SUPPORTING INFORMATION

Toward Reducing hERG Affinities for DAT Inhibitors with a Combined Machine Learning and Molecular Modelling Approach

Kuo Hao Lee[1,#], Andrew D. Fant[1,#], Jiqing Guo[2], Andy Guan[1], Joslyn Jung[1], Mary Kudaibergenova[3], Williams E. Miranda[3], Therese Ku[4], Jianjing Cao[4], Soren Wacker[2,3,5], Henry J. Duff[2], Amy Hauck Newman[4], Sergei Y. Noskov[3], Lei Shi[1,*]

[1]Computational Chemistry and Molecular Biophysics Section, Molecular Targets and Medications Discovery Branch, National Institute on Drug Abuse – Intramural Research Program, National Institutes of Health, Baltimore, MD 21224, USA

[2]Libin Cardiovascular Institute of Alberta, Cumming School of Medicine, University of Calgary, Calgary, AB T2N 4N1, Canada

[3]Centre for Molecular Simulation, Department of Biological Sciences, University of Calgary, Calgary, AB T2N 1N4, Canada

[4]Medicinal Chemistry Section, Molecular Targets and Medications Discovery Branch, National Institute on Drug Abuse – Intramural Research Program, National Institutes of Health, Baltimore, MD 21224, USA

[5]Achlys Inc., 7-126 Li Ka Shing Center for Health and Innovation, Edmonton, AB T6G 2E1, Canada

[#]contributed equally

*corresponding author: lei.shi2@nih.gov

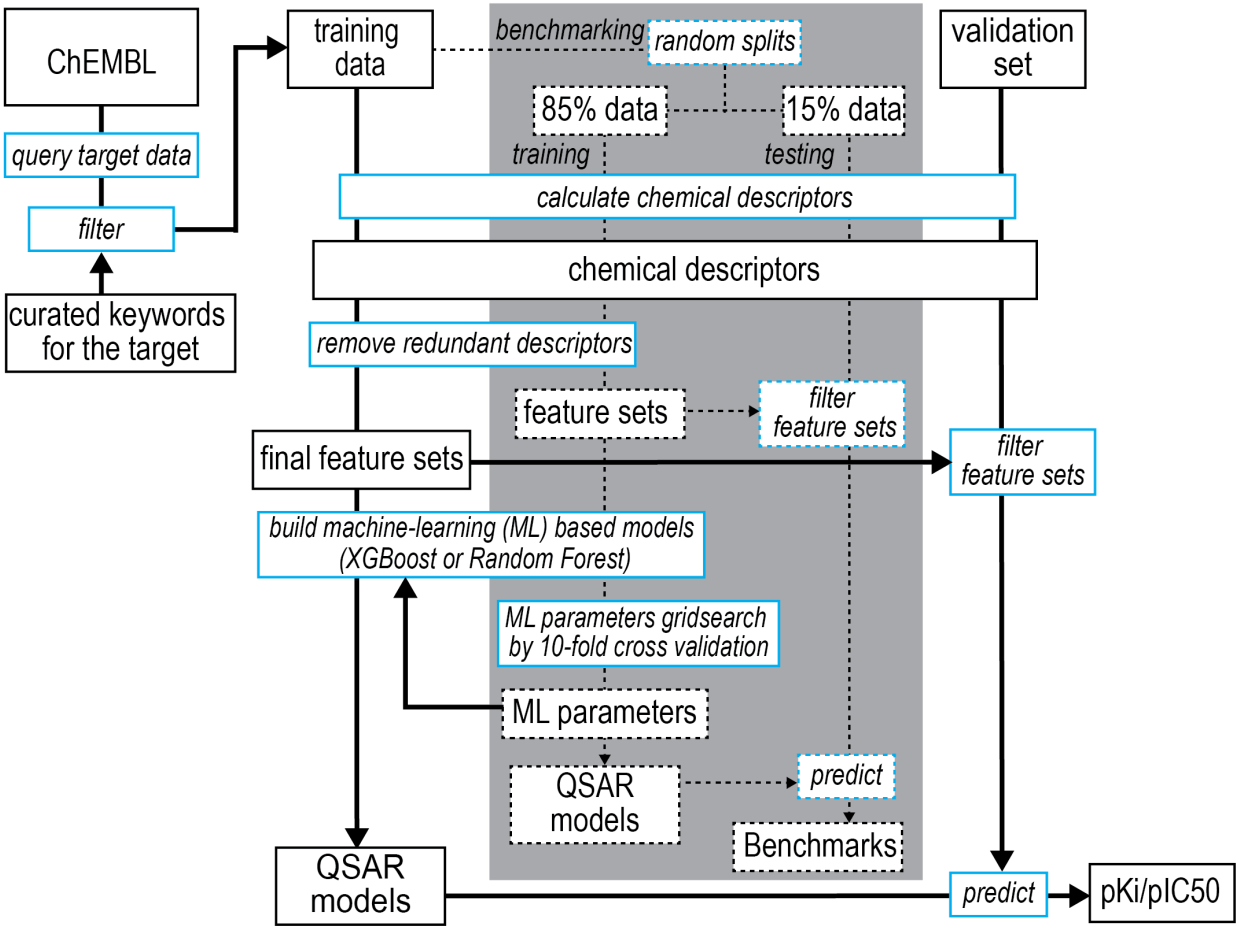

Figure S1. The workflow of building the machine learning-based QSAR models and using them to make predictions. Black boxes indicate data, blue boxes denote operations.

