

## Supporting Information

### **Machine learning classification can reduce false positives in structure-based virtual screening**

Yusuf O. Adeshina<sup>1,2</sup>, Eric J. Deeds<sup>2,3</sup>, and John Karanicolas<sup>1\*</sup>

<sup>1</sup>Program in Molecular Therapeutics, Fox Chase Cancer Center, Philadelphia, PA 19111

<sup>2</sup>Center for Computational Biology, <sup>3</sup>Department of Molecular Biosciences,  
University of Kansas, Lawrence, KS 66045

\*To whom correspondence should be addressed.

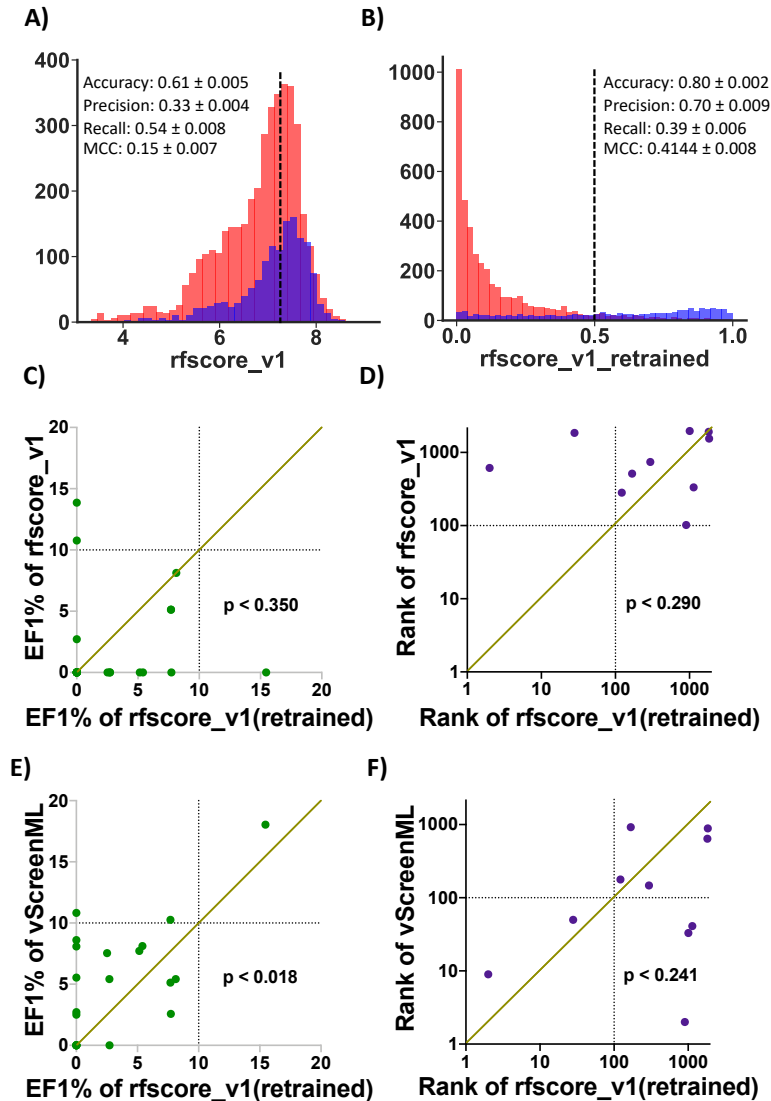
E-mail: john.karanicolas@fccc.edu

## Supplemental Figures

<b>Rosetta energies (6)</b>		<b>ChemAxon (4)</b>		<b>BINANA (13)</b>	
fa_atr		fsp3		$\alpha$ -helix side chain flexibility	
fa_rep		Polar surface area		$\beta$ -strand side chain flexibility	
fa_sol		Van der Waals surface area		Other side chain flexibility	
fa_elec		pienergy		$\alpha$ -helix back-bone flexibility	
hbond_bb_sc				$\beta$ -strand back-bone flexibility	
hbond_sc				other back-bone flexibility	
				Electrostatics	
				Number of hydrogen bonds	
				Hydrophobic contacts	
				Pi-pi interaction	
				T-stacking	
				Pi-cation	
				Salt-bridge	
<b>Rosetta Struct. Quantities (8)</b>		<b>SZYBKI (1)</b>		<b>RF-Score (36)</b>	
interface_Energy		Ligand conformational entropy change upon binding		Multiple distance-dependent atom counts	
total_BSA					
interface_HB					
total_packstats					
interface_unsat					
total_pose_exposed_SASA					
interface_hydrophobic_sasa					
interface_polar_sasa					

