Supporting Information

HERGAI: An Artificial Intelligence Tool for Structure-Based Prediction of hERG Inhibitors

Viet-Khoa Tran-Nguyen,[#,*] Ulrick Fineddie Randriharimanamizara,[#] Olivier Taboureau*

Université Paris Cité, CNRS UMR 8251, INSERM ERL 1133, F-75013 Paris, France

*Corresponding authors: Viet-Khoa Tran-Nguyen (viet-khoa.tran-nguyen@u-paris.fr) and Olivier Taboureau (olivier.taboureau@u-paris.fr).

[#]These authors contributed equally.

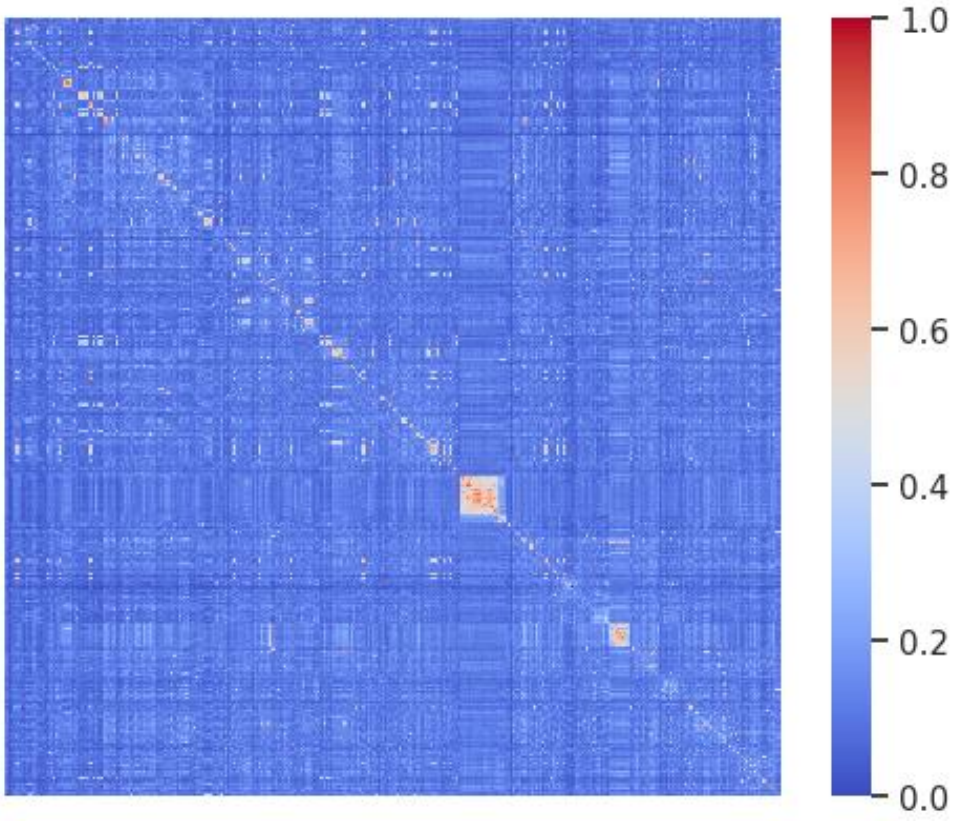

Figure S1. Heatmap illustrating the pairwise Tanimoto similarity values of 1,937 true active molecules included in our final data set, in terms of Morgan fingerprints (2,048 bits, radius 2). Only 1.50% of all active ligand pairs (n=1,875,016) had Tanimoto similarity above 0.30, highlighting the structural diversity of our active ligands.

