Supplementary Information

hERG toxicity prediction in early drug discovery using Extreme Gradient Boosting and Isometric Stratified Ensemble Mapping

Gabriela Falcón-Cano[1]; Aliuska Morales-Helguera[1]; Heather Lambert[1]; Miguel-Ángel Cabrera-Pérez[2] & Christophe Molina[1]*

[1]PIKAÏROS, S.A.; 31650, Saint Orens de Gameville, France
[2]Departamento de Ciencias Farmacéuticas, Facultad de Ciencias, Universidad Católica del Norte, Angamos, Antofagasta 0610, Chile

*Corresponding Author: E-mail: christophe.molina@pikairos.com

Table of Contents

- **Supplementary Information 5**. Detailed description of the model performance for ET I. (XLSX)
- **Supplementary Information 6**. Detailed description of the model performance for EV I. (XLSX)
- **Supplementary Information 7**. Detailed description of the model performance for EV II. (XLSX)
- **Supplementary Information 8**. Contents of the statistical results in Supplementary Information 4 – 7. (XLSX)

Statistics	Definition	Equation
Sensitivity or Recall	$SE = \frac{TP}{FN+TP}$	(1)
Specificity	$SP = \frac{TN}{FP + TN}$	(2)
Positive Predictive Value	$PPV = \frac{TP}{TP+FP}$	(3)
Negative Predictive Value	$NPV = \frac{TN}{TN+FN}$	(4)
F-Measure	$F - Measure\ = \frac{2}{SE^{-1} + PR^{-1}}$	(5)
Accuracy	$ACC = \frac{(TP+TN)}{(TP+TN+FP+FN)}$	(6)
Balanced Accuracy	$BACC = \frac{(\mathrm{SE} + \mathrm{SP})}{2}$	(7)
Geometric Mean	$G - Mean = (SE \cdot SP)^{1/2}$	(8)
Matthews Correlation Coefficient	$MCC = \frac{TP \cdot TN - FP \cdot FN}{((TP+FP)(TP+FN)(TN+FP)(TN+FN))^{1/2}}$	(9)
Cohen's kappa	$CK = \frac{ACC-PE}{1-PE}$	(10)

Table S1. Evaluation metrics used for classification. TP, TN, FP and FN are true positives, true negatives, false positives, and false negatives, respectively. PE is the hypothetical probability of chance agreement [1]. Statistics such as CK and MCC take values within [-1, 1], while the rest are within the interval [0, 1].

Molecular Descriptor	Description	Class	Median	Min	Max
ESOL	Estimated SOLubility (logS) for aqueous solubility using LOGPcons	Inhibitor	-4.67	-9.01	0.81
		Non-inhibitor	-3.83	-10.04	2.72
GATS3s	Geary autocorrelation of lag 3 weighted by I-state	Inhibitor	0.84	0.27	1.68
		Non-inhibitor	0.92	0.25	2.29
GATS5m	Geary autocorrelation of lag 5 weighted by mass	Inhibitor	0.93	0.24	2.15
		Non-inhibitor	1.01	0	3.55
GATS5s	Geary autocorrelation of lag 5 weighted by I-state	Inhibitor	1.01	0.25	2.99
		Non-inhibitor	1.2	0	4.16
GATS6m	Geary autocorrelation of lag 6 weighted by mass	Inhibitor	0.98	0.31	3.33
		Non-inhibitor	1.02	0	5.59
GATS6s	Geary autocorrelation of lag 6 weighted by I-state	Inhibitor	0.96	0.15	2.66
		Non-inhibitor	1.11	0	5.62
MATS1i	Moran autocorrelation of lag 1 weighted by ionization potential	Inhibitor	-0.14	-0.63	0.4
		Non-inhibitor	-0.11	-1.24	0.54
MATS5e	Moran autocorrelation of lag 5 weighted by Sanderson electronegativity	Inhibitor	-0.02	-0.62	0.46
		Non-inhibitor	-0.04	-1.07	0.92
MATS7m	Moran autocorrelation of lag 7 weighted by mass	Inhibitor	-0.03	-2.81	1.23
		Non-inhibitor	-0.05	-3.51	3.5
MaxssO	Maximum ssO	Inhibitor	4.88	0	6.89
		Non-inhibitor	4.78	0	7.11
mindssC	Mimimum dssC	Inhibitor	-0.01	-2.03	1.55
		Non-inhibitor	-0.26	-3.59	1.86
minssCH2	Mimimum ssCH2	Inhibitor	0.48	-1.84	1.37
		Non-inhibitor	0.2	-2.76	1.55
nRNH2	Number of primary amines (aliphatic)	Inhibitor	0	0	2
		Non-inhibitor	0	0	5
nRNHR	Number of secondary amines (aliphatic)	Inhibitor	0	0	2
		Non-inhibitor	0	0	4