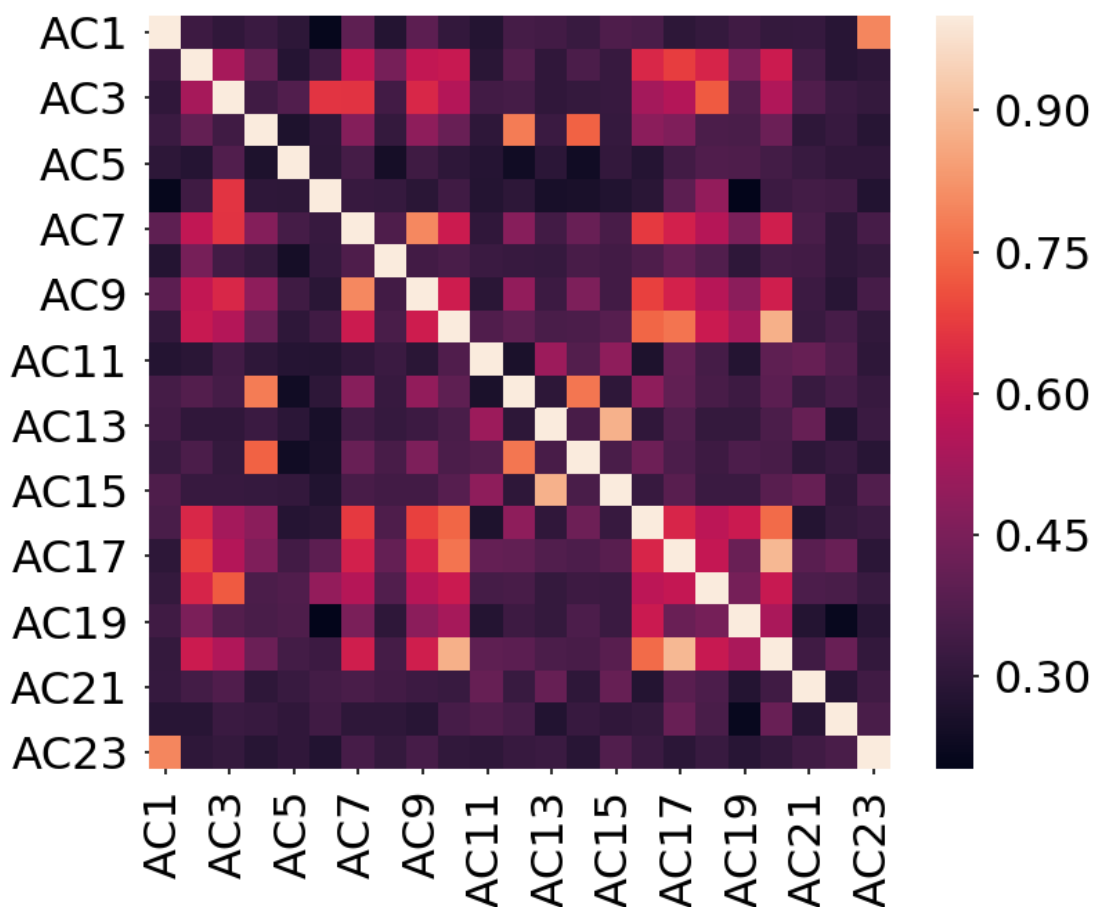
**68 TOTAL  
FEATURES**

**Figure S1: Features incorporated into vScreenML.** These features derive from six sources: Rosetta energy terms, Rosetta structural quantifiers, RF-Score's rfscore\_v1 features, BINANA's analysis of intermolecular contacts, ChemAxon's cxcalc features, OpenEye's SZYBKI conformational entropy term.

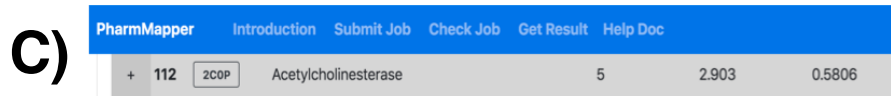
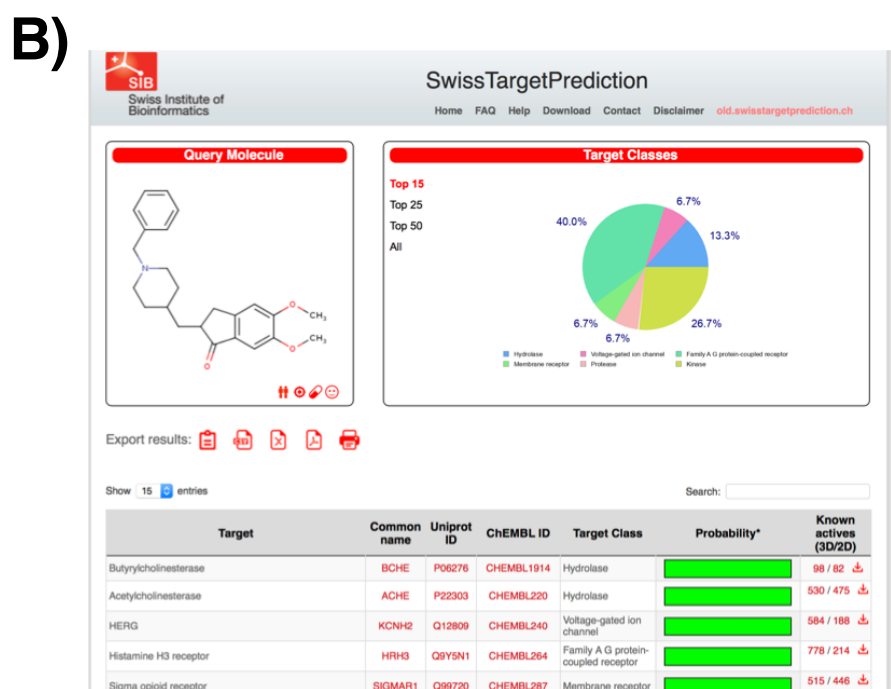
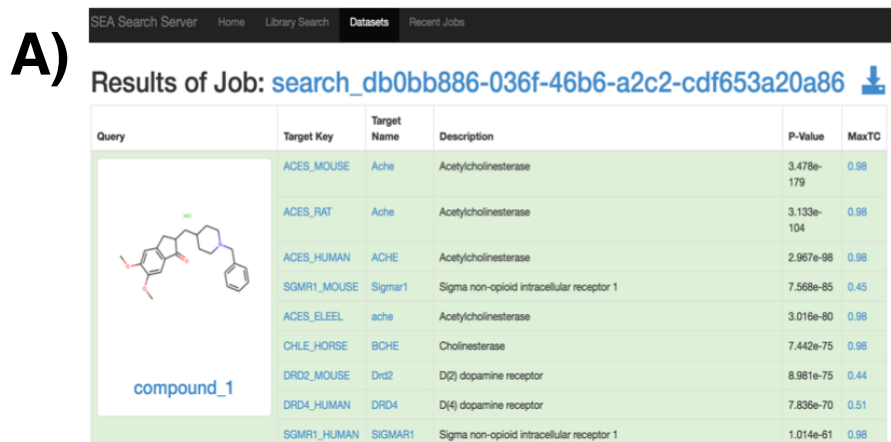