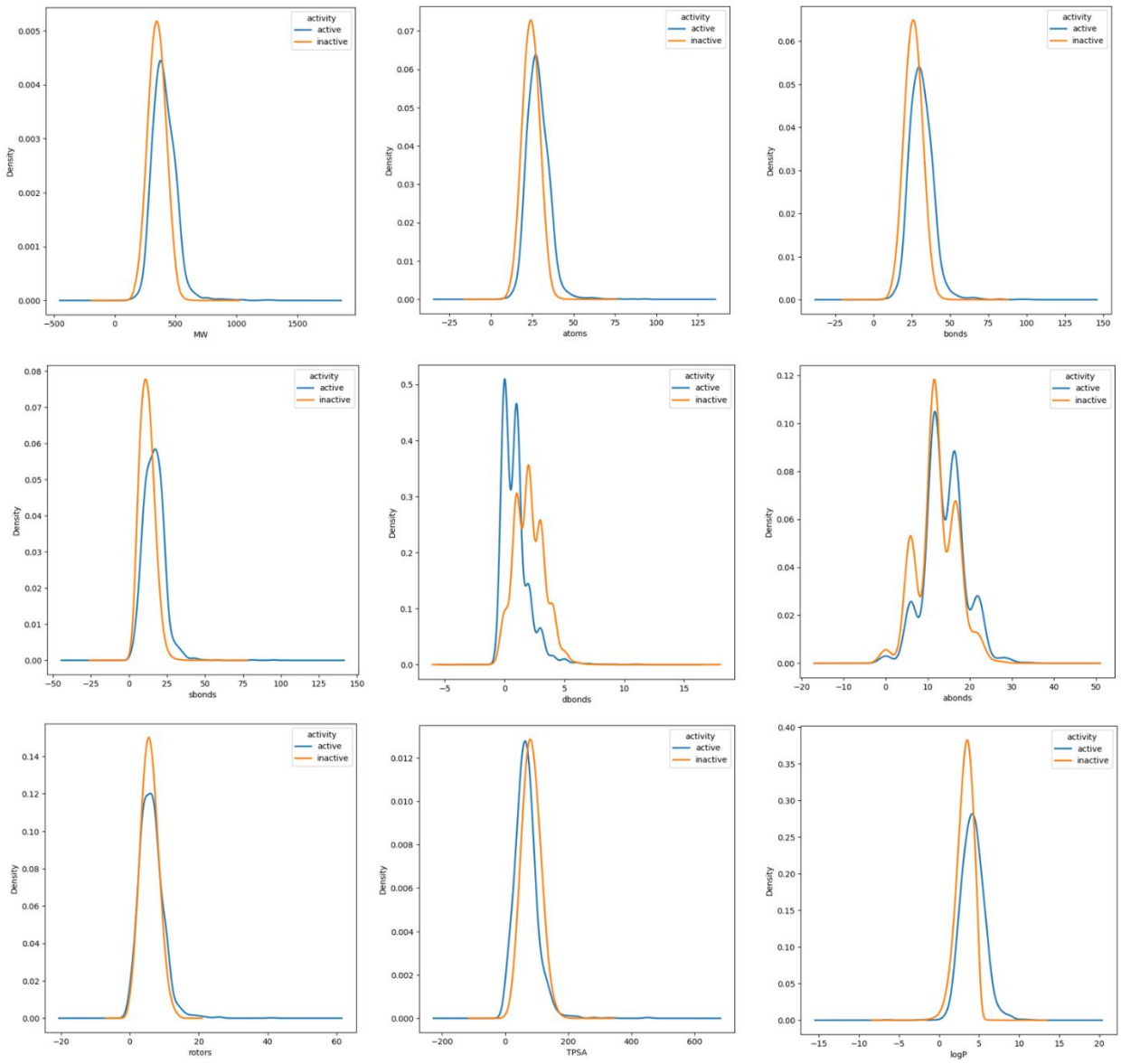

Figure S2. Distribution plots of nine physicochemical properties calculated from all molecules of our final data set, using Open Babel (v.3.1.1). These properties include: molecular weight (MW), number of atoms (atoms), number of bonds (bonds), number of single bonds (sbonds), number of double bonds (dbonds), number of aromatic bonds (abonds), number of rotatable bonds (rotors), topological polar surface area (TPSA), and lipophilicity (logP). The y axis of each plot represents the density of probability. A substantial overlap of these properties can be observed, proving that the actives and the inactives shared wide physicochemical property ranges.