nRNR2	Number of tertiary amines (aliphatic)	Inhibitor	1	0	4
		Non-inhibitor	0	0	4
P_VSA_charge_10	P_VSA-like on partial charges, bin 10	Inhibitor	5.24	0	61.25
		Non-inhibitor	4.84	0	134.56
P_VSA_charge_7	P_VSA-like on partial charges, bin 7	Inhibitor	50.8	0	208.73
		Non-inhibitor	42.97	0	313.63
P_VSA_LogP_5	P_VSA-like on LogP, bin 5	Inhibitor	37	0	183.45
		Non-inhibitor	26.94	0	225.05
P_VSA_MR_7	P_VSA- like on MR (molar refractivity), bin 7	Inhibitor	22.59	0	208.18
		Non-inhibitor	16.79	0	249.91
peoe_VSA8	P_VSA- like on PEOE (partial charges), bin 8	Inhibitor	30.71	0	96.87
		Non-inhibitor	17.92	0	109.14
SdssC	Sum of E-states of atoms of type =C< within a molecule	Inhibitor	0	-4.24	6.77
		Non-inhibitor	-0.22	-7.76	11.14
slogp_VSA3	P_VSA- like on LogP, bin 3	Inhibitor	10.44	0	55.7
		Non-inhibitor	10.02	0	71.5

Table S2. Sorted list of 22 variables selected after initial RVS for training the reduced model**.** Minimum and maximum values of the molecular descriptors for the inhibitor and non-inhibitor classes in the training set, along with their median values.

	Delre's model [2]	**DeepHIT [3]**	**CardioTox [4]**	**Feng et al. [5]**	**Ogura et al. [6]**	**This study**
BACC	0.76	0.79	0.51	0.75	0.83	0.87
SE	0.65	0.81	0.25	0.51	0.67	0.83
SP	0.86	0.78	0.76	0.99	0.99	0.91
MCC	0.22	0.25	0.0	0.66	0.73	0.41
Passed Molecules	85 746 [a]	87,306 [b]	87,306 [b]	87,361 [c]	87,361 [c]	87,306 [b]
Method	BRF + SVM + GBDT	DNN	DNN	GBDT + DNN	Weighted SVM	Balanced XGBoost Consensus
Available	KNIME Workflow	Python Code	Python Code	Python Code	Web Server	Web Server
Reprod	Yes	No	Yes	No	Yes	Yes