**Figure S2: Retraining rfscore\_v1 using D-COID.** (A) Overlaid histograms for scores obtained when scoring active complexes (*blue*) and decoy complexes (*red*) from D-COID using the original rfscore\_v1. Performance measures are presented as the average of 10 experiments, each of which uses different partitions for 10-fold cross-validation. Uncertainty is presented as 95% confidence intervals.

(B) Overlaid histograms after re-training rfscore\_v1. (C) Comparison of the original and re-weighted versions of rfscore\_v1 applied to the DEKOIS benchmark. (D) Comparison of the original and re-weighted versions of rfscore\_v1 applied to the PPI benchmark. (E) Comparison of re-weighted rfscore\_v1 versus vScreenML on the DEKOIS benchmark. (F) Comparison of re-weighted rfscore\_v1 versus vScreenML on the PPI benchmark. p-values were computed using the two-tailed Wilcoxon Signed-Rank test.



**Figure S3: Heatmap showing similarity between the 23 compounds tested as candidate AChE inhibitors.** All compounds are included in this heatmap, labels include only odd numbers for clarity. Similarity is measured via 2D fingerprints. As highlighted by this analysis, several subsets of these compounds are similar to one another: AC1/AC23 group together, AC3/AC18 group together, AC4/AC12/AC14 group together, AC7/AC9/AC16 group together, AC10/AC16/AC17/AC20 group together, and AC13/AC15 group together. On the other hand, some of these compounds are completely unrelated to any others in this set (e.g., AC5, AC21, AC22). The similarity shown here is also confirmed by visual inspection of the compounds themselves (Table S6).



**Figure S4: Positive control for target identification methods.** We confirmed that all three methods would successfully identify AChE as the target of a known AChE inhibitor (donepezil, ChEMBL1678). **(A)** Similarity Ensemble Approach (SEA). **(B)** SwissTargetPrediction. **(C)** PharmMapper. We note that AChE was only ranked #112 among the PharmMapper hits because the 3D conformations it built for donepezil were not sufficiently well-matched to the active conformation to produce a better ranking.

## Supplemental Tables

Performance measure	Optimized	Non-optimized
Accuracy	0.90 ± 0.0022	0.89 ± 0.0026
Precision	0.89 ± 0.0076	0.85 ± 0.0079
Recall	0.71 ± 0.0070	0.66 ± 0.0081
AUC	0.84 ± 0.0034	0.81 ± 0.0042
F1-Score	0.79 ± 0.0049	0.74 ± 0.0064
MCC	0.74 ± 0.0062	0.68 ± 0.0078

**Table S1: Effect of hyperparameter tuning on vScreenML.** The performance is shown for the optimized and non-optimized vScreenML models (both use XGBoost, with the complete feature set. Performance measures are presented as the average of 100 trained models, each of which derived from 10-fold cross-validation (see *Methods*). Uncertainty is presented as 95% confidence intervals. In all cases, performance measures were calculated for a subset of the data that was held out from the training step.

Parameter	Pre-optimization (default value)	Optimized value
Learning rate	0.3	0.01
Min_child_weight	1	1
Max_depth	6	7
Gamma	0	0.1
Subsample	1	0.5
Colsample_bytree	1	0.4
Lambda	1	Default
Alpha	0	Default
Scale_pos_weight	1	1
N_estimators	100	1945

**Table S2: Results of hyperparameter tuning.** The parameters resulting from optimization for vScreenML are shown.

Learning scheme	Accuracy	Precision	Recall	AUC	MCC
XGBoost	0.89 ± 0.0019	0.85 ± 0.0059	0.66 ± 0.0052	0.81 ± 0.0027	0.68 ± 0.0054
Gradient Boosting	0.89 ± 0.0024	0.85 ± 0.0073	0.66 ± 0.0062	0.81 ± 0.0034	0.68 ± 0.0069
Random Forest	0.86 ± 0.0022	0.84 ± 0.0069	0.56 ± 0.0071	0.76 ± 0.0037	0.61 ± 0.0070
Extra Trees	0.86 ± 0.0021	0.84 ± 0.0061	0.55 ± 0.0074	0.76 ± 0.0037	0.61 ± 0.0068
Linear Discriminant Analysis	0.86 ± 0.0030	0.78 ± 0.0089	0.62 ± 0.0077	0.78 ± 0.0042	0.61 ± 0.0087
Quadratic Discriminant Analysis	0.35 ± 0.0042	0.28 ± 0.0014	0.98 ± 0.0029	0.56 ± 0.0030	0.17 ± 0.0064
Gaussian Naïve Bayes	0.50 ± 0.0097	0.32 ± 0.0038	0.90 ± 0.0053	0.63 ± 0.0061	0.25 ± 0.0097
K-nearest Neighbor (kNN)	0.74 ± 0.0018	0.45 ± 0.0065	0.24 ± 0.0053	0.57 ± 0.0025	0.18 ± 0.0060
DUMB	0.75	0.00	0.00	-	0.00

**Table S3: Performance from alternate learning schemes.** Using the complete vScreenML feature set, alternate learning schemes were evaluated. Performance measures are presented as the average of 100 trained models, each of which derived from 10-fold cross-validation (see *Methods*). Uncertainty is presented as 95% confidence intervals. In all cases, performance measures were calculated for a subset of the data that was held out from the training step.