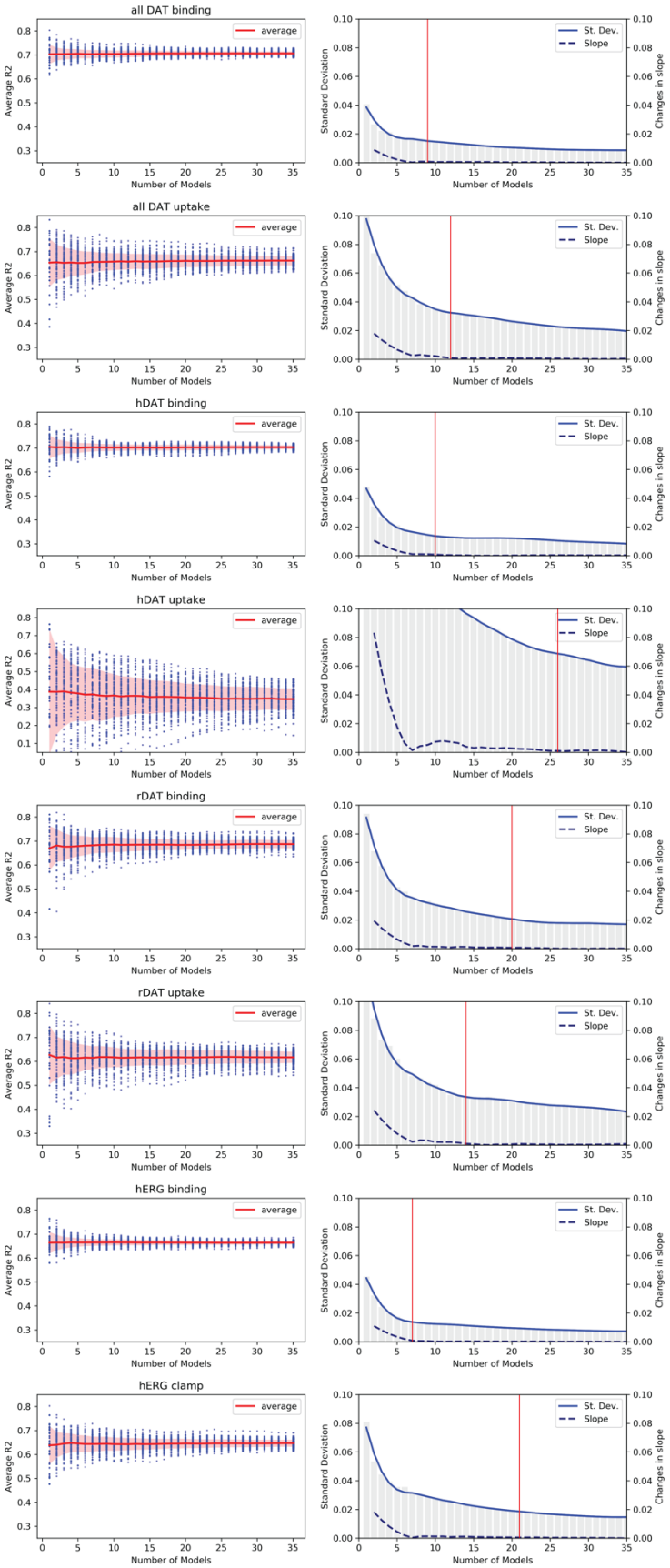
all DAT binding
all DAT uptake
hDAT binding
hDAT uptake
rDAT binding
rDAT uptake
hERG binding
hERG clamp
average
St. Dev.
Slope
Average R2
Standard Deviation
Changes in slope
Number of Models

Figure S2. The number of models needed for each set of XGBoost regression models. For each set of models, we randomly select certain number (n=1, …, 35) of models by a bootstrapping sampling with 100 repeats, and calculated the averages (red curves) and standard deviations (pink areas) of R^2 values for n models. On the right panels, slope is the change of the standard deviations of R^2 shown on the corresponding left panels for each number of models. The number of models for each dataset that starts to have the slope <0.001 was indicated by red bar, which we define as the converging number of models.