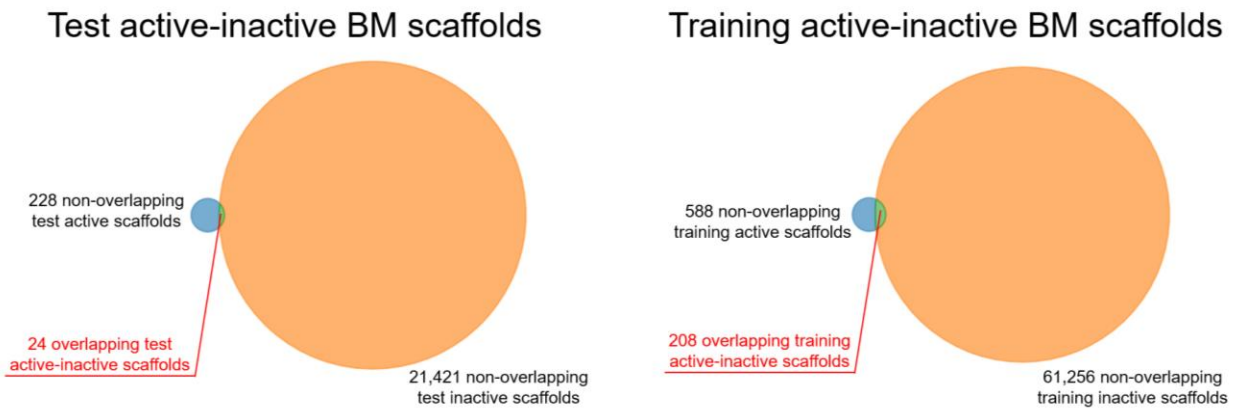

Figure S3. Venn diagrams illustrating the BM scaffolds shared between the active and inactive sets. The presence of overlapping active-inactive scaffolds, particularly in the test set, increases the difficulty of our benchmark and reflects real-world screening libraries, where active compounds often differ from inactives by only subtle structural variations that may not be captured by BM scaffolds.