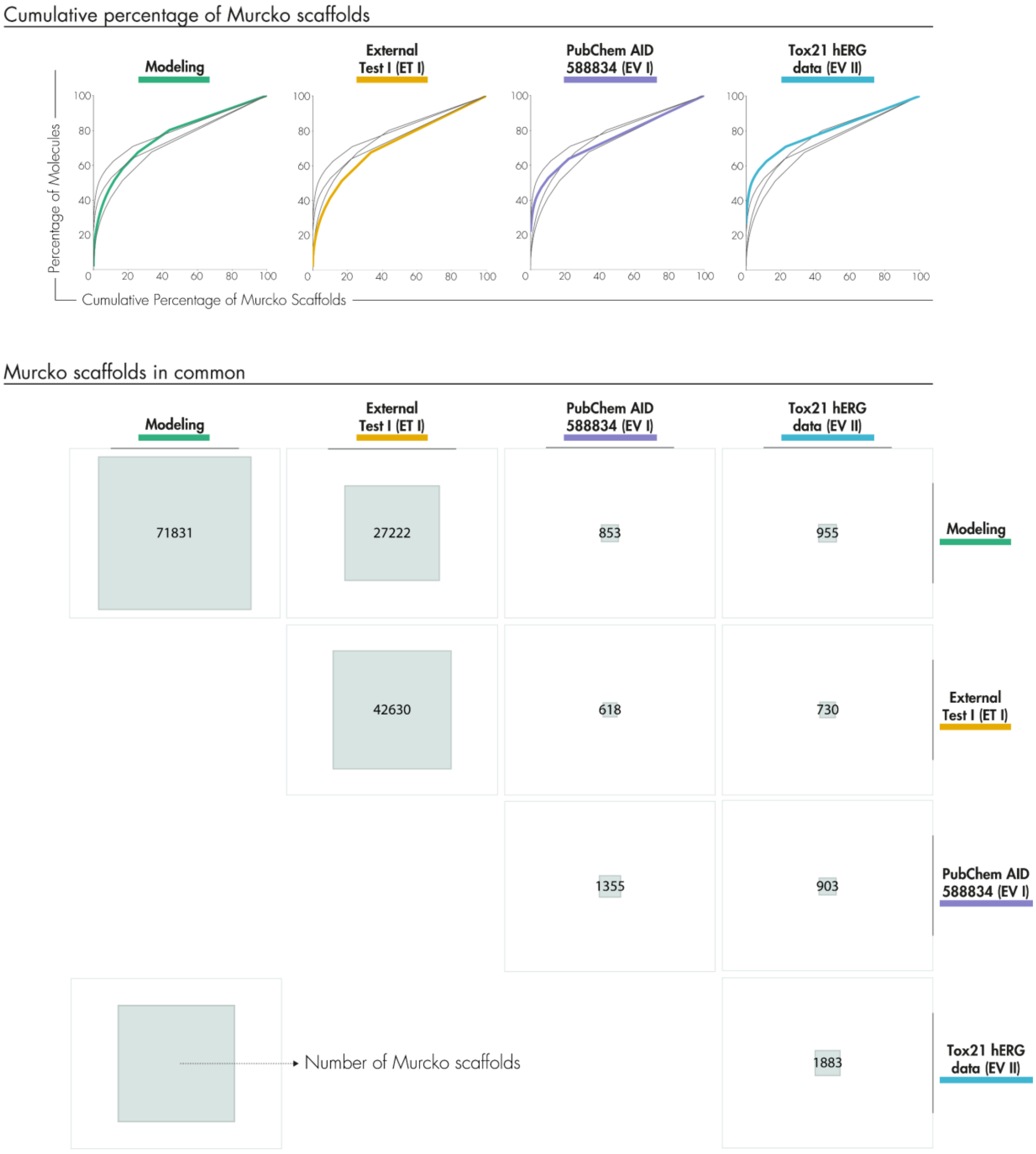
Table S3. Comparison of the XGBoost ensemble (majority vote as the output model) with other published models based on the external set (ET I) published in [6]. [a] 1,560 molecules were removed from the curated ET I (N=87,306) due to form part of the Training Set of Delre et al [2]. [c] This number reflects the original number of molecules in the External Set of Ogura et al [6]. [b] After curation of the original Ogura's External Set, we detected 55 duplicates, so the data size of the curated External Set is 87,306. BACC: Balanced Accuracy; SE: Sensitivity; SP: Specificity; MCC: Matthews correlation coefficient; and Reprod: Reproducibility, BRF: balanced random forest, SVM: Support Vector Machine, GBDT: Gradient Boosting Decision Tree, DNN: deep neural network

No.	Variables	Rank	Median (Rank)
1	peoe_VSA8	1, 1, 1	1
2	ESOL	2, 2, 2	2
3	SdssC	3, 3, 3	3
4	MaxssO	4, 4, 4	4
5	nRNR2	5, 6, 5	5
6	MATS1i	6, 9, 6	6
7	nRNHR	7, 7, 7	7
8	nRNH2	8, 8, 8	8

Table S4. Ordered list of the most important variables selected by the RVE procedure. The full method was repeated three times with shuffled data. The rank column shows the variable positions in each run.