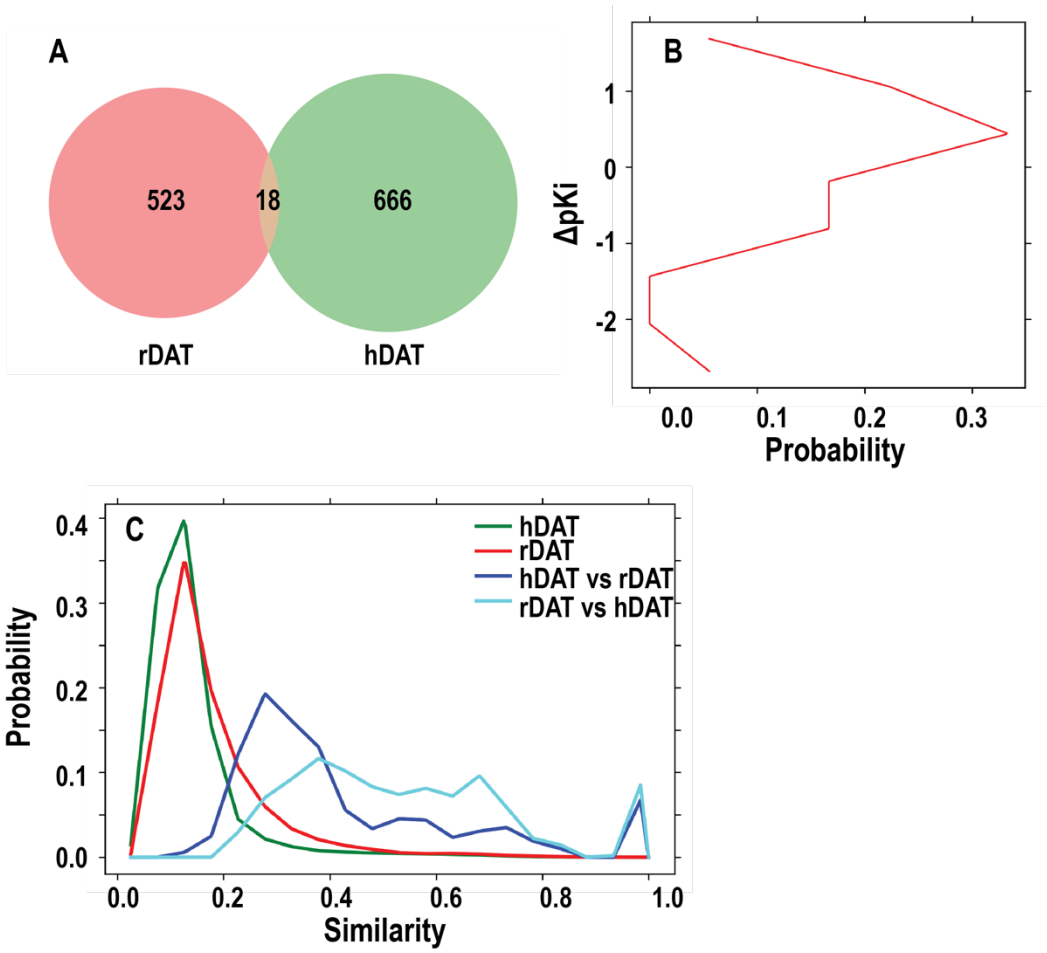

Figure S3. Comparison of the rDAT and hDAT binding datasets. (A) rDAT and hDAT binding datasets have 18 overlapping molecules. (B) Distribution of ΔpKi of 18 molecules. The ΔpKi was calculated using $pKi_{rDAT} - pKi_{hDAT}$. (C) Distribution of pairwise similarity. The Tanimoto similarity was calculated based on the Morgan fingerprint of each molecule.

