Supplemental Material

Predicting hERG Channel Blockers with Directed Message Passing Neural Networks

Mengyi Shan,[//,†] Chen Jiang,[//,‡] Jing Chen,[†,⊥] Lu-Ping Qin,[,†] Jiang-Jiang Qin,[*,§]*

and Gang Cheng[,†]*

Table of content

For a better understanding of Table 1 and Table 2, we sorted out the size, the ratio of blockers to non-blockers and the source of the whole database used in the article in Table S1.

Table S1. Detailed description of the all databases.

Refs.	Dataset			Data source
Cai et al.[1]	training size: 6311 (P: N=3,485 : 2826)	test size: 789 (P: N=435 : 354)	validation size: 789 (P: N=435 : 354)	ChEMBL(patch-clamp), radioligand binding measurements，hERG K+ channel binding affinity, and literature derived
Doddareddy et al.[2]	D3 training size: 2389 (P: N=1,004 : 1,385)	D3 test size: 155 (P:N=108 :147)		literature derived
Siramshetty et al.[3]	training size: 8154 (P: N=2,164:5, 990)	validation size: 839 (P: N=53 : 786)		ChEMBL (multiple), NCATS (flux assay), FDA Pharmacological and Safety Reviews (patch-clamp)
	training size: 8154 (P: N=2,164:5, 990)	FDA-1 size: 177 (P: N=15 : 162)		
Hou et al. [4]	training I size: 392 (P: N=352 : 40)	test I size: 195 (P: N=175 : 20)		patch-clamp
	training II size: 392 (P: N=352 : 40)	test II size: 195 (P: N=175 : 20)		
Ogura et al.[5]	training size: 203,941 (P: N=6,923 : 196,918)	test size: 87,361 (P: N=2,966 : 84,395)		ChEMBL, GOSTAR, NCGC, hERGCentral, hERG integrated dataset
Karim et al.[6]	training size: 12620 (P: N=6,643 : 5, 977)	Test-set I size: 44 (P: N=30 : 14)		BindingDB database, ChEMBL, and literature derived
	training size: 12620 (P: N=6,643 : 5, 977)	Test-set II size: 44 (P: N=11 : 30)		
	training size: 12620 (P: N=6,643 : 5, 977)	Test-set III size: 740 (P: N=30 : 710)		

P: Positive; N: Negative; NCATS: National Center for Advancing Translational Sciences; NCGC: NIH Chemical Genomics Center.

The predictive ability and applicable scope of the model are greatly related to the chemical spatial diversity of the data set. To visualize the distribution of datasets in the chemical space, two-dimensional (2D) principal component analysis[7] (PCA) was employed on different datasets with ECFP descriptors as input. As shown in Figure S1, high chemical diversity and overlap among the compounds within Cai's[1], Siramshetty's[3] and Hou's[4] datasets reflects the effectiveness of the applicability domain of the model.[8] Moreover, the PCA approach was also implemented on the Cai's, Siramshetty's and Hou's dataset with threshold values of 20 μM in Figure S2 to indicate that ECFP contribute to distinguishing the blockers and non-blockers to some extent.