ACC: Accuracy
AD: Applicability Domain
ADL: Applicability Domain Level
AUC: Area Under the Curve - used in the context of ROC curves, it represents the area under the curve plotting the true positive rate against the false positive rate
BRF: Balanced Random Forest
CiPA: Comprehensive In Vitro Pro-Arrhythmia Assay
CK: Cohen’s Kappa
CL: Consensus level
DL: Deep Learning
DT: Decision Tree
DNN: Deep Neural Network
ESOL: Estimated Solubility (logS) for aqueous solubility using LOGPcons
ETA: Extended topochemical atom
EV: External validation
FDA: Food and Drug Administration
FN: False Negatives
FP: False Positives
G-Mean: Geometric Mean of Sensitivity and Specificity
GBDT: Gradient Boosting Decision Tree
hERG: Human Ether-à-go-go-Related Gene
HTS: High throughput screening
IC_{50}: Inhibitory Concentration at 50%
ICH: International Council for Harmonization
INCHI: International Chemical Identifier
ISE: Isometric Stratified Ensemble
kNN: k Nearest Neighbors
MATS1i: Moran autocorrelation of lag 1 weighted by ionization potential
MaxssO: Maximum atom-type E-State: -O-
MBDS: Multiple Balanced Data Sampling
MCC: Matthews Correlation Coefficient
MOE: Molecular Operating Environment
NCATS: National Center for Advancing Translational Science
NN: Neural networks
NPV: Negative Predictive Value
nRNH2: Number of primary amines (aliphatic)
nRNHR: Number of secondary amines (aliphatic)
nRNR2: Number of tertiary amines (aliphatic)
OECD: Organization for Economic Co-operation and Development
peoe_VSA8: P_VSA- like on PEOE (partial charges), bin 8
PPV: Positive Predictive Value or Precision
QSAR: Quantitative Structure-Activity Relationship
R: Imbalance Ratio
ROC: Receiver Operating Characteristic
RVE: Recursive Variable Elimination
RVS: Recursive Decorrelated Variable Selection
SdssC: Sum of E-states of atoms of type =C< within a molecule
SE: Sensitivity
SI: Supplementary Information
SMILES: Simplified molecular-input line-entry system
SMOTE: Synthetic Minority over-sampling Technique
SP: Specificity
SVM: Support Vector Machine
TdP: Torsades de Pointes
TN: True negatives
TP: True Positives
VSA: van der Waals surface area
XGBoost: eXtreme Gradient Boosting

Table S5. Glossary of Terms



	Modeling	External Test I (ET I)	PubChem AID 588834 (EV I)	Tox21 hERG data (EV II)
Modeling	71831	27222	853	955
External Test I (ET I)		42630	618	730
PubChem AID 588834 (EV I)			1355	903
Tox21 hERG data (EV II)				1883

Figure S1. Murcko Scaffold Distribution and Overlap Across Datasets. This figure presents the distribution and overlap of Murcko scaffolds among four datasets: Modeling, External Test I (ET I), PubChem AID 588834 (EV I), and Tox21 hERG Data (EV II). The top panel shows cumulative percentage plots of Murcko scaffolds for each dataset, illustrating how scaffold diversity is distributed within each set. The bottom panel provides a matrix-like visualization of the number of shared Murcko scaffolds between datasets, where the size of each square is proportional to the number of common scaffolds. The largest overlap is observed between Modeling and External Test I (ET I), while smaller intersections appear for PubChem and Tox21 datasets. This visualization highlights scaffold diversity across datasets and their structural similarities, which are crucial for assessing the applicability and generalizability of predictive models.