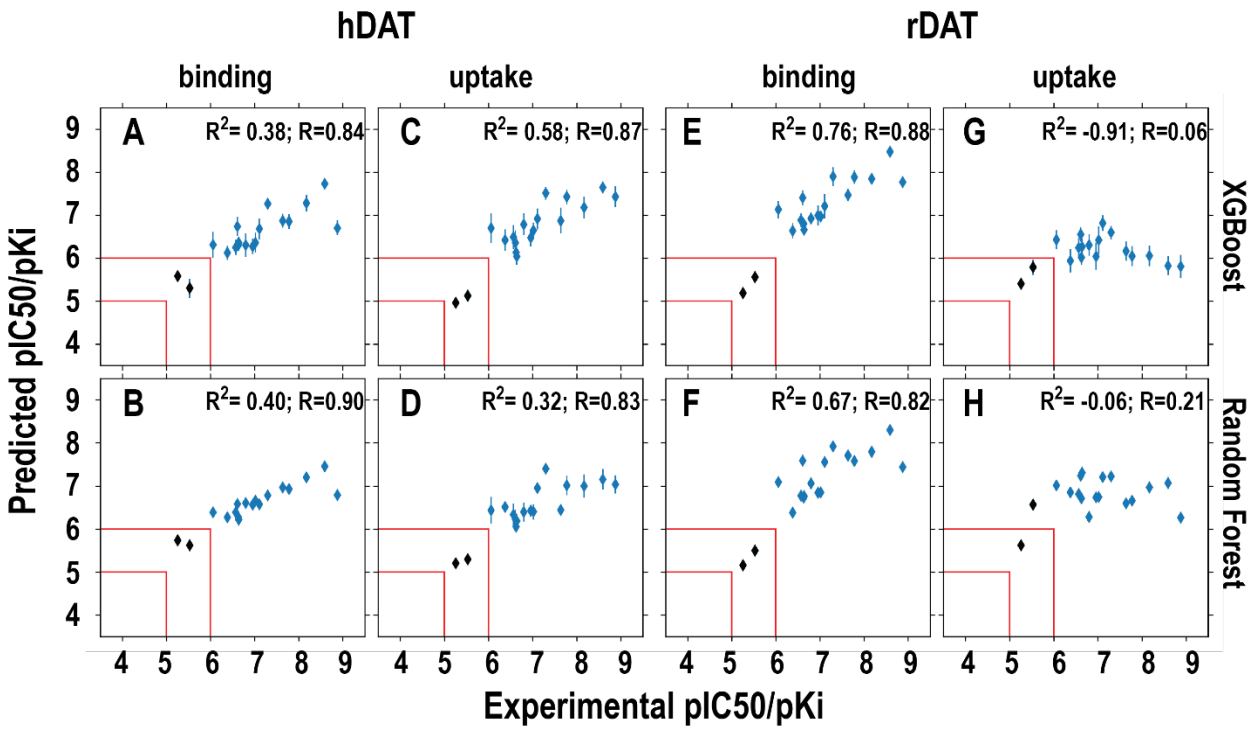

Figure S4. Correlations between the predicted and experimentally measured hDAT and rDAT affinities. See Figure 3 for the color scheme.

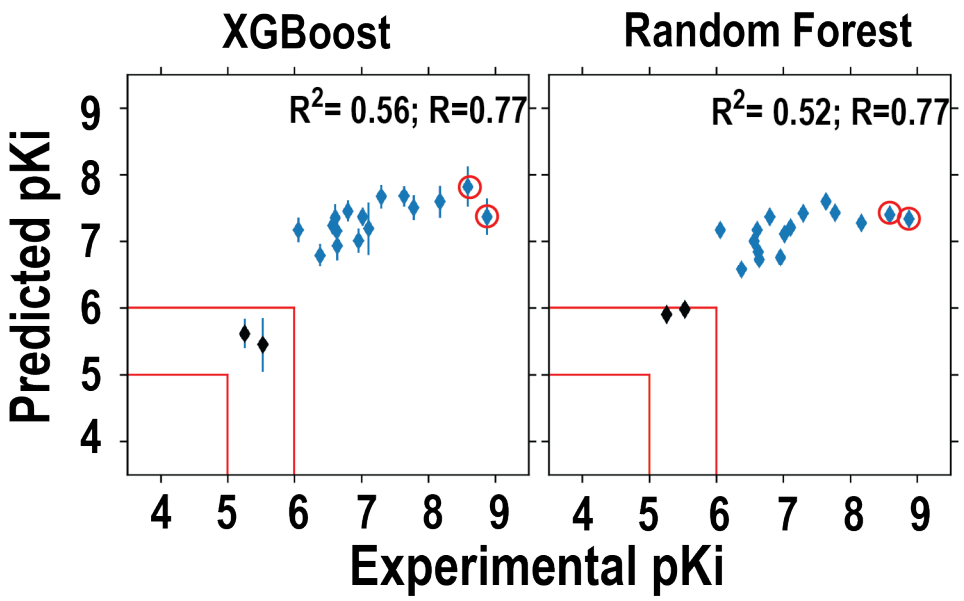

Figure S5. Correlations between the predicted and experimentally measured DAT affinities using models built with the in-house DAT dataset. When removing the two high affinity points (pKi ≥ 8.5, highlighted with red circles), the R values increase to 0.84 for the predictions with the XGBoost models and 0.82 for those with the RF models.

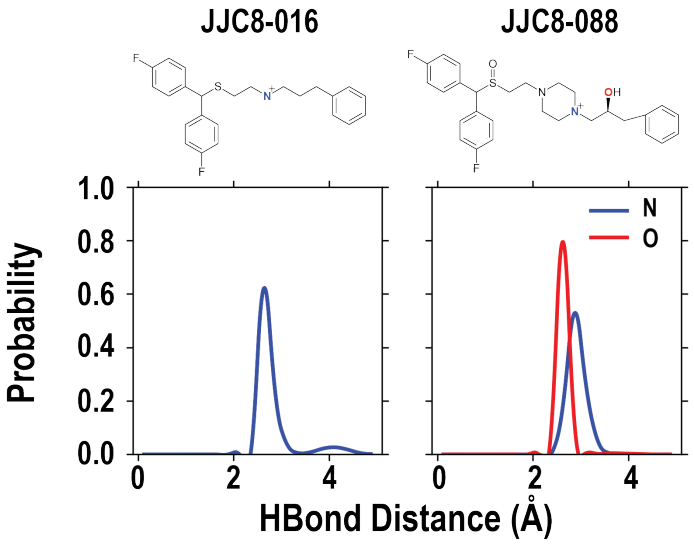

Figure S6. MD simulations of hDAT show that the nitrogen atom on JJC8-016 can form one H-bond with Asp79, while the protonated nitrogen and the hydroxy group of JJC8-088 can form two stable H-bonds with Asp79. The blue and red represent the H-bond interaction from the nitrogen and oxygen atoms on the ligand.