Features	Accuracy	Precision	Recall	AUC	MCC
Rosetta (reweighted)	0.85 ± 0.0019	0.76 ± 0.0060	0.58 ± 0.0093	0.76 ± 0.0040	0.57 ± 0.0061
RF (reweighted)	0.78 ± 0.0021	0.65 ± 0.0103	0.31 ± 0.0065	0.63 ± 0.0032	0.34 ± 0.0076
BINANA (reweighted)	0.83 ± 0.0010	0.74 ± 0.0064	0.50 ± 0.0046	0.72 ± 0.0015	0.51 ± 0.0027
LigPro+Szybki+RF+BINANA	0.86 ± 0.0011	0.80 ± 0.0033	0.56 ± 0.0063	0.76 ± 0.0027	0.59 ± 0.0035
Rosetta+NC	0.86 ± 0.0025	0.80 ± 0.0066	0.61 ± 0.0109	0.78 ± 0.0050	0.61 ± 0.0078
Rosetta+NC+LigPro	0.86 ± 0.0021	0.80 ± 0.0067	0.61 ± 0.0090	0.78 ± 0.0040	0.61 ± 0.0064
Rosetta+NC+LigPro+Szybki	0.86 ± 0.0021	0.79 ± 0.0068	0.61 ± 0.0086	0.78 ± 0.0039	0.61 ± 0.0064
Rosetta+NC+LigPro+Szybki+RF	0.88 ± 0.0027	0.83 ± 0.0081	0.66 ± 0.0087	0.81 ± 0.0044	0.67 ± 0.0081
Rosetta+NC+LigPro+Szybki+BINANA	0.88 ± 0.0030	0.85 ± 0.0064	0.62 ± 0.0123	0.79 ± 0.0059	0.65 ± 0.0092
Rosetta+NC+LigPro+Szybki+RF+BINANA (non-optimized vScreenML)	0.89 ± 0.0026	0.85 ± 0.0079	0.66 ± 0.0081	0.81 ± 0.0042	0.68 ± 0.0078

**Table S4: Performance with restricted feature sets.** Examination of models in which all features from a given origin are added/removed en masse; all models are trained using XGBoost. Performance measures are presented as the average of 100 trained models, each of which derived from 10-fold cross-validation (see *Methods*). Uncertainty is presented as 95% confidence intervals. In all cases, performance measures were calculated for a subset of the data that was held out from the training step.

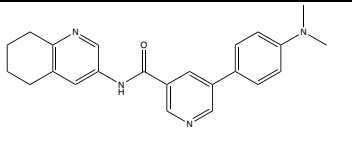
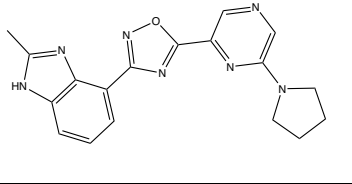
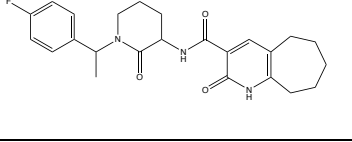
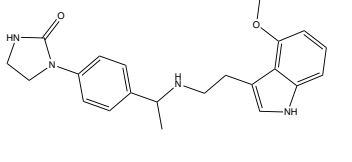


DEKOIS testcase: PDB (protein name)	Closest protein in D-COID	Sequence identity (%)	EF1% from vScreenML
3hng (VEGFR1)	4ag8 (VEGFR2)	76.7	10.3
1hov (MMP-2)	1xuc (MMP-13)	66.5	2.5
3ny9 ( $\beta$ 2 adrenergic receptor)	2y00 ( $\beta$ 1 adrenergic receptor)	65.3	0
3kk6 (COX-1)	3ln0 (COX-2)	64.6	10.8
1nhz (glucocorticoid receptor)	2zk5 (PPARG)	62.5	5.4
1xp0 (PDE5A)	2jc6 (CAMK1D)	52.4	8.1
1z11 (CYP2A6)	5w0c (CYP2C9)	50.1	0
3tfq (corticosteroid 11 $\beta$ -dehydrogenase 1)	2bpm (CDK2 kinase)	50.0	8.6
2oo8 (angiopoietin-1 receptor)	5am6 (FGFR1)	47.1	5.1
1b8o (purine nucleoside phosphorylase)	3o4v (MTH/SAM nucleosidase)	43.5	7.5
2w31 (Bcl-2)	5myg (peregrin bromodomain)	42.9	5.5
1hw8 (HMG-CoA reductase)	1jla (HIV-1 RT)	42.4	24.6
2afx (human glutaminyl cyclase)	4f9v (fly glutaminyl cyclase)	42.1	0
3ewj (TNF-alpha convertase)	1xuc (MMP-13)	39.5	2.7
3v8s (ROCK1 kinase)	4nus (Rsk2 kinase)	39.2	18.0
3eml (adenosine receptor A2a)	5a8e ( $\beta$ 1 adrenergic receptor)	39.2	7.7
2z94 (SARS-CoV protease)	5ccs (cyclophilin D)	37.9	0
1uze (angiotensin-converting enzyme ACE)	4ag8 (VEGFR2)	34.3	21.4
3klm (estradiol 17 $\beta$ -dehydrogenase 1)	4bo0 (ACP reductase)	29.6	5.4
2wcg (glucosylceramidase)	5c8z (zearalenone hydrolase)	27.9	2.6
1w4r (thymidine kinase)	3zv9 (enterovirus 3C protease)	25.4	0
1r4l (angiotensin-converting enzyme ACE2)	4wnp (ULK1 kinase)	23.3	8.1
1uou (thymidine phosphorylase)	3wmc ( $\beta$ -GlcNAcase)	21.7	0