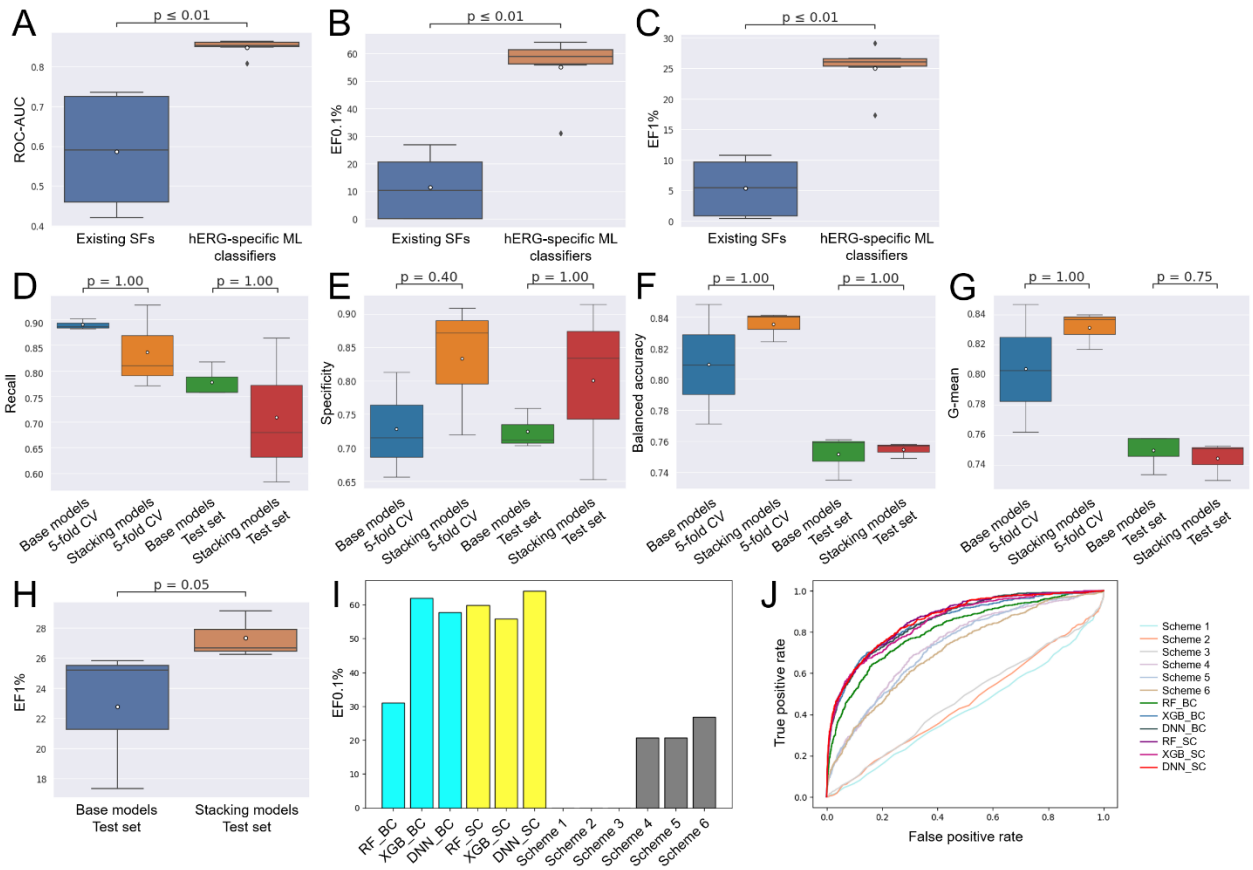

Figure S4. Performance of our six AI-based binary classifiers for hERG inhibitor prediction. Their screening power was evaluated on the external test set and compared to that of six VS schemes involving existing SFs, in terms of ROC-AUC **(A)**, EF0.1% **(B)**, and EF1% **(C)**: all p-values being lower than 0.05 (Mann-Whitney U test with Bonferroni correction), the difference between their screening power was statistically significant. In terms of classification power, the stacking models gave comparable performance to their base components, according to the recall **(D)**, the specificity **(E)**, the balanced accuracy **(F)**, and the G-mean **(G)**, both during 5-fold CV and on the external test set: all p-values being higher than 0.05 (Mann-Whitney U test with Bonferroni correction). However, the stacking models gave significantly better EF1% than the base ones **(H)**: p-value being equal to 0.05 (Mann-Whitney U test with Bonferroni correction). The EF0.1% given by all hERG-specific ML models and the six VS schemes involving existing generic SFs are portrayed in **(I)**. For the EF1%, see Figure 4C. The ROC curves illustrating the screening performance of our six AI-based classifiers, alongside those of the six VS schemes employing existing SFs are provided in **(J)**.