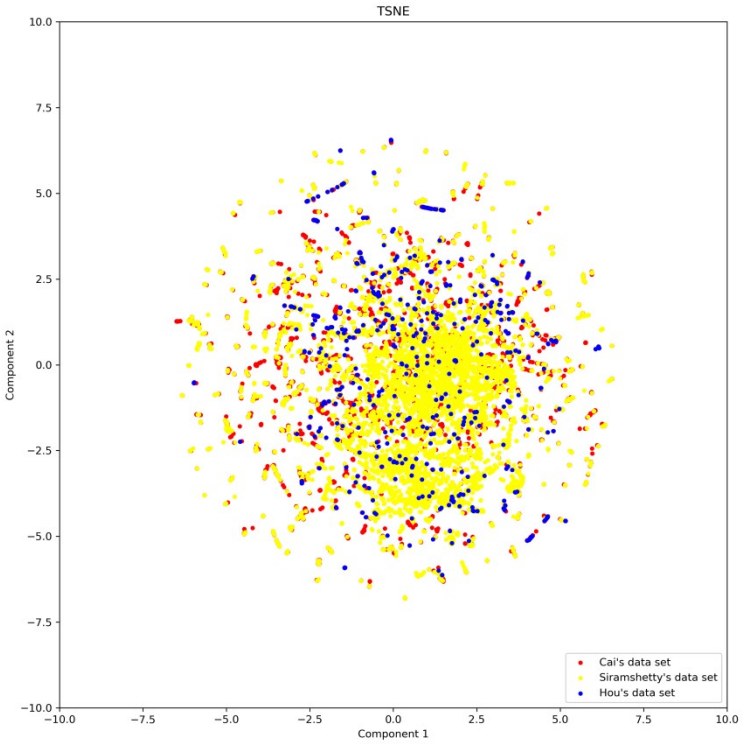

Figure S1. Two-dimensional principal component analysis plots based on Cai's (red), Siramshetty's (blue) and Hou's data sets (yellow).

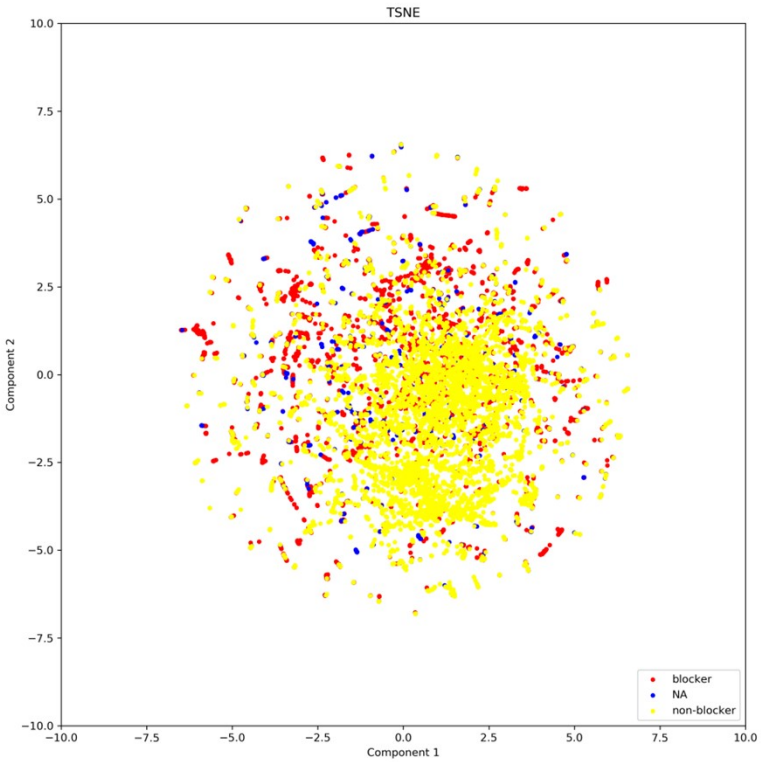

Figure S2. Two-dimensional principal component analysis plots based on blocker (red), NA (blue) and non-blocker (yellow). NA: not available

Table S2. The 206 descriptors in moe206.