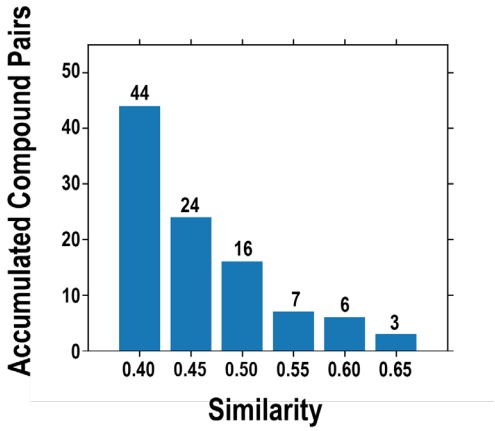

Figure S7. The compound pairs found in the ChEMBL datasets showing opposite affinity trends at hERG and DAT. We use the criteria of one compound is >90 fold better in DAT, and the other compound is >2 fold better in hERG. The accumulation of the numbers of compound pairs with Tanimoto similarity cutoff is reported in the bar plot.

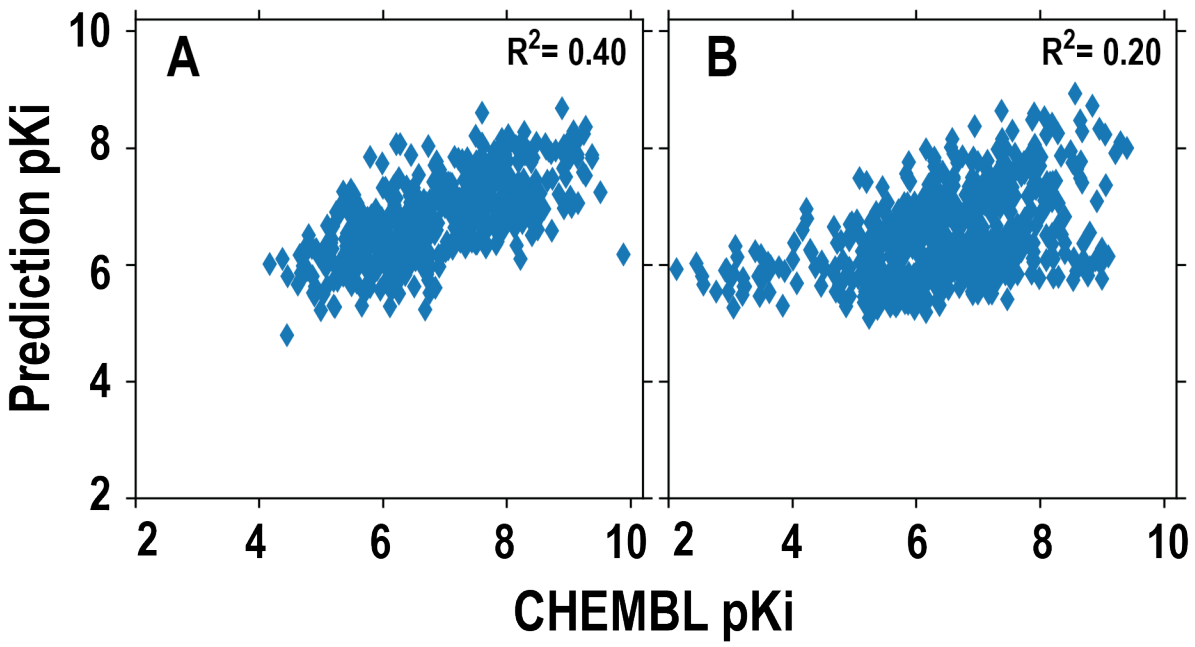

Figure S8. Cross predictions between hDAT and rDAT binding XGBoost models. (A) Predictions of the hDAT binding models on the rDAT binding dataset. (B) Predictions of the rDAT binding models on the hDAT binding dataset.

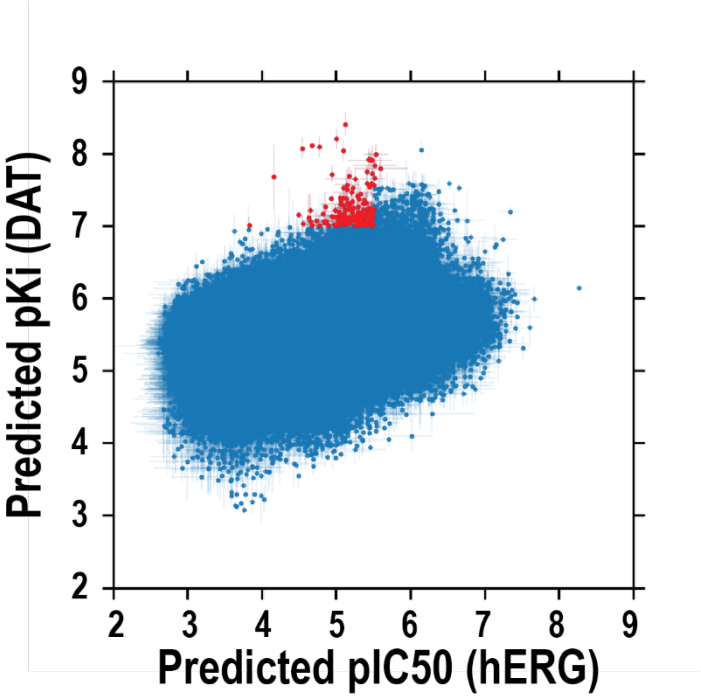

Figure S9. Counter screening of the NCI open database compounds. The hERG clamp models and all-DAT binding models are used for the prediction, with the training data including both the ChEMBL and the validation datasets.