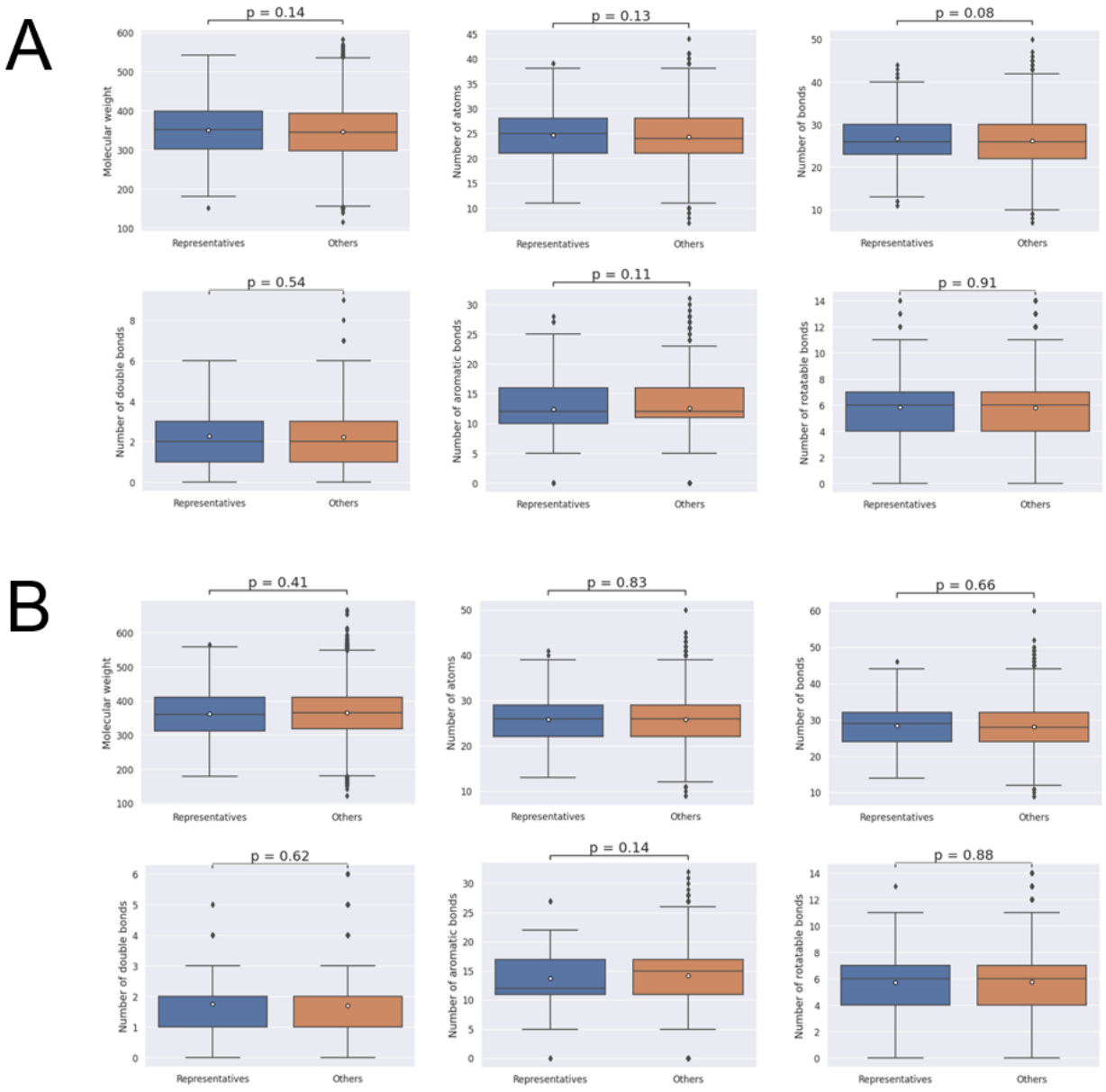

Figure S5. Comparisons of key structural features between the chosen 1% subset and the remaining 99% of true negative (TN) **(A)** and false positive (FP) **(B)** instances. Statistical analyses were performed using the Mann-Whitney U test with Bonferroni correction. All p-values exceeded 0.05, implying that the chosen 1% did not differ significantly from the other 99% in terms of these features. Along with Figure 5, the observations herein suggest that the selected TN and FP instances were representative of the entire inactive test population.