descriptor	description	descriptor	description
apol	Sum of atomic polarizabilities	vdw_vol	Van der Waals volume (A**3)
ast_fraglike	Astex Fragment-like Test	BCUT_PEOE_0	PEOE Charge BCUT (0/3)
ast_fraglike_ext	Astex Fragment-like Test (Extended)	BCUT_PEOE_1	PEOE Charge BCUT (1/3)
ast_violation	Astex Fragment-like Violation Count	BCUT_PEOE_2	PEOE Charge BCUT (2/3)
ast_violation_ext	Astex Fragment-like Violation Count (Extended)	BCUT_PEOE_3	PEOE Charge BCUT (3/3)
a_acc	Number of H-bond acceptor atoms	BCUT_SLOGP_0	LogP BCUT (0/3)
a_acid	Number of acidic atoms	BCUT_SLOGP_1	LogP BCUT (1/3)
a_aro	Number of aromatic atoms	BCUT_SLOGP_2	LogP BCUT (2/3)
a_base	Number of basic atoms	BCUT_SLOGP_3	LogP BCUT (3/3)
a_count	Number of atoms	BCUT_SMR_0	Molar Refractivity BCUT (0/3)
a_don	Number of H-bond donor atoms	BCUT_SMR_1	Molar Refractivity BCUT (1/3)
a_donacc	Number of H-bond donor + acceptor atoms	BCUT_SMR_2	Molar Refractivity BCUT (2/3)
a_heavy	Number of heavy atoms	BCUT_SMR_3	Molar Refractivity BCUT (3/3)
a_hyd	Number of hydrophobic atoms	bpol	Difference of bonded atom polarizabilities
a_IC	Atom information content (total)	b_1rotN	Number of rotatable single bonds
a_ICM	Atom information content (mean)	b_1rotR	Fraction of rotatable single bonds
a_nB	Number of boron atoms	b_ar	Number of aromatic bonds
a_nBr	Number of bromine atoms	b_count	Number of bonds
a_nC	Number of carbon atoms	b_double	Number of double bonds
a_nCl	Number of chlorine atoms	b_heavy	Number of heavy-heavy bonds
a_nF	Number of fluorine atoms	b_max1len	Maximum single-bond chain length
a_nH	Number of hydrogen atoms	b_rotN	Number of rotatable bonds
a_nI	Number of iodine atoms	b_rotR	Fraction of rotatable bonds
a_nN	Number of nitrogen atoms	b_single	Number of single bonds
a_nO	Number of oxygen atoms	b_triple	Number of triple bonds
a_nP	Number of phosphorus atoms	chi0	Atomic connectivity index (order 0)
a_nS	Number of sulfur atoms	chi0v	Atomic valence connectivity index (order 0)
balabanJ	Balaban averaged distance sum connectivity	chi0v_C	Carbon valence connectivity index (order 0)
chi1	Atomic connectivity index (order 1)	chi0_C	Carbon connectivity index (order 0)
chi1v	Atomic valence connectivity index (order 1)	h_log_pbo	Sum of log (1 + p-bond orders)
chi1v_C	Carbon valence connectivity index (order 1)	h_mr	Molar Refractivity
chi1_C	Carbon connectivity index (order 1)	h_pavgQ	Average total charge (pH=7)
chiral	Number of chiral centers	h_pKa	Acidity (pH=7)
chiral_u	Number of unconstrained chiral centers	h_pKb	Basicity (pH=7)
density	Mass density (AMU/A**3)	h_pstates	Entropic state count (pH=7)
diameter	Largest vertex eccentricity in graph	h_pstrain	Protonation state strain energy (pH=7)
FCharge	Sum of formal charges	Kier1	First kappa shape index