**Table S5: Similarity of proteins in DEKOIS benchmark to training set.** The complete DEKOIS set comprises 81 proteins. Some were present in our D-COID training set as well, and were therefore excluded when we carried out the benchmark experiment. For each of 23 protein targets included in our benchmark, we present the closest protein present in our D-COID training set (on the basis of sequence identity). The EF1% values listed here correspond to those presented in **Figure 3a** (zeros correspond to cases in which none of the active compounds were ranked in the top 1%). The fact that vScreenML does not exhibit superior performance when a closer protein homolog is available suggests that its performance does not rely on identifying a related protein homolog in the training set.

Compound	Structure	SMILES string
AC1		<chem>COC=1C=CC=C2NC=C(C(C)C)C=3C=CC(=CC3)N4C=NC=N4)C12</chem>
AC2		<chem>CN1N=C(C=C1NC(=O)C=2C=NC=C(C2)C=3C=CC(Cl)=CC3)C(C)(C)C</chem>
AC3		<chem>CC=1C=CC=C(C1)C=2C=NC=C(C2)C(=O)NC3=NC(=CS3)C4CCNCC4</chem>
AC4		<chem>FC=1C=CC(NC=2C=CC(NC(=O)C=3C=NC(Cl)=CN3)=C4C=NC=CC24)=CC1</chem>
AC5		<chem>CC1=NC=2C(=CC=CC2N1)C3=NOC(=N3)C=4C=NC=C(N4)N5CCCC5</chem>
AC6		<chem>ClC1=C(NC=2C=CC(Cl)=CC12)C(=O)NC3=NC(=CS3)C4CCNCC4</chem>
AC7		<chem>CC=1C=CC=C(C1)C=2C=NC=C(C2)C(=O)NC=3C=CC=4N=CNC4C3</chem>
AC8		<chem>CC1=CC(=NN1C=2C=C(C)C=C(C)C2)C(=O)N3C[C@H]([C@@H](C3)C=4C=CC(Cl)=CC4)C(=O)O</chem>
AC9		<chem>CC=1C=CC=C(C1)C=2C=NC=C(C2)C(=O)NC=3C=CC(=CC3)C4=CNC=N4</chem>
AC10		<chem>CN(C)C=1C=CC(=CC1)C=2C=NC=C(C2)C(=O)NC=3C=NC=4CCC(N)CC4C3</chem>

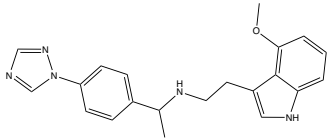
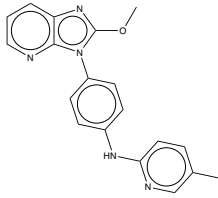
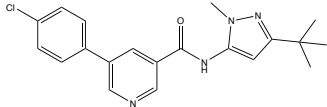
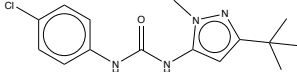
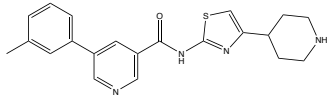
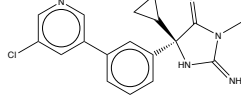
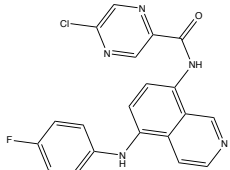
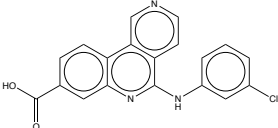
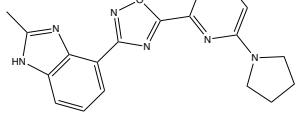
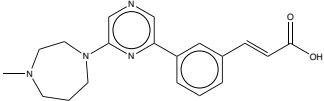
AC11		<chem>O=C(NCC1CCCC=2C=CC=NC12)N3CCC4=C(C3)N=NN4C=5C=CC=CC5</chem>
AC12		<chem>NC(=O)C1=CNC(=C1)C(=O)NC=2C=CC(NC=3C=CC(F)=CC3)=C4C=CN=CC24</chem>
AC13		<chem>O=C(CC1=NOC=2C=CC=CC12)N3CCC4=C(C3)N=CN4CCC=5C=CC=CC5</chem>
AC14		<chem>NC(=O)C1=CC(=NN1C=2C=CC=CC2)C(=O)NC=3C=CC(NC=4C=CC(F)=CC4)=C5C=CN=CC35</chem>
AC15		<chem>COC=1C=CC=2ON=C(CC(=O)N3CCC4=C(C3)N=CN4CCC=5C=CC=CC5)C2C1</chem>
AC16		<chem>CN(C)C=1C=CC(=CC1)C=2C=NC=C(C2)C(=O)NC=3C=CC(F)=C(C)C3</chem>
AC17		<chem>C1C=1C=CC(=CC1)C=2C=NC=C(C2)C(=O)NC=3C=NC=4CCCCC4C3</chem>
AC18		<chem>CC=1C=CC(=CC1)C=2C=NC=C(C2)C(=O)NC3=NC(=NN3C)C4CCNCC4</chem>
AC19		<chem>CN(C)C=1C=CC(=CC1)C=2C=NC=C(C2)C3=NN=C(O3)C4=C(C)N=C5C=CC=CN45</chem>