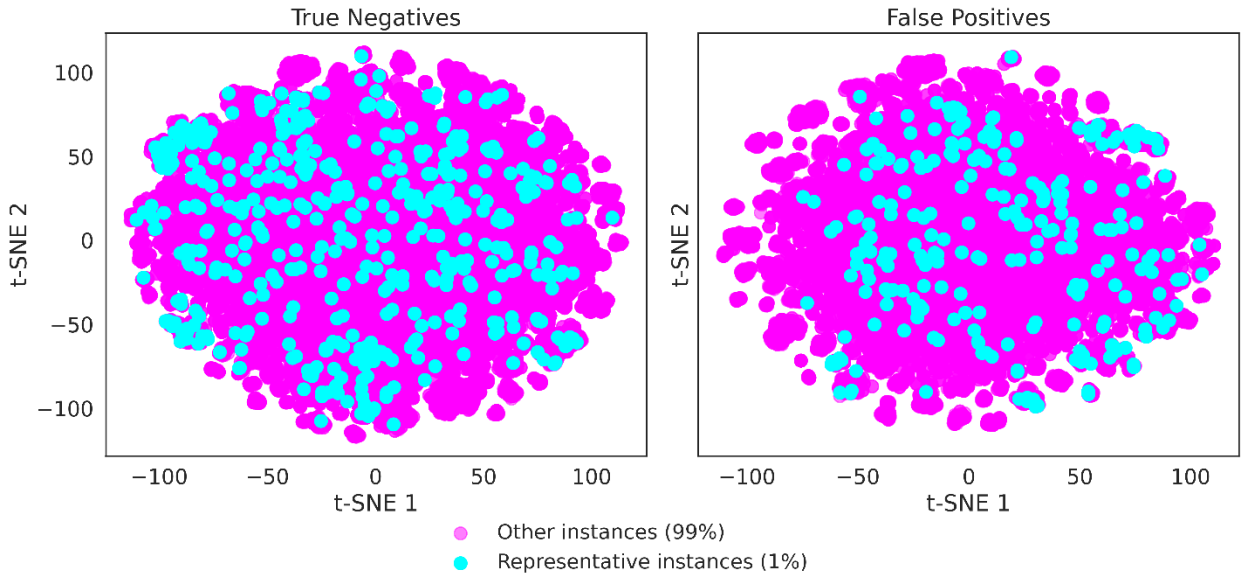

Figure S6. 2D representation of the chemical space defined by the first two embedded dimensions obtained from t-SNE dimensionality reduction. This analysis was applied to our representative instances (1%) and the remaining 99% of the inactive test data using Morgan fingerprints (2,048 bits, radius 2) as descriptors, and the 'TSNE' function from 'sklearn' (v.1.7.0) with 'n_components' set to 2 and 'random_state' fixed at 42 (all other parameters were left at default values). Similar to our PCA analyses (Figure 5), the representatives are observed to be well distributed across the space occupied by the rest of the inactive test data, including the peripheral regions, thereby confirming the representativeness of the selected 1% inactive test subset in chemical space.

Table S1. Overview of hERG inhibitor prediction tools published since 2021. For a comprehensive overview of hERG blocker predictors published in 2020 or earlier, please refer to reference [16] cited in our article.

Tool	Availability	Data sets	Methodologies	Advantages	Disadvantages/issues
CardioTox net Karim et al (2021) [17]	GitHub repository: CardioTox	Training set: 12,620 molecules. Three external test sets collected from published articles.	Convolution 1D neural network, fully connected neural network, graph convolutional neural network, meta ensemble	Trained on multiple molecular feature types, the model effectively combines insights from different chemical representations.	Two of the three test sets are small (<50 molecules) and have unrealistically high active-to-inactive ratios, likely leading to overestimated screening performance. The training data used are limited in size, despite the availability of much larger public data sets.
DMPNN-hERG Shan et al (2022) [18]	GitHub repository: DMPNN-hERG	Training set: 7,889 molecules from Cai et al (2019) [15]. Six external test sets collected from published articles.	A directed message passing neural network (D-MPNN) that builds molecular representations by operating on the molecular graph, using an edge-dependent (rather than atom-dependent) neural network for message propagation	Combining D-MPNN with moe206 2D descriptors leverages the strengths of both learned graph-based representations and traditional 2D features. This hybrid approach enhances the model's capacity to capture both structural connectivity and molecular properties, leading to improved screening performance.	Some test sets have overly high active-to-inactive ratios, which can inflate screening performance estimates. The training data used are relatively small, even though much larger public data sets are available.

HergSPred Zhang et al (2022) [11]	Web service (icdrug.com)	Internal data set: 12,850 compounds (6,907 actives and 5,943 inactives). Five external test sets collected from published articles.	Random forest (using Scikit-learn), extreme gradient boosting (using xgboost), deep neural network (using Keras), ensemble averaging	The study offers a structural feature analysis that serves as a valuable guide for molecular design.	No source code is publicly available; only a web service is provided, which supports only small-scale input, limiting its utility on large data sets. Some test sets have unrealistically high active-to-inactive ratios, likely leading to inflated screening performance. The training data are limited in size.
Ylipää et al (2023) [20]	Not available	Training set: 203,853 compounds. Test set: 87,366 compounds.	Random forest, support vector machine, extreme gradient boosting, gated recurrent unit-based deep neural network, graph neural network (GNN)	The models were trained and evaluated on a large data set comparable in size to those found in public repositories. The GNN method eliminates the need for manual feature engineering, requiring minimal human intervention, offering a promising foundation for fully automated workflows.	No source code or web service is available, limiting the usability of the best model (GNN) for both retrospective analyses and prospective applications.
AttenhERG Yang et al (2024) [21]	GitHub repository: AttenhERG	Internal data set: 14,322 molecules (8,488 actives and 5,834 inactives).	Graph-based molecular representations, attentive encoding mechanisms, and uncertainty evaluation analysis	AttenhERG is an advanced framework that enhances both the accuracy and interpretability of hERG channel blocker prediction. Through the	Some test sets exhibit unrealistically high active-to-inactive ratios, potentially leading to inflated screening performance estimates. The