GCUT_PEOE_0	PEOE Charge GCUT (0/3)	Kier2	Second kappa shape index
GCUT_PEOE_1	PEOE Charge GCUT (1/3)	Kier3	Third kappa shape index
GCUT_PEOE_2	PEOE Charge GCUT (2/3)	KierA1	First alpha modified shape index
GCUT_PEOE_3	PEOE Charge GCUT (3/3)	KierA2	Second alpha modified shape index
GCUT_SLOGP_0	LogP GCUT (0/3)	KierA3	Third alpha modified shape index
GCUT_SLOGP_1	LogP GCUT (1/3)	KierFlex	Molecular flexibility
GCUT_SLOGP_2	LogP GCUT (2/3)	lip_acc	Lipinski Acceptor Count
GCUT_SLOGP_3	LogP GCUT (3/3)	lip_don	Lipinski Donor Count
GCUT_SMR_0	Molar Refractivity GCUT (0/3)	lip_druglike	Lipinski Druglike Test
GCUT_SMR_1	Molar Refractivity GCUT (1/3)	lip_violation	Lipinski Violation Count
GCUT_SMR_2	Molar Refractivity GCUT (2/3)	logP(o/w)	Log octanol/water partition coefficient
GCUT_SMR_3	Molar Refractivity GCUT (3/3)	logS	Log Solubility in Water
h_ema	Sum of EHT acceptor strengths	mr	Molar refractivity
h_emd	Sum of EHT donor strengths	mutagenic	Mutagenicity
h_emd_C	Sum of EHT carbon donor strengths	nmol	Number of molecules
h_logD	Octanol/water distribution coefficient (pH=7)	opr_brigid	Oprea Rigid Bond Count
h_logP	Octanol/water partition coefficient	opr_leadlike	Oprea Leadlike Test
h_logS	Log solubility in water	opr_nring	Oprea Ring Count
h_log_dbo	Sum of log (1 + d-bond orders)	opr_nrot	Oprea Rotatable Bond Count
PC+	Total positive partial charge	opr_violation	Oprea Violation Count
PC-	Total negative partial charge	PEOE_VSA_PNEG	Total polar negative vdw surface area
PEOE_PC+	Total positive partial charge	PEOE_VSA_POL	Total polar vdw surface area
PEOE_PC-	Total negative partial charge	PEOE_VSA_POS	Total positive vdw surface area
PEOE_RPC+	Relative positive partial charge	PEOE_VSA_PPOS	Total polar positive vdw surface area
PEOE_RPC-	Relative negative partial charge	petitjean	(diameter - radius) / diameter
PEOE_VSA+0	Total positive 0 vdw surface area	petitjeanSC	(diameter - radius) / radius
PEOE_VSA+1	Total positive 1 vdw surface area	Q_PC+	Total positive partial charge
PEOE_VSA+2	Total positive 2 vdw surface area	Q_PC-	Total negative partial charge
PEOE_VSA+3	Total positive 3 vdw surface area	Q_RPC+	Relative positive partial charge
PEOE_VSA+4	Total positive 4 vdw surface area	Q_RPC-	Relative negative partial charge
PEOE_VSA+5	Total positive 5 vdw surface area	Q_VSA_FHYD	Fractional hydrophobic vdw surface area
PEOE_VSA+6	Total positive 6 vdw surface area	Q_VSA_FNEG	Fractional negative vdw surface area
PEOE_VSA-0	Total negative 0 vdw surface area	Q_VSA_FPNEG	Fractional polar negative vdw surface area
PEOE_VSA-1	Total negative 1 vdw surface area	Q_VSA_FPOL	Fractional polar vdw surface area
PEOE_VSA-2	Total negative 2 vdw surface area	Q_VSA_FPOS	Fractional positive vdw surface area
PEOE_VSA-3	Total negative 3 vdw surface area	Q_VSA_FPPOS	Fractional polar positive vdw surface area

