EMCIP: An Ensemble Model for Cdr1 Inhibitor Prediction Leveraging Traditional Machine Learning and Deep Learning Methodologies

The-Chuong Trinh[1], Pierre Falson[2], Viet-Khoa Tran-Nguyen[3]* and Ahcène Boumendjel[1]*

[1]Univ. Grenoble Alpes, INSERM, LRB, Grenoble, 38000, France

[2]Drug Resistance & Membrane Proteins Group, CNRS-Lyon 1 University Laboratory UMR 5086, IBCP, 69367, CEDEX Lyon 07, France

[3]Université Paris Cité, CNRS UMR8251, INSERM U1133, F-75013 Paris, France

*Corresponding authors: viet-khoa.tran-nguyen@u-paris.fr, Ahcene.Boumendjel@univ-grenoble-alpes.fr

ABSTRACT

The emergence of antifungal-resistant *Candida* strains necessitates novel therapeutic strategies, usually tackling the overexpression of Cdr1, an ATP-binding cassette efflux pump. This study describes the development of EMCIP, a new ensemble model for Cdr1 inhibitor prediction leveraging multiple traditional machine learning (ML) algorithms and a multi-instance 3D graph neural network. It utilized various molecular feature types and learned from ligand conformations represented as 3D molecular graphs. On a test set structurally dissimilar to the training data, its average precision was 0.755, its F1-score was 0.714, the area under the receiver operating characteristic curve was 0.884, and the balanced accuracy was 0.799. It gave a low false positive rate of 0.1236 on another test set outside the training chemical space, indicating its ability to avoid false positives. This work highlights the potential of stacking ensemble ML and offers a rigorous general workflow to build ligand-based predictive ML models for other targets.

Keywords: Cdr1 inhibitors, antifungal resistance, drug discovery, machine learning, deep learning, graph neural network, stacking ensemble.

Table 1. Evaluation results on the ET and HT sets of our traditional ML models and MIL-3D-GNN.

Feature type	Algorithm	ET-AP	ET-F1-score	ET-ROC-AUC	ET-BaAcc	HT-FPR
RDK5	CatBoost	0.630	0.63	0.841	0.748	0.0948
RDK6	XGB	0.505	0.5	0.787	0.690	0.0966
RDK7	LR	0.487	0.507	0.816	0.709	0.1010
Mordred	CatBoost	0.696	0.704	0.847	0.787	0.0821
Avalon	CatBoost	0.619	0.679	0.852	0.780	0.0749
Ph4_gobbi	MLP	0.561	0.593	0.765	0.729	0.1742
Graph	MIL-3D-GNN	0.698	0.588	0.868	0.721	0.1381

ET: External test set, Ph4_gobbi: Gobbi pharmacophore, BaAcc: balanced accuracy, MIL-3D-GNN: multi-instance 3D graph neural network, HT-FPR: false positive rate on the hard test (HT) set.

Table 2. The performance of 20 stacking models employing 10 ML algorithms as meta-learners on the validation, ET, and HT sets.

Algorithm used as meta-learner of each stacking model	Number of base models*	Ligand set	AP	F1-score	ROC-AUC	BaAcc	HT-FPR
LR	7	Validation	0.879	0.889	0.915	0.900	
	7	ET	0.695	0.588	0.883	0.721	0.0650
	6	Validation	0.867	0.889	0.907	0.900	
	6	ET	0.608	0.588	0.859	0.721	0.0686
KNN	7	Validation	0.838	0.889	0.895	0.900	
	7	ET	0.568	0.642	0.795	0.752	0.0839
	6	Validation	0.838	0.865	0.895	0.894	
	6	ET	0.566	0.604	0.795	0.733	0.0948
SVM	7	Validation	0.876	0.842	0.904	0.888	
	7	ET	0.742	0.679	0.867	0.780	0.1173
	6	Validation	0.860	0.842	0.898	0.888	
	6	ET	0.663	0.679	0.781	0.780	0.0993
RF	7	Validation	0.852	0.889	0.898	0.900	
	7	ET	0.699	0.655	0.884	0.764	0.0957
	6	Validation	0.838	0.865	0.880	0.894	
	6	ET	0.629	0.679	0.824	0.780	0.0993
ExT	7	Validation	0.861	0.842	0.908	0.888	
	7	ET	0.670	0.655	0.858	0.764	0.1038
	6	Validation	0.838	0.865	0.874	0.894	
	6	ET	0.642	0.679	0.781	0.780	0.1011
Ada	7	Validation	0.594	0.744	0.859	0.859	
	7	ET	0.531	0.678	0.788	0.788	0.1146
	6	Validation	0.791	0.865	0.894	0.894	
	6	ET	0.611	0.727	0.802	0.802	0.0930