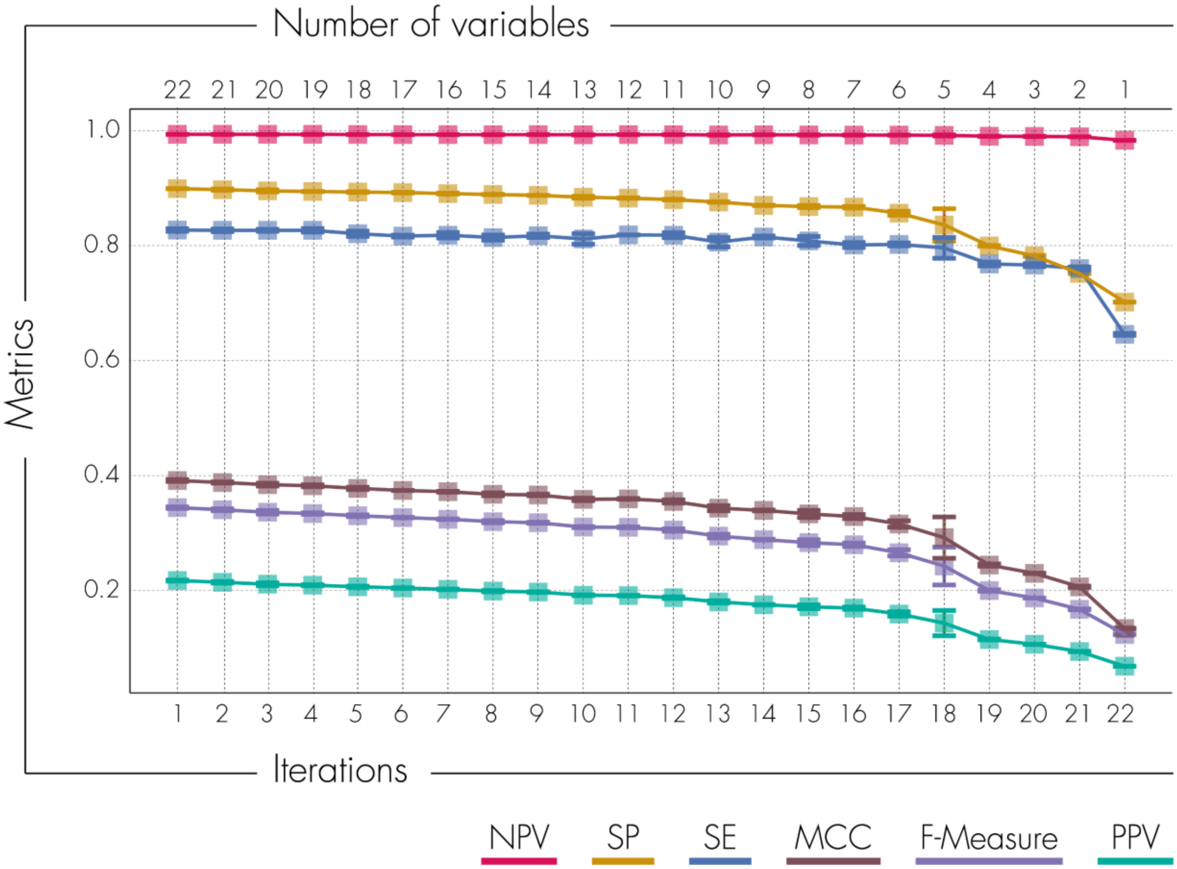
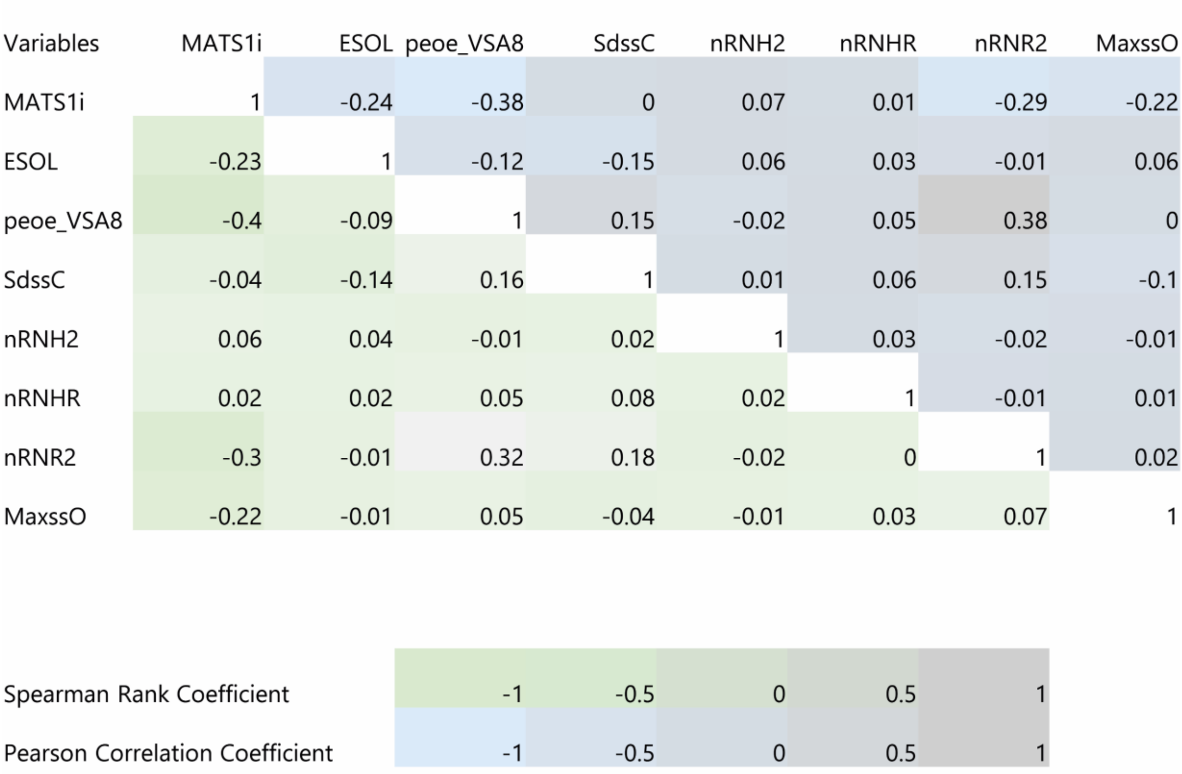