PEOE_VSA-4	Total negative 4 vdw surface area	Q_VSA_HYD	Total hydrophobic vdw surface area
PEOE_VSA-5	Total negative 5 vdw surface area	Q_VSA_NEG	Total negative vdw surface area
PEOE_VSA-6	Total negative 6 vdw surface area	Q_VSA_PNEG	Total polar negative vdw surface area
PEOE_VSA_FHYD	Fractional hydrophobic vdw surface area	Q_VSA_POL	Total polar vdw surface area
PEOE_VSA_FNEG	Fractional negative vdw surface area	Q_VSA_POS	Total positive vdw surface area
PEOE_VSA_FPNEG	Fractional polar negative vdw surface area	Q_VSA_PPOS	Total polar positive vdw surface area
PEOE_VSA_FPOL	Fractional polar vdw surface area	radius	Smallest vertex eccentricity in graph
PEOE_VSA_FPOS	Fractional positive vdw surface area	reactive	Reactivity
PEOE_VSA_FPPOS	Fractional polar positive vdw surface area	rings	Number of rings
PEOE_VSA_HYD	Total hydrophobic vdw surface area	RPC+	Relative positive partial charge
PEOE_VSA_NEG	Total negative vdw surface area	RPC-	Relative negative partial charge
SlogP	Log Octanol/Water Partition Coefficient	rsynth	Synthetic Feasibility
SlogP_VSA0	Bin 0 SlogP (-10 ,-0.40]	vsa_acid	VDW acidic surface area (A**2)
SlogP_VSA1	Bin 1 SlogP (-0.40,-0.20]	vsa_base	VDW basic surface area (A**2)
SlogP_VSA2	Bin 2 SlogP (-0.20, 0.00]	vsa_don	VDW donor surface area (A**2)
SlogP_VSA3	Bin 3 SlogP (0.00, 0.10]	vsa_hyd	VDW hydrophobe surface area (A**2)
SlogP_VSA4	Bin 4 SlogP (0.10, 0.15]	vsa_other	VDW other surface area (A**2)
SlogP_VSA5	Bin 5 SlogP (0.15, 0.20]	vsa_pol	VDW polar surface area (A**2)
SlogP_VSA6	Bin 6 SlogP (0.20, 0.25]	Weight	Molecular weight (CRC)
SlogP_VSA7	Bin 7 SlogP (0.25, 0.30]	weinerPath	Weiner path number
SlogP_VSA8	Bin 8 SlogP (0.30, 0.40]	weinerPol	Weiner polarity number
SlogP_VSA9	Bin 9 SlogP (0.40,10]	zagreb	Zagreb index
SMR_VSA0	Bin 0 SMR (0.000,0.110]	SMR	Molar Refractivity
SMR_VSA2	Bin 2 SMR (0.260,0.350]	SMR_VSA1	Bin 1 SMR (0.110,0.260]
SMR_VSA4	Bin 4 SMR (0.390,0.440]	SMR_VSA3	Bin 3 SMR (0.350,0.390]
SMR_VSA6	Bin 6 SMR (0.485,0.560]	SMR_VSA5	Bin 5 SMR (0.440,0.485]
TPSA	Topological Polar Surface Area (A**2)	SMR_VSA7	Bin 7 SMR (0.560,10]
VAdjMa	Vertex adjacency information (mag)	VAdjEq	Vertex adjacency information (equal)
VDistMa	Vertex distance magnitude index	VDistEq	Vertex distance equality index
vsa_acc	VDW acceptor surface area (A**2)	vdw_area	Van der Waals surface area (A**2)

Table S3. The 53 descriptors in MOE53.

descriptor	descriptor
LabuteASA	SlogP_VSA5
PEOE_VSA1	SlogP_VSA6
PEOE_VSA2	SlogP_VSA7
PEOE_VSA3	SlogP_VSA8
PEOE_VSA4	SlogP_VSA9
PEOE_VSA5	SlogP_VSA10
PEOE_VSA6	SlogP_VSA11
PEOE_VSA7	EState_VSA1
PEOE_VSA8	EState_VSA2
PEOE_VSA9	EState_VSA3
PEOE_VSA10	EState_VSA4
PEOE_VSA11	EState_VSA5
PEOE_VSA12	EState_VSA6
PEOE_VSA13	EState_VSA7
SMR_VSA1	EState_VSA8
SMR_VSA2	EState_VSA9
SMR_VSA3	EState_VSA10
SMR_VSA4	VSA_EState1
SMR_VSA5	VSA_EState2
SMR_VSA6	VSA_EState3
SMR_VSA7	VSA_EState4
SMR_VSA8	VSA_EState5
SMR_VSA9	VSA_EState6
SlogP_VSA1	VSA_EState7
SlogP_VSA2	VSA_EState8
SlogP_VSA3	VSA_EState9
SlogP_VSA4	