Grad	7	Validation	0.838	0.865	0.885	0.894	
	7	ET	0.630	0.679	0.824	0.780	0.0839
	6	Validation	0.838	0.865	0.885	0.894	
	6	ET	0.629	0.679	0.824	0.780	0.1002
XGB	7	Validation	0.655	0.751	0.846	0.833	
	7	ET	0.536	0.593	0.820	0.729	0.1760
	6	Validation	0.655	0.757	0.846	0.833	
	6	ET	0.536	0.593	0.820	0.720	0.1760
CatBoost	7	Validation	0.866	0.800	0.901	0.877	
	7	ET	0.723	0.667	0.883	0.776	0.1471
	6	Validation	0.863	0.821	0.902	0.883	
	6	ET	0.599	0.679	0.870	0.780	0.1300
MLP	7	Validation	0.881	0.865	0.916	0.894	
	7	ET	0.701	0.642	0.880	0.752	0.1146
	6	Validation	0.867	0.842	0.905	0.888	
	6	ET	0.599	0.630	0.854	0.748	0.1119
P-values**			0.04	0.41	0.03	0.34	0.95

*: In case the number of base models was 6, the DL model MIL-3D-GNN was excluded.

**: P-values were calculated using the Wilcoxon signed-rank test. The sample included evaluation values of 20 meta-models (10 employing all seven base models and 10 employing all but MIL-3D-GNN) on both the ET set and the validation set. A p-value below 0.05 indicated statistically significant differences between the two compared scenarios: one using MIL-3D-GNN and the other excluding it.

ET: External test set, HT: Hard test set, FPR: false positive rate.