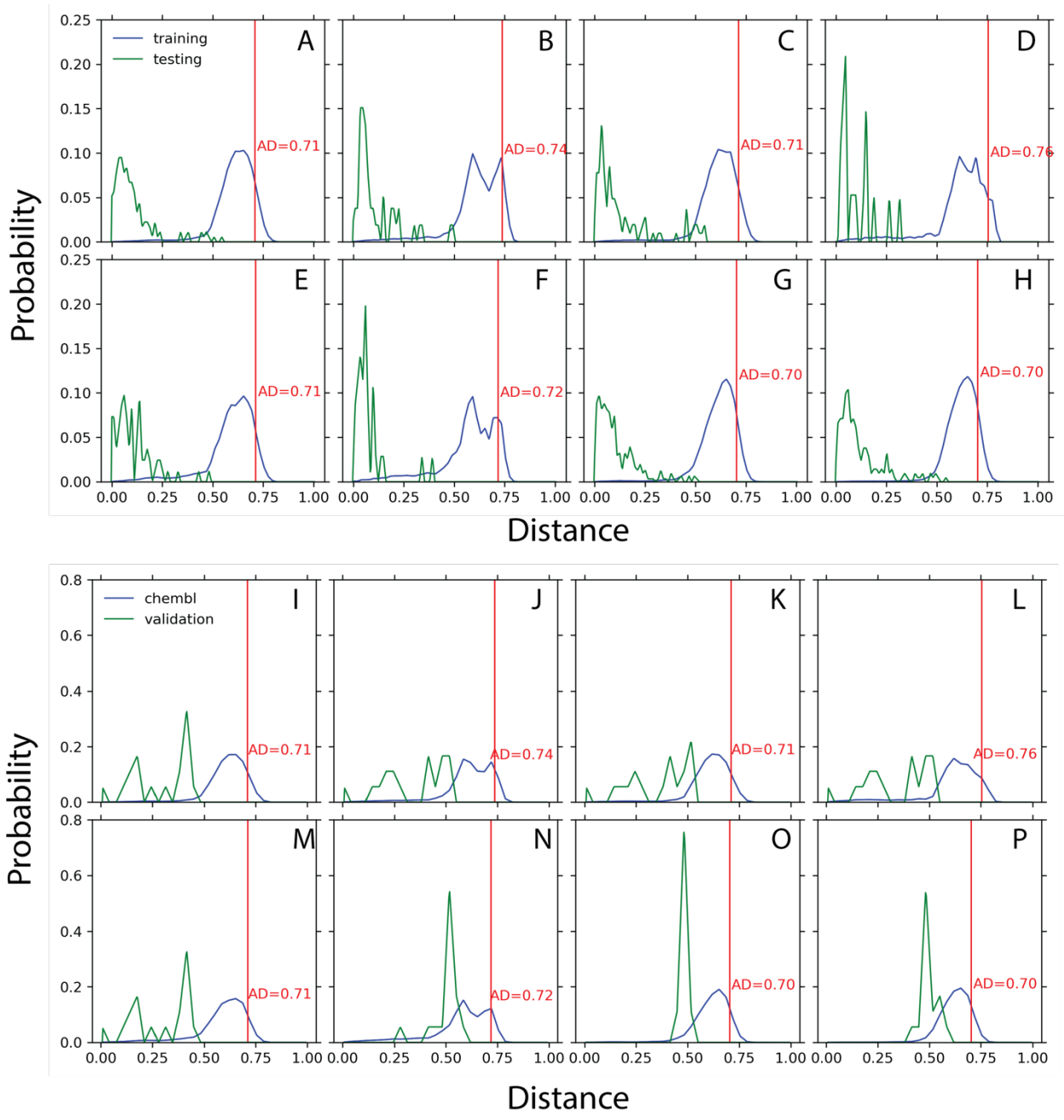

Figure S10. The testing and validation datasets are covered by the applicability domains of the QSAR models built from different datasets. For the models used in the benchmarking, examples of using one random splitting show that the testing datasets are covered by the applicability domains of corresponding training datasets for the (A) all-DAT binding, (B) all-DAT uptake, (C) hDAT binding, (D) hDAT uptake, (E) rDAT binding, (F) rDAT uptake, (G) hERG binding and (H) hERG clamp. The validation dataset is also covered by the applicability domains of the datasets used to build the final models: (I) all-DAT binding, (J) all-DAT uptake , (K) hDAT binding, (L) hDAT uptake, (M) rDAT binding, (N) rDAT uptake, (O) hERG binding and (P) hERG clamp.