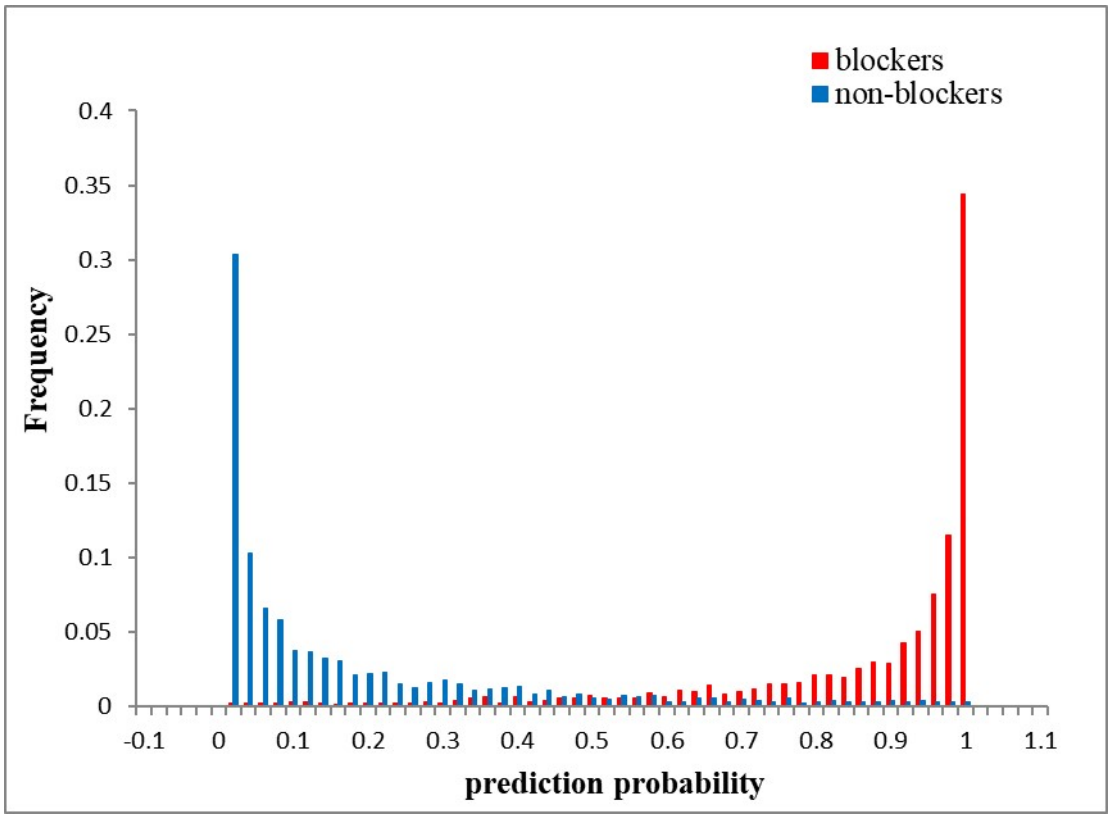

Figure S3. The prediction probability distribution of hERG blockers for Cai's dataset (at threshold = 10 μM) by D-MPNN.

Cardiotoxicity caused by hERG blocked is one of the most significant aspects of ADMET properties[9], and many physico-chemical properties are closely related to ADMET, So we applied six molecular descriptors to ADMET predictions, includes the number of rotatable bonds (NumRot), the number of hydrogen bond acceptor (HBD), molecular weight (MW), the number of molecular hydrogen bond donor (HBA), the octanol/water partition coefficient (LogP), the number of heavyatoms (NumHeavyAtoms). Comparing with blockers, non-blockers have more hydrogen bond acceptors and donors, while a small logP indicates that non-inhibitors are more hydrophilic. In general, for any descriptor, there is still a large overlap in the distribution of inhibitors and non-inhibitors, which indicates that no single molecular property can effectively distinguish between inhibitors and non-inhibitors.