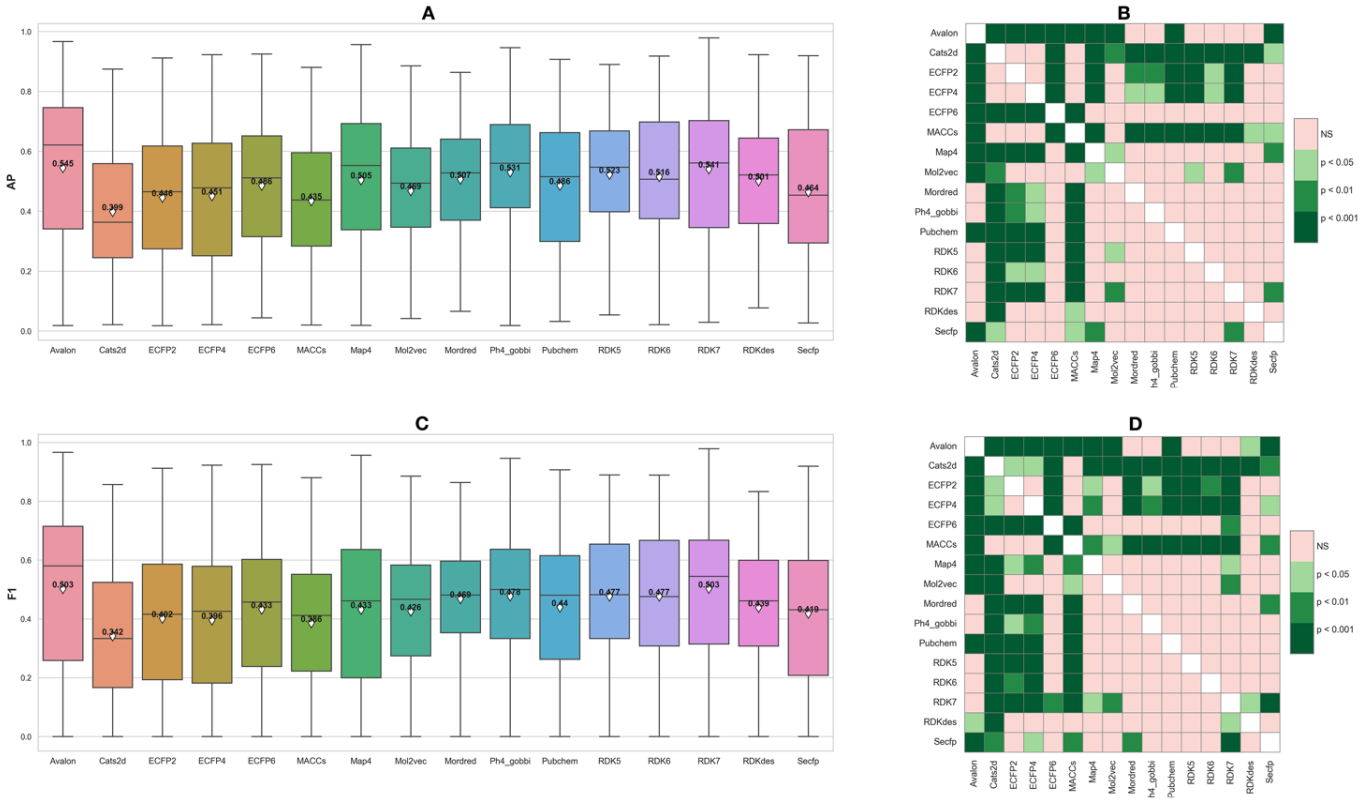


Fig. 1. Boxplots and Wilcoxon heatmaps visualizing the results of our molecular representation meta-analysis. (A) Boxplots comparing the BM 10-fold CV results across 16 types of LB structural representation based on AP. (B) Heat map illustrating the results of Wilcoxon signed-rank tests based on AP. (C) Boxplots comparing the BM 10-fold CV results across 16 types of molecular representation based on F1-scores. (D) Heat map illustrating the results of Wilcoxon signed-rank tests based on F1-scores. In (B) and (D), the pink cells represent statistically insignificant differences between two molecular feature types (p-values≥0.05). Conversely, the green cells indicate statistically significant differences between the two compared feature types (p-values<0.05). P-values are provided in Tables S1, S2.