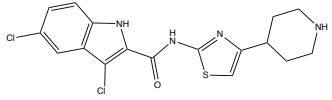
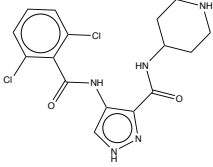
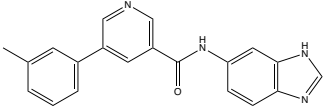
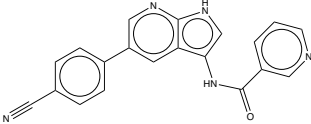
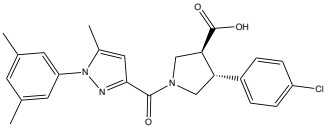
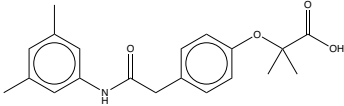
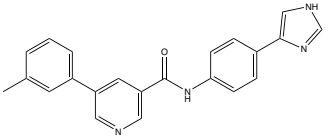
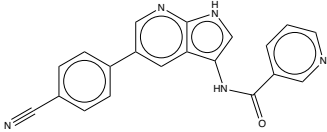
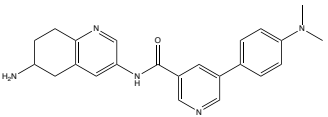
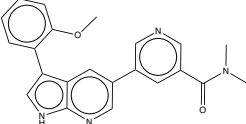
AC20		<chem>CN(C)C=1C=CC(=CC1)C=2C=NC=C(C2)C(=O)NC=3C=NC=4CCCCC4C3</chem>
AC21		<chem>CC1=NC=2C(=CC=CC2N1)C3=NOC(=N3)C=4C=NC=C(N4)N5CCCC5</chem>
AC22		<chem>CC(N1CCCC(NC(=O)C2=CC=3CCCCC3NC2=O)C1=O)C=4C=CC(F)=CC4</chem>
AC23		<chem>COC=1C=CC=C2NC=C(CCNC(C)C=3C=CC(=CC3)N4CCNC4=O)C12</chem>

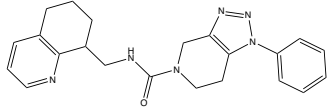
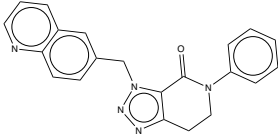
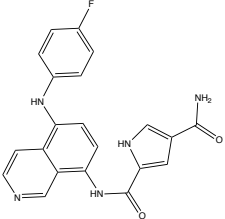
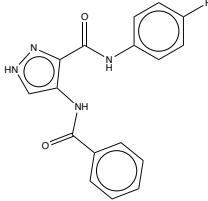
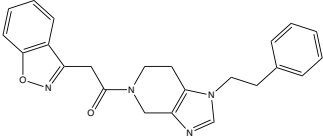
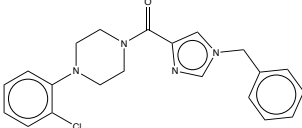
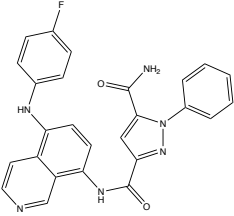
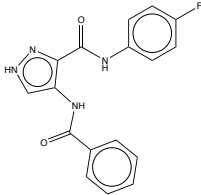
**Table S6: Compounds tested as candidate AChE inhibitors.** Chemical structures and SMILES strings are presented for each of the 23 compounds selected by vScreenML as a candidate AChE inhibitor.

Compound	rfscore_v1	rfscore_v2	rfscore_v3	nnscore	PLEClinear	PLECnn	PLECrf	rfscore_VS
AC1	18711	14707	16971	10184	484	3590	8779	12311
AC10	13709	8242	9480	17452	8671	2186	724	9336
AC11	8815	5095	6460	1371	10659	10432	5159	7553
AC14	3081	248	1871	540	2663	3188	5583	6253
AC17	15601	12686	15651	14166	10370	4883	4967	10158
AC19	11425	8951	8250	15378	5732	480	8448	9842
AC21	5699	12409	7174	6435	4992	3012	168	10310
AC22	1595	948	137	1401	12695	11360	6603	4391
AC23	15335	11442	15529	7018	1044	3520	7056	5809
AC6	4938	1049	4610	14151	16169	15061	3204	17445
AC8	6085	3317	1331	5377	6781	4983	3518	1877
AC9	18334	8948	16295	4917	5288	980	14428	14319
<b>Average rank:</b>	<b>10277</b>	<b>7337</b>	<b>8647</b>	<b>8199</b>	<b>7129</b>	<b>5306</b>	<b>5720</b>	<b>9134</b>