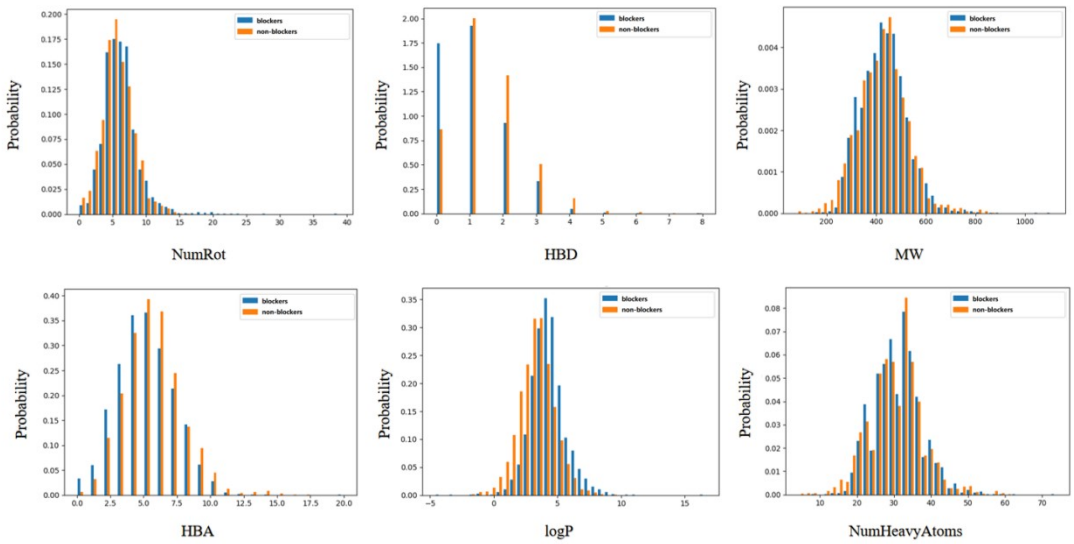

Figure S4. Comparing the chemical properties of blockers and non-blockers. NumRot: the number of rotatable bonds; HBD: the number of hydrogen bond acceptor; MW: molecular weight; HBA: the number of molecular hydrogen bond donor; LogP: the octanol/water partition coefficient; NumHeavyAtoms: the number of heavyatoms.

Table S4. The full name, bits and source of the descriptors.

abbreviation	full	bits	refs. (DOI)
ECFP4	Extended Connectivity Fingerprint	2048	10.1021/ci100050t
ECFP6	Extended Connectivity Fingerprint	4096	10.1021/ci100050t
FCFP4	Functional class fingerprints	2048	10.1021/ci100050t
MACCS	Molecular Accession System keys	166	10.1021/ci010132r
PubchemFP	PubChem fingerprint	881	list_fingerprints.pdf (ohio-state.edu)
RDkit2D	RDkit 2D normalized	200	https://rdkit.org/docs/index.html
MOE53	Molecular Operating Environment	53	10.1186/s13321-018-0258-y
moe206	Molecular Operating Environment	206	10.1007/s10822-012-9570-1
Mol2vec	Mol2vec	300	10.1021/acs.jcim.7b00616

1. Cai, C.; Guo, P.; Zhou, Y.; Zhou, J.; Wang, Q.; Zhang, F.; Fang, J.; Cheng, F. Deep Learning-Based Prediction of Drug-Induced Cardiotoxicity. *J. Chem. Inf. Model.* **2019,** *59*, (3), 1073-1084.
2. Doddareddy, M. R.; Klaasse, E. C.; Shagufta; Ijzerman, A. P.; Bender, A. Prospective validation of a comprehensive in silico hERG model and its applications to commercial compound and drug databases. *ChemMedChem* **2010,** *5*, (5), 716-29.
3. Siramshetty, V. B.; Nguyen, D. T.; Martinez, N. J.; Southall, N. T.; Simeonov, A.; Zakharov, A. V. Critical Assessment of Artificial Intelligence Methods for Prediction of hERG Channel Inhibition in the "Big Data" Era. *J. Chem. Inf. Model.* **2020,** *60*, (12), 6007-6019.
4. Wang, S. Q.; Sun, H. Y.; Liu, H.; Li, D.; Li, Y. Y.; Hou, T. J. ADMET Evaluation in Drug Discovery. 16. Predicting hERG Blockers by Combining Multiple Pharmacophores and Machine Learning Approaches. *Mol. Pharm.* **2016,** *13*, (8), 2855-2866.
5. 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.* **2019,** *9*, (1), 12220.
6. 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. Cheminformatics* **2021,** *13*, (1), 60.
7. Ma, S.; Dai, Y. Principal component analysis based methods in bioinformatics studies. *Brief. Bioinformatics* **2011,** *12*, (6), 714-22.
8. Weaver, S.; Gleeson, M. P. J. J. o. M. G.; Modelling. The importance of the domain of applicability in QSAR modeling. **2008,** *26*, (8), 1315-1326.
9. Ferreira, L. L. G.; Andricopulo, A. D. ADMET modeling approaches in drug discovery. *Drug Discov. Today* **2019,** *24*, (5), 1157-1165.