Table S1. The filters and the numbers of datapoints after applying each filter.

	hERG		all-DAT		hDAT		rDAT	
Starting data	20695		13273		7138		5909	
After confidence score filter	20056		8392		5832		2442	
After assay type filter	19330		8369		5809		2442	
	IC50	Ki	IC50	Ki	IC50	Ki	IC50	Ki
After Ki / IC50 filter	10454	2466	3432	3418	2506	2316	887	1087
After standard units filter	8957	1546	2671	2659	1745	1557	887	1087
After activity relationship type fixes	6685	1206	2253	2368	1381	1323	833	1030

	hERG				all-DAT				hDAT				rDAT			
	binding		clamp		binding		uptake		binding		uptake		binding		uptake	
	IC50	Ki	IC50	Ki	IC50	Ki	IC50	Ki	IC50	Ki	IC50	Ki	IC50	Ki	IC50	Ki
After assay description filter	2456	730	2021	52	845	1616	822	484	383	935	597	156	442	666	218	328
Reserving hERG calibration compounds	n/a	n/a	1968	49	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
After data set size filter	2337	678	1542	44	807	1561	784	481	353	895	564	153	434	649	204	328
Desalting pass	2337	678	1542	44	805	1559	784	480	351	893	564	152	434	649	204	328
After oddball element filter	2337	678	1542	44	805	1559	784	480	351	893	564	152	434	649	204	328
After molecular weight filter	2317	667	1519	44	793	1540	767	477	339	874	560	149	434	649	191	328
After pChEMBL value filter	2317	667	1519	44	793	1540	767	477	339	874	560	149	434	649	191	328
After edge case filter	2317	667	1519	44	793	1540	767	477	339	874	560	149	434	649	191	328
After deduplication pass	2043	634	1405	44	538	1189	554	350	279	684	414	126	260	541	140	229
Excluding 5-6 pKi/pIC50	1137	334	783	42	434	887	383	219	213	503	277	45	222	424	110	177
Binders	549	251	284	41	417	798	294	200	199	438	234	38	219	401	64	165
Nonbinders	588	83	499	1	17	89	89	19	14	65	43	7	3	23	46	12

Table S2. Keywords used in assay description filter to divide the data into hERG binding, hERG clamp, DAT binding, and DAT uptake datasets.

Dataset	Description keyword
hERG binding	[3H] Astemizole [3H]astemizole [3H]-astemizole [3H] astemizole [3H]Astemizole radiolabeled astemizole [3H]bufuralol [3H]Dofetilide [3H] dofetilide [3H]dofetilide 3H-dofetilide [3H]-dofetilide [3H]-Dofetilide radiolabeled dofetilide Displacement of dofetilide Displacement of labeled dofetilide Inhibition of dofetilide binding [3H]dofetidile Displacement of doferilide Cy3b-Dofetilide-based [35S]MK-499 [35S]MK499 [35S]-MK-499 35[S] MK-499 MK499 Displacement of MK-499 radiolabeled MK-499 radio-labeled MK-499 MK-0499 [35S]N-[(4R)-1'-[(2R)-6-cyano-1,2,3,4-tetrahydro-2-naphthalenyl]-3,4-dihydro-4-hydroxyspiro[2H-1-benzopyran-2,4'-piperidin]-6-yl]methanesulfonamide [35S]N-[(4R)-10-[(2R)-6-cyano-1,2,3,4-tetrahydro-2-naphthyl]-3,4-dihydro-4-hydroxyspiro[2H-1-benzopyran-2,40-piperidin]-6-yl]methanesulfonamide [35S]N-[(4R)-1'-[(2R)-6-cyano-1,2,3,4-tetrahydro-2-naphthyl]-3,4-dihydro-4-hydroxyspiro[2H-1-benzopyran-2,4'-piperidin]6-yl]methanesulfonamide 3,7-Bis[2-(4-nitro[3,5-3H]phenyl)ethyl]-3,7-diazabicyclo[3.3.1]nonane 3,7-Bis[2-(4-nitro[3,5]-[3H]phenyl)ethyl]-3,7-diazabicyclo[3.3.1]nonane radioligand displacement radioligand binding assay radioligand-binding competition Inhibition of binding to hERG Displacement of dofetidine Inhibition of Cy3B-labeled ligand binding Displacement of Tracer Red

Dataset	Description keyword
hERG clamp	manual electrophysiology electrophysiology assay electrophysiological assay electrophysiology study whole-cell plate-based electrophysiology patch plate method clamp PatchXpress Q-patch Qpatch patch express assay ion works assay IONWORKS IonWorks ionworks HT assay
DAT binding	BCTP [3H]BTCP radiolabeled BTCP CIT mazindol Mazindol Vanoxerine [125I]PE2I IPT [125I]N-(3'-iodopropen-2'yl)-2-beta-carbomethoxy-3-beta-(4-chlorophenyl)tropane CFT WIN- WIN-35 WIN35428 WIN5428 WIN 35428 WIN 35,428 WIN-35428 WIN-35,428 WIN35,428 [125I]methyl 3-(4-iodophenyl)-8-methyl-8-aza-bicyclo[3.2.1]octane-2-carboxylate GBR GBR12935 GBR-12935 GBR-12,935 RT155 RTI55 RTI -55 RTI-55 RTI-121
DAT uptake	[3H]dopamine reuptake [3H]-dopamine [3H]dopamine [3H]DA [3H]-DA Inhibition of dopamine (DA) uptake

Table S3. Benchmarks of the models trained with the in-house DAT binding dataset.