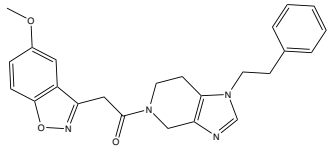
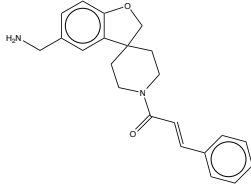
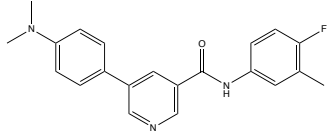
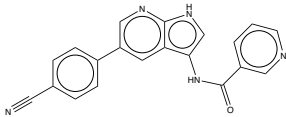
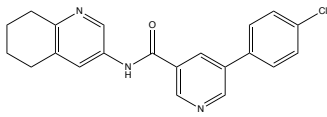
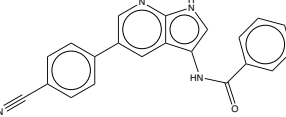
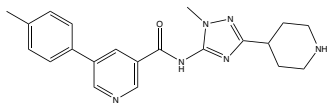
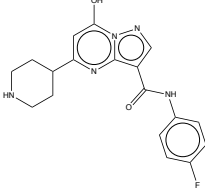
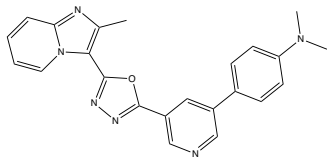
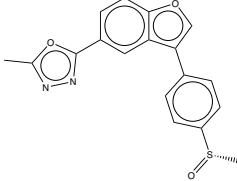
**Table S7: Ranking of first-round compounds selected by vScreenML by other methods.** Using each of the scoring functions methods evaluated in this study, we ranked the 20,000 docked models from the first round of AChE screening (corresponding data is not shown for the second round of screening, because at that point already vScreenML had been used to focus the search). Because we did not explicitly test the compounds selected by these other methods, we cannot say whether these other compounds are true AChE inhibitors that were missed by vScreenML. However, because none of these methods assign favorable (low) ranking to the same compounds selected by vScreenML (the compounds presented in this table), we can conclude that the particular compounds selected by vScreenML would not have been selected by these other methods.

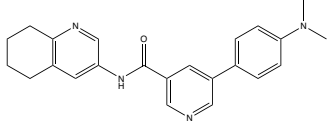
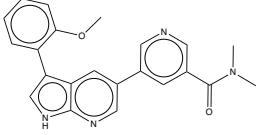
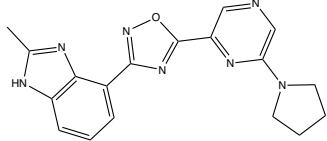
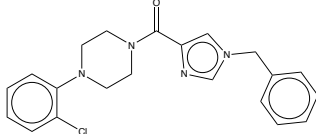
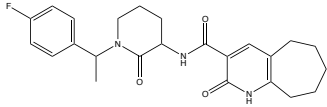
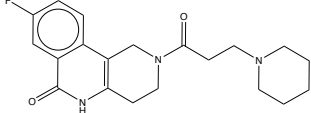
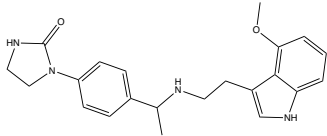
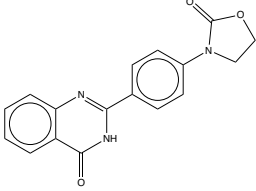
Compound	Structure	Closest analog in D-COID
AC1		 <p data-bbox="1011 499 1149 531">4p1r (2KR)</p> <p data-bbox="943 569 1219 600">Phosphodiesterase 10A</p>
AC2		 <p data-bbox="1008 751 1157 783">1kv1 (BMU)</p> <p data-bbox="984 821 1182 852">p38 MAP kinase</p>
AC3		 <p data-bbox="1011 1031 1149 1062">4djx (0KQ)</p> <p data-bbox="984 1100 1177 1131">Beta-secretase 1</p>
AC4		 <p data-bbox="1008 1339 1154 1371">3pe1 (3NG)</p> <p data-bbox="1008 1409 1154 1440">CK2 kinase</p>
AC5		 <p data-bbox="1016 1619 1146 1650">2xj2 (985)</p> <p data-bbox="1003 1688 1159 1719">Pim-1 kinase</p>

AC6		 <p>2vu3 (LZE) CDK2 kinase</p>
AC7		 <p>514q (LKB) AAK1 kinase</p>
AC8		 <p>1g9v (RQ3) Hemoglobin</p>
AC9		 <p>514q (LKB) AAK1 kinase</p>
AC10		 <p>2qoh (P3Y) Abl kinase</p>

<p>AC11</p>		 <p>4deh (0JK) c-Met kinase</p>
<p>AC12</p>		 <p>2vto (LZ8) CDK2 kinase</p>
<p>AC13</p>		 <p>4psq (2WL) Retinol-binding protein 4</p>
<p>AC14</p>		 <p>2vto (LZ8) CDK2 kinase</p>



AC15		 <p>2zec (11N) Tryptase beta 2</p>
AC16		 <p>514q (LKB) AAK1 kinase</p>
AC17		 <p>514q (LKB) AAK1 kinase</p>
AC18		 <p>4k0y (10A) Pim-1 kinase</p>
AC19		 <p>3gb2 (G3B) GSK3beta kinase</p>

AC20		 <p>2qoh (P3Y)</p> <p>Abl kinase</p>
AC21		 <p>4psq (2WL)</p> <p>Retinol-binding protein 4</p>
AC22		 <p>3ki6 (G9L)</p> <p>Cholix toxin</p>
AC23		 <p>4buw (F33)</p> <p>Tankyrase-2</p>

**Table S8: Similarity of compounds selected by vScreenML to D-COVID set.** For each compound selected by vScreenML as a candidate AChE inhibitor, we show the closest compound from D-COVID (used in training vScreenML). In each case, the PDB and ligand ID are shown, along with the protein to which this ligand was bound in the solved structure. None of these compounds come from complexes with proteins related to AChE, ruling out any concern that vScreenML might have recognized some subtle features of the binding sites from proteins related to AChE.