Figure S2. Recursive Variable Elimination (RVE) applied iteratively to the Recursive Decorrelated Variable Selection (RVS) -selected descriptors to identify the minimum number of variables required for effective hERG inhibition prediction. Error bars represent one standard deviation for each metric, with results based on internal set predictions. The bottom x-axis represents the iteration number, and the top x-axis the number of remaining variables at each iteration, while the y-axis represents the metric values, which range from 0 to 1. RVS model performance on the internal test set was evaluated across multiple iterations and metrics. Early on, metrics like; selectivity (SE) and specificity (SP) remained stable above 0.8, but dropped sharply after 18 iterations as variables were eliminated. A statistical comparison over 22 iterations, using the Friedman test and Dunn-Bonferroni test, identified that significant differences from iteration 0 first appear in the iteration 15, according to Matthews Correlation Coefficient (MCC), the measure analyzed (p < 0.05). This result suggests that the model at iteration 14 is not statistically different from the model trained with 22 descriptors, indicating that only 8 variables are relevant to predict the target variable, using the current methodology.



Variables	MATS1i	ESOL	peoe_VSA8	SdssC	nRNH2	nRNHR	nRNR2	MaxssO
MATS1i	1	-0.24	-0.38	0	0.07	0.01	-0.29	-0.22
ESOL	-0.23	1	-0.12	-0.15	0.06	0.03	-0.01	0.06
peoe_VSA8	-0.4	-0.09	1	0.15	-0.02	0.05	0.38	0
SdssC	-0.04	-0.14	0.16	1	0.01	0.06	0.15	-0.1
nRNH2	0.06	0.04	-0.01	0.02	1	0.03	-0.02	-0.01
nRNHR	0.02	0.02	0.05	0.08	0.02	1	-0.01	0.01
nRNR2	-0.3	-0.01	0.32	0.18	-0.02	0	1	0.02
MaxssO	-0.22	-0.01	0.05	-0.04	-0.01	0.03	0.07	1

Spearman Rank Coefficient	-1	-0.5	0	0.5	1
Pearson Correlation Coefficient	-1	-0.5	0	0.5	1

Figure S3. Correlation coefficients (Pearson and Spearman Rank) between each pair of variables used in the lowest dimensionality model (8 variables).

References

1. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* **20**, 37–46 (1960).
2. Delre, P. *et al.* Ligand-based prediction of hERG-mediated cardiotoxicity based on the integration of different machine learning techniques. *Front. Pharmacol.* **13**, 951083 (2022).
3. Ryu, J. Y., Lee, M. Y., Lee, J. H., Lee, B. H. & Oh, K.-S. DeepHIT: a deep learning framework for prediction of hERG-induced cardiotoxicity. *Bioinformatics* **36**, 3049–3055 (2020).
4. Karim, A., Lee, M., Balle, T. & Sattar, A. CardioTox net: a robust predictor for hERG channel blockade based on deep learning meta-feature ensembles. *J. Cheminform.* **13**, 60 (2021).
5. Feng, H. & Wei, G.-W. Virtual screening of DrugBank database for hERG blockers using topological Laplacian-assisted AI models. *Comput. Biol. Med.* **153**, 106491 (2023).
6. Ogura, K., Sato, T., Yuki, H. & Honma, T. Support Vector Machine model for hERG inhibitory activities based on the integrated hERG database using descriptor selection by NSGA-II. *Sci Rep* **9**, 12220 (2019).