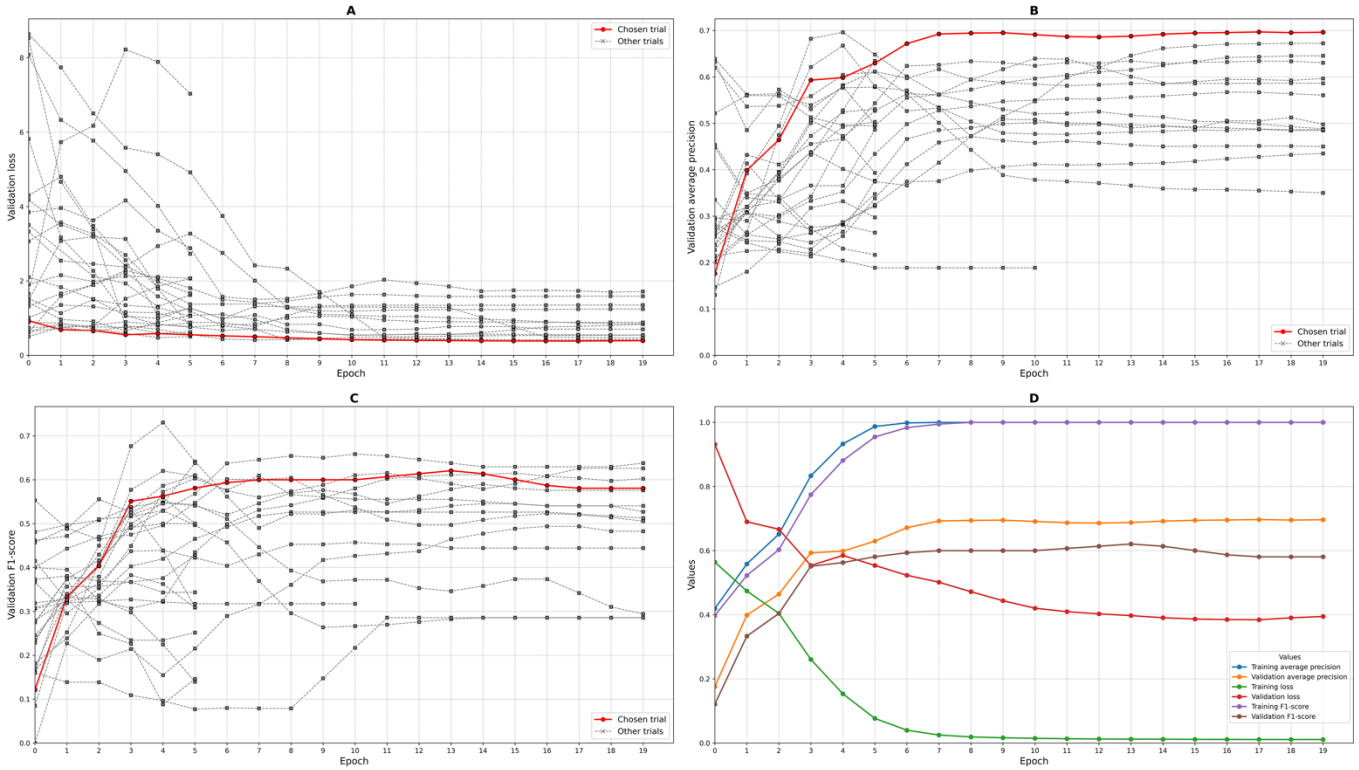
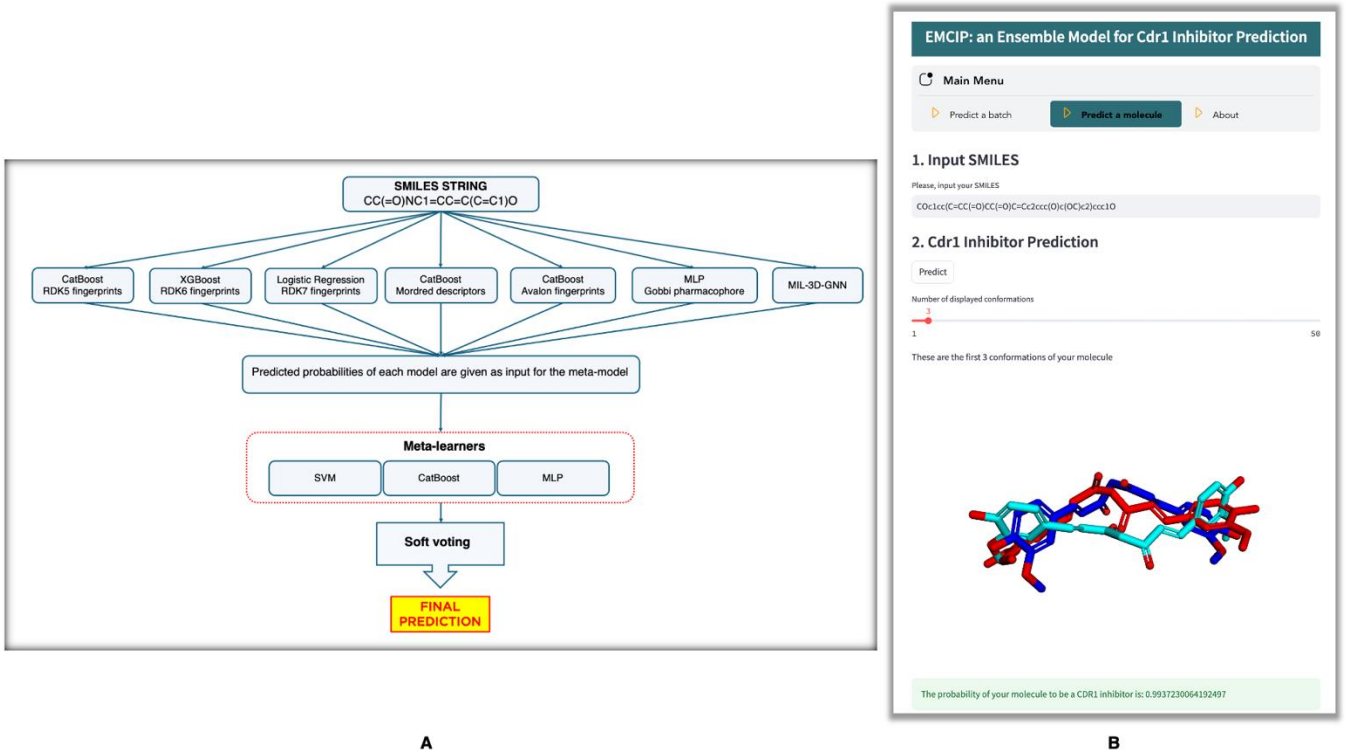


Fig. 2. The results of hyperparameter tuning across 30 trials in Phase 2. (A) The records of validation BCE losses across 20 epochs of 30 trials. (B) The records of validation AP values across 20 epochs of 30 trials. (C) The records of validation F1-scores across 20 epochs of 30 trials. (D) The learning curve when the configuration obtained from the 13th trial was used. Some trials were pruned, using the median stopping rule implemented in the 'MedianPruner' of Optuna, when their results were not promising, as illustrated by shorter recorded lines in panels (A), (B), and (C).



