188	0.864	0.917	0.189	0.313
WeightedEnsemble_L2	0.877	0.163	0.877	0.918	0.213	0.345
Dual activity prediction model						
SVM (ECFP)	0.809	0.006	0.826	0.658	0.164	0.485
RF (ECFP)	0.562	0.000	0.658	0.316	0.578	0.409
NeuralNetFastAI	0.992	0.632	0.813	0.632	0.147	0.238
WeightedEnsemble_L2	0.995	0.641	0.818	0.641	0.201	0.305

Table S4: Metrics per model on PRB and Pathogen box data

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

GBM (MACCS)	0.511	0.663	0.774	0.337	0.623
LR (MACCS)	0.296	0.911	0.977	0.089	0.662
Dual activity prediction model on PRB box					
SVM class-weighted (ECFP)	0.525	0.172	0.333	0.828	0.266
RF (ECFP)	0.240	0.017	0.059	0.983	0.100
GBM (ECFP)	0.487	0.135	0.275	0.865	0.250
LR class-weighted (ECFP)	0.593	0.309	0.510	0.691	0.281
SVM class-weighted (MACCS)	0.521	0.341	0.412	0.659	0.220
RF (MACCS)	0.444	0.160	0.235	0.840	0.202
GBM (MACCS)	0.501	0.246	0.333	0.754	0.221
LR class-weighted (MACCS)	0.571	0.550	0.725	0.450	0.264
Dual activity prediction model on Pathogen box					
SVM class-weighted (ECFP)	0.449	0.185	0.247	0.815	0.281
RF (ECFP)	0.226	0.011	0.052	0.989	0.095
GBM (ECFP)	0.444	0.170	0.237	0.830	0.277
LR class-weighted (ECFP)	0.467	0.359	0.340	0.641	0.291
SVM class-weighted (MACCS)	0.492	0.289	0.340	0.711	0.317
RF (MACCS)	0.408	0.104	0.186	0.896	0.252
GBM (MACCS)	0.408	0.152	0.196	0.848	0.242
LR class-weighted (MACCS)	0.567	0.463	0.598	0.537	0.414

Table S5: Top 100 shared and unique ECFP features among models

Models sharing ECFP features	**Number of ECFP features shared**	**ECFP features shared**
ABS activity prediction models		
GBM, LR, RF	40	200; 161; 298; 194; 78; 349; 291; 114; 153; 81; 311; 346; 495; 323; 345; 191; 178; 234; 314; 388; 326; 209; 256; 195; 21; 80; 99; 457; 277; 377; 232; 295; 50; 375; 303; 496; 239; 491; 332; 213
GBM, RF	20	90; 463; 434; 325; 226; 321; 76; 67; 236; 112; 487; 367; 216; 324; 173; 70; 484; 387; 328; 265
LR, RF	8	424; 93; 262; 176; 74; 447; 221; 310
GBM, LR	17	102; 27; 458; 431; 471; 172; 193; 72; 351; 201; 149; 302; 499; 250; 305; 59; 42
RF	32	206; 71; 125; 190; 437; 313; 231; 89; 343; 11; 275; 197; 337; 101; 459; 62; 241; 69; 119; 264; 92; 414; 167; 39; 260; 380; 155; 53; 158; 354; 399; 242
GBM	23	127; 259; 243; 35; 65; 203; 371; 145; 154; 96; 126; 37; 63; 455; 162; 410; 334; 68; 25; 46; 257; 3; 408
LR	35	18; 366; 95; 220; 20; 347; 29; 374; 58; 129; 2; 14; 131; 282; 23; 364; 159; 170; 428; 180; 341; 438; 87; 56; 470; 144; 283; 254; 105; 248; 111; 146; 456; 4; 317
GBM, LR, RF	40	200; 161; 298; 194; 78; 349; 291; 114; 153; 81; 311; 346; 495; 323; 345; 191; 178; 234; 314; 388; 326; 209; 256; 195; 21; 80; 99; 457; 277; 377; 232; 295; 50; 375; 303; 496; 239; 491; 332; 213