		Five external test sets: four collected from published articles and one in-house data set.		incorporation of uncertainty estimation, it offers greater reliability than existing benchmark models.	training data are limited in size despite the availability of substantially larger public data sets.
hERGAT Lee and Yoo (2025) [22]	GitHub repository: hERGAT	Internal data set: 23,381 compounds (14,183 actives and 9,198 inactives). Three external test sets collected from published articles.	A gated recurrent unit combined with a graph attention mechanism	The model captures intricate atomic and molecular interactions, and its interpretability is enhanced through analysis of highlighted substructures, offering meaningful insights into their contributions to hERG activity.	All external test sets present unrealistically high active-to-inactive ratios, which may result in overestimated screening performance. The training data are relatively small, even though significantly larger public data sets are available.
Falcón-Cano et al (2025) [23]	Web service (hypercubane.re)	Internal data set: 291,219 molecules (9,890 actives and 281,329 inactives). Two external validation sets: one from PubChem, the other collected from the US federal Tox21 program.	A refined machine learning algorithm integrating recursive methods for variable selection, extreme gradient boosting and isometric stratified ensemble (ISE) mapping	Model interpretability is improved by employing recursive methods for variable selection. Prospective validation on external data sets confirms strong generalization even with fewer descriptors. The ISE map offers a model-agnostic approach to assess prediction confidence and applicability domain, further increasing model interpretability.	The web service restricts input to a maximum of 100 molecules per session, limiting its scalability. Additionally, the source code is not openly accessible (e.g., via GitHub) and is only available under a negotiated license.

Table S2. The optimal hyperparameters used to train our six binary classification models. The decision threshold for each classifier was also tuned along with the hyperparameters.

Binary classifier	Optimal hyperparameters and decision threshold
RF_BC	'criterion': 'gini', 'max_depth': 6, 'n_estimators': 2600, 'threshold': 0.32799228309084005
XGB_BC	'max_depth': 4, 'n_estimators': 86.0, 'reg_alpha': 0.5, 'reg_lambda': 1, 'threshold': 0.06986083857271219
DNN_BC	'batch_size': 68.0, 'dropout1': 0.3912375678578729, 'dropout2': 0.4672821567914654, 'dropout3': 0.382443636406835, 'optimizer': 'Adadelta', 'units1': 352.8790957431371, 'units2': 507.20805366280905, 'units3': 242.0267251614644, 'threshold': 0.14415252008378787
RF_SC	'criterion': 'gini', 'max_depth': 7, 'n_estimators': 3600, 'threshold': 0.4747112704737022
XGB_SC	'max_depth': 4, 'n_estimators': 100.0, 'reg_alpha': 5,

	'reg_lambda': 0.5, 'threshold': 0.3341022941083429
DNN_SC	'batch_size': 46.0, 'dropout1': 0.3294908082433152, 'dropout2': 0.4453136608614301, 'dropout3': 0.5668191107970695, 'optimizer': 'rmsprop', 'units1': 131.561382068786, 'units2': 227.11758746306032, 'units3': 81.34015005515602, 'threshold': 0.20844117646625016

Table S3. Standard deviations of different evaluation metrics across validation folds during 5-fold CV, and MCC performance of our six AI/ML binary classifiers. As explained in the manuscript, these near-zero MCC values do not fairly reflect the classification power of our models on a severely class-imbalanced test set (active-to-inactive ratio of 1/154) based on an imperfect reference standard, and are thus not suitable for evaluating model performance.

	RF_BC	XGB_BC	DNN_BC	RF_SC	XGB_SC	DNN_SC
Standard deviations (during 5-fold CV)						
Recall	0.019	0.041	0.062	0.100	0.088	0.048
Specificity	0.254	0.196	0.143	0.131	0.129	0.173
BA [a]	0.121	0.091	0.058	0.040	0.049	0.080
G-mean	0.157	0.117	0.068	0.043	0.054	0.102
MCC	0.064	0.053	0.048	0.113	0.073	0.046
MCC						
5-fold CV [b]	0.113	0.125	0.160	0.271	0.197	0.128
Test set	0.083	0.091	0.096	0.140	0.110	0.086

[a] Balanced accuracy

[b] Average values across the five folds during 5-fold CV

Table S4. The eight strongest hERG blockers in our test set (IC_{50}s not exceeding 0.1 μM) and their most similar active ligands used in model training, in terms of Morgan fingerprints (2,048 bits, radius 2). The Tanimoto similarity scores are provided, proving the structural difference between them.

Strongest hERG blocker in the test set (SID)	Most similar active ligand in the training set (SID)	Tanimoto similarity
103626969	103528194	0.305
377003148	103609934	0.281
377003149	404958464	0.299
377003131	404958431	0.293
377003143	312363999	0.333
377003150	103609934	0.292
377003146	123089189	0.311
336951486	442142298	0.533