A B

Fig. 3. Our final Ensemble Model for Cdr1 Inhibitor Prediction (EMCIP). (A) The architecture of EMCIP. Screened molecules, represented by SMILES strings, were preprocessed for different models (fingerprints, descriptors, 3D molecular graphs). Seven base models then predicted each molecule's probability of being a Cdr1 inhibitor. These predictions were fed into three stacking models (SVM, CatBoost, MLP) for further refinement. Finally, soft voting combined the probabilities from these models to deliver the final classification. (B) The graphical user interface (GUI) for EMCIP. This GUI is accessible and user-friendly to those unfamiliar with programming.

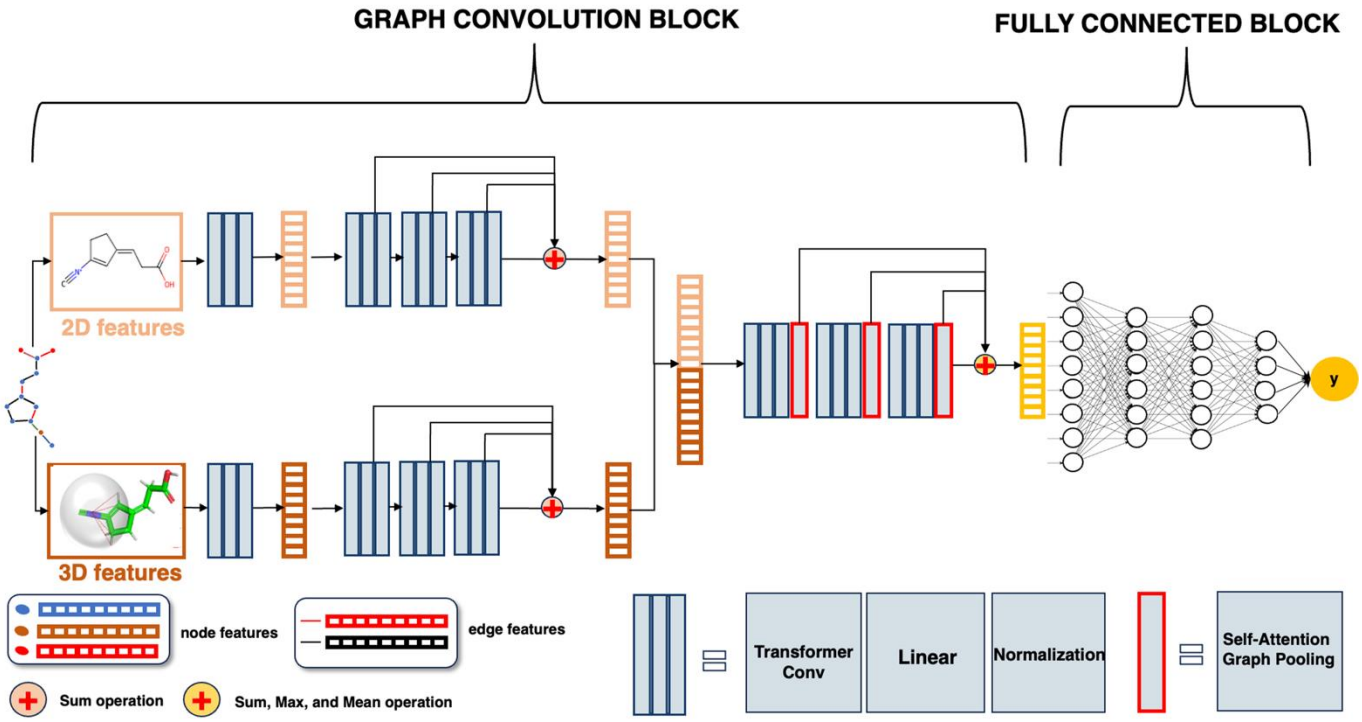


Fig. 4. The architecture of our MIL-3D-GNN model.