Models	Metrics	XGBoost			Random Forest		
		Ave.	St.Dev.	Best	Ave.	St.Dev.	Best
Regression	R^2	0.48	0.13	0.71	0.46	0.15	0.67
	RMSE	0.61	0.09	--	0.63	0.10	--
Classification	Accuracy	0.97	0.02	--	0.98	0.02	--
	Sensitivity	0.99	0.01	--	1.00	0.01	--
	Specificity	0.00	0.00	--	0.20	0.41	--
	F Score	0.98	0.01	--	0.99	0.01	--

Ave., averages of 35 models for each dataset for the regression modeling, or 25 models for each dataset for the classification modeling (see Methods and Figure S2); S.D., standard deviation.

Table S4. Benchmarks of the XGBoost classification models trained with equal numbers of binders and non-binders from the all-DAT binding dataset. We randomly reduced the number of binders to match number of nonbinders, and the randomization was performed 9 times to prepare 9 different training datasets. For each dataset, models were built using 25 different random splittings. The averages and standard deviations of the benchmarks were then calculated for the resulting 225 models. Compared to Table 2, the accuracy is not as good as using the entire dataset, but the sensitivity and specificity are improved.

	Ave.	St. Dev.
Accuracy	0.87	0.07
Sensitivity	0.88	0.09
Specificity	0.87	0.09
F Score	0.86	0.07

Table S5. Most correlated descriptors for DAT and hERG ligands.

Dataset	10 most positively correlated features		10 most negatively correlated features	
	Descriptor	R	Descriptor	R
DAT binding	NumAliphaticHeterocycles	0.50	BalabanJ	-0.41
	NumSaturatedHeterocycles	0.49	SlogP_VSA11	-0.27
	Chi3n	0.44	SlogP_VSA1	-0.23
	Chi4n	0.44	qed	-0.21
	Chi3v	0.43	fr_allylic_oxid	-0.21
	Chi4v	0.42	SMR_VSA9	-0.21
	RingCount	0.41	TPSA	-0.18
	NumSaturatedRings	0.40	NHOHCount	-0.18
	Chi2n	0.39	fr_NH2	-0.17
	Chi2v	0.39	PEOE_VSA1	-0.17
hERG clamp	fr_unbrch_alkane	0.34	fr_COO	-0.34
	MolLogP	0.30	fr_COO2	-0.34
	EState_VSA5	0.21	TPSA	-0.33
	MinAbsEStateIndex	0.20	fr_Ar_COO	-0.33
	VSA_EState5	0.19	VSA_EState2	-0.28
	PEOE_VSA7	0.18	NumHDonors	-0.26
	EState_VSA8	0.18	NOCount	-0.25
	fr_sulfide	0.18	fr_C_O	-0.25
	NumRotatableBonds	0.16	NHOHCount	-0.23
	PEOE_VSA6	0.16	PEOE_VSA2	-0.23
DAT validation	RingCount	0.86	BalabanJ	-0.80
	Kappa2	0.84	qed	-0.79
	NumRotatableBonds	0.83	fr_priamide	-0.61
	Chi1	0.83	NHOHCount	-0.61
	MolMR	0.82	PEOE_VSA12	-0.61
	Chi1n	0.82	fr_NH2	-0.61
	LabuteASA	0.82	SMR_VSA4	-0.61
	Kappa3	0.81	fr_amide	-0.49
	HeavyAtomCount	0.81	fr_C_O_noCOO	-0.49
	Chi3n	0.80	fr_C_O	-0.49
hERG validation	NumRotatableBonds	0.87	qed	-0.88
	MolLogP	0.81	BalabanJ	-0.81
	Kappa2	0.79	SMR_VSA4	-0.70

Kappa3	0.76	fr_priamide	-0.70
RingCount	0.76	NHOHCount	-0.70
Chi1	0.74	fr_NH2	-0.70
NumAromaticCarbocycles	0.73	PEOE_VSA12	-0.70
fr_benzene	0.73	TPSA	-0.69
NumAromaticRings	0.73	FpDensityMorgan1	-0.61
Chi1n	0.72	VSA_EState2	-0.58

Table S6. Summary of MD simulations.

Protein	Ligand	Number of runs	Simulation length
DAT	JJC8016	7	15.5 μs
	JJC8088	5	11.1 μs
hERG	JJC8016	6	4.86 μs
	JJC8088	3	3.6 μs