GBM, RF	20	90; 463; 434; 325; 226; 321; 76; 67; 236; 112; 487; 367; 216; 324; 173; 70; 484; 387; 328; 265
Dual-activity prediction models		
GBM, LR, RF	30	200; 194; 78; 434; 349; 291; 58; 81; 311; 495; 345; 191; 282; 326; 195; 21; 455; 99; 324; 377; 149; 310; 303; 496; 328; 491; 318; 257; 332; 242
GBM, RF	25	118; 55; 84; 95; 325; 203; 261; 153; 82; 178; 79; 364; 176; 256; 216; 61; 232; 167; 70; 100; 484; 392; 473; 213; 408
LR, RF	19	424; 20; 397; 65; 459; 241; 471; 218; 404; 314; 486; 227; 488; 447; 477; 342; 360; 421; 13
GBM, LR	21	127; 445; 376; 220; 337; 289; 2; 323; 236; 234; 209; 80; 379; 263; 474; 370; 116; 380; 46; 436; 456
RF	26	298; 31; 340; 76; 339; 361; 112; 187; 121; 344; 355; 479; 367; 460; 297; 334; 312; 56; 260; 1; 375; 416; 254; 155; 354; 3
GBM	24	90; 366; 233; 190; 259; 275; 321; 346; 69; 487; 43; 457; 255; 277; 91; 438; 87; 214; 293; 365; 141; 222; 25; 250
LR	30	276; 27; 161; 458; 89; 157; 29; 431; 432; 67; 327; 212; 126; 483; 37; 428; 383; 410; 201; 50; 221; 173; 499; 68; 283; 450; 85; 185; 390; 439

Table S6: Activity predictions of compounds with low chemical similarity to training set

Compound	Activity	Predicted activity	Predicted probability
ABS activity SVM model			
Test set compound 1	ABS activity/Hit	Inactive	0.003860
Test set compound 2	ABS activity/Hit	Active	1.000000
Test set compound 3	ABS activity/Hit	Active	0.984875
Test set compound 4	ABS activity/Hit	Active	0.990663
Test set compound 5	ABS activity/Hit	Active	0.990675
PRB compound 1	ABS activity/Hit	Active	0.999999
PRB compound 2	ABS activity/Hit	Inactive	0.639942
PRB compound 3	ABS activity/Hit	Active	0.920255
PRB compound 4	ABS activity/Hit	Active	0.987174
PRB compound 5	ABS activity/Hit	Active	0.995573
Dual-activity SVM model			
Test set compound 1	Dual-activity/Gametocytocidal	Inactive	0.020803
Test set compound 2	Dual-activity/Gametocytocidal	Active	0.958620
Test set compound 3	Dual-activity/Gametocytocidal	Active	0.537336
Test set compound 4	Dual-activity/Gametocytocidal	Active	0.982600
Test set compound 5	Dual-activity/Gametocytocidal	Active	0.962076
PRB compound 1	Dual-activity/Gametocytocidal	Active	0.950257
PRB compound 2	Dual-activity/Gametocytocidal	Active	0.973742
PRB compound 3	Dual-activity/Gametocytocidal	Inactive	0.302202
PRB compound 4	Dual-activity/Gametocytocidal	Inactive	0.011141
PRB compound 5	Dual-activity/Gametocytocidal	Inactive	0.233355

Table S7: Current ML models for infectious parasitic diseases

Model	Model name	Disease agent	Target on which training data is based on	Constraints	Advantages	Ref
Naïve Bayes	MAIP	*P. falciparum* (ABS)	Phenotypic screening data/Whole-cell inhibition activity	• Compound activity restrictive to the biology of the parasite stage on which phenotypic screening was conducted. • Chemical space of specific targets may not be well defined within whole-cell models. • Active compound MoA unknown. • Larger training dataset required.	• Chemical space of multiple targets can be captured. • Due to whole-cell chemical space captured, predicted actives may still be active towards resistant strains. • Compound activity relating to transport into cell captured. • Multiple compounds with different and novel MoA can be identified.	1
Random Forest	NA	*P. falciparum* (ABS and Liver stages)				2
Random forest and DNN	DeepMalaria	*P. falciparum* (ABS)				3
ANN & KPLS	NA	*Trypanosoma cruzi*				4
Bayesian	NA	*Schistosoma mansoni*				5
DNN	NA	*P. falciparum*	*P. falciparum* Ion pump (*Pf*ATP4)	• Compound activity restricted to one/few targets. • Compound activity may fail/decrease due to failed transport/permeability of compound into cell or active site. • Chemical space of target may change due to gene mutations to target protein in response to drugs and hence resistant strains may be outside the scope of the model. • Compound MoA is restricted	• More defined and finite chemical space for activity. • MoA of active compounds is known • Smaller set of training data required	6
Naïve Bayes and Ensemble models	NA	*P. falciparum*	*Plasmodium falciparum* enoyl acyl carrier protein reductase (*Pf*ENR)			7
Random Forest	NA	*Leishmania*	*L.mexicana* pyruvate kinase enzyme (*Lm*PK)			8

References

(1) 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 Cheminform* **2021**, *13* (1), 13. DOI: 10.1186/s13321-021-00487-2.

(2) 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. DOI: 10.1021/acsinfecdis.2c00189.

(3) Keshavarzi Arshadi, A.; Salem, M.; Collins, J.; Yuan, J. S.; Chakrabarti, D. DeepMalaria: Artificial Intelligence Driven Discovery of Potent Antiplasmodials. *Front Pharmacol* **2019**, *10*, 1526, Original Research. DOI: 10.3389/fphar.2019.01526.

(4) de Souza, A. S.; Ferreira, L. L. G.; de Oliveira, A. S.; Andricopulo, A. D. Quantitative Structure–Activity Relationships for Structurally Diverse Chemotypes Having Anti-Trypanosoma cruzi Activity. *International Journal of Molecular Sciences* **2019**, *20* (11), 2801.

(5) Zorn, K. M.; Sun, S.; McConnon, C. L.; Ma, K.; Chen, E. K.; Foil, D. H.; Lane, T. R.; Liu, L. J.; El-Sakkary, N.; Skinner, D. E.; et al. A Machine Learning Strategy for Drug Discovery Identifies Anti-Schistosomal Small Molecules. *ACS Infectious Diseases* **2021**, *7* (2), 406-420. DOI: 10.1021/acsinfecdis.0c00754.

(6) Tse, E. G.; Aithani, L.; Anderson, M.; Cardoso-Silva, J.; Cincilla, G.; Conduit, G. J.; Galushka, M.; Guan, D.; Hallyburton, I.; Irwin, B. W. J.; et al. An Open Drug Discovery Competition: Experimental Validation of Predictive Models in a Series of Novel Antimalarials. *J Med Chem* **2021**, *64* (22), 16450-16463. DOI: 10.1021/acs.jmedchem.1c00313 From NLM.

(7) Shah, P.; Tiwari, S.; Siddiqi, M. I. Integrating molecular docking, CoMFA analysis, and machine-learning classification with virtual screening toward identification of novel scaffolds as Plasmodium falciparum enoyl acyl carrier protein reductase inhibitor. *Medicinal Chemistry Research* **2014**, *23* (7), 3308-3326. DOI: 10.1007/s00044-014-0910-7.

(8) Jamal, S.; Scaria, V. Cheminformatic models based on machine learning for pyruvate kinase inhibitors of Leishmania mexicana. *BMC Bioinformatics* **2013**, *14* (1), 329. DOI: 10.1186/1471-2105-14-329.