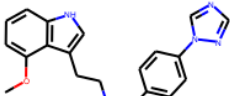
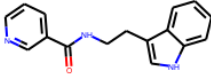
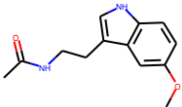
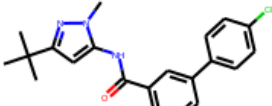
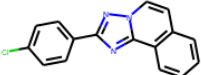
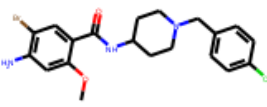
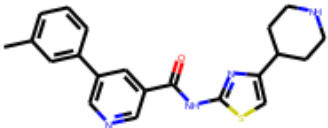
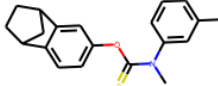
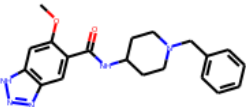
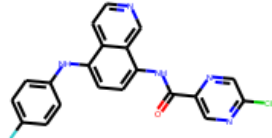
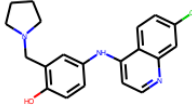
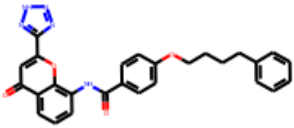
Compound	SEA predictions	SwissTarget predictions
AC1	Peptidyl-prolyl cis-trans isomerase B Kinesin-like protein KIF20A 5-hydroxytryptamine receptor 1B 5-hydroxytryptamine receptor 1D Apelin receptor	Serotonin receptor Tyrosine-protein kinase (JAK1 and JAK2) CaM kinase II Calcium sensing receptor Glutamate [NMDA] receptor
AC2	Sodium channel protein type 10 subunit alpha Sodium channel protein type 2 subunit alpha Stromal interaction molecule 1 Luciferin 4-monoxygenase Mitogen-activated protein kinase kinase kinase 5	Phosphodiesterase 5A Histone deacetylase 6 Serotonin 2a (5-HT2a) receptor P2X purinoceptor 7 Acetyl-CoA carboxylase 2
AC3	Thrombopoietin receptor GMP synthase [glutamine-hydrolyzing] Sodium channel protein type 10 subunit alpha Atypical chemokine receptor 3 Mas-related G-protein coupled receptor member X1	Melanin-concentrating hormone receptor 1 Rho-associated protein kinase 2 Cyclin-dependent kinase 4 Serine/threonine-protein kinase Aurora-A Monoamine oxidase B
AC4	Beta-secretase 2 Metabotropic glutamate receptor 4 Beta-secretase 1 Arachidonate 15-lipoxygenase Metabotropic glutamate receptor 4	Focal adhesion kinase 1 Tyrosine-protein kinase SRC Quinone reductase 2 Adenosine A1 receptor Adenosine A2a receptor
AC5	Serine/threonine-protein kinase pim-2 Serine/threonine-protein kinase pim-3 Calcium-activated potassium channel subunit alpha-1 Serine/threonine-protein kinase pim-1	DGAT1 protein Cholecystokinin B receptor Adenosine A1 receptor Adenosine A3 receptor Neurokinin 3 receptor
AC6	Glycogen phosphorylase, liver form Trace amine-associated receptor 1 NAD-dependent protein deacetylase HST2 Thrombopoietin receptor Glycogen phosphorylase, muscle form	Phosphodiesterase 5A Cyclin-dependent kinase 7 Protein kinase C theta Urotensin II receptor Melanin-concentrating hormone receptor 1
AC7	Potassium channel subfamily K member 9 NAD-dependent protein deacetylase sirtuin-3, mitochondrial NAD-dependent protein deacetylase sirtuin-2 Sodium channel protein type 10 subunit alpha Amine oxidase [flavin-containing] B	Melanin-concentrating hormone receptor 1 Sodium channel protein type X alpha subunit Microtubule-associated protein tau Alpha-synuclein Cyclin-dependent kinase 5/CDK5 activator 1

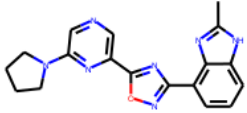
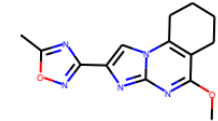
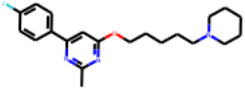
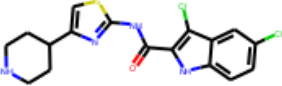
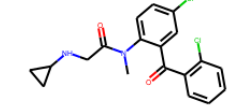
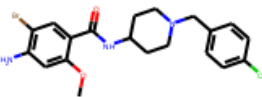
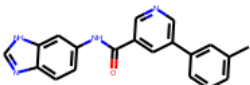
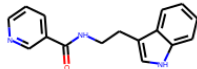
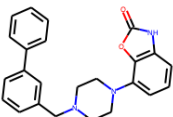
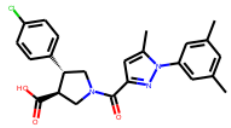
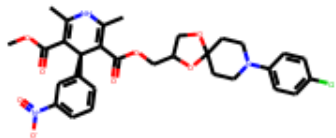
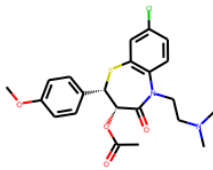
AC8	<p>Melanocortin receptor 4</p> <p>Transporter</p> <p>Replication protein A 70 kDa DNA-binding subunit</p> <p>Sodium-dependent dopamine transporter</p> <p>Sodium-dependent serotonin transporter</p>	<p>Lysosomal protective protein</p> <p>Thromboxane A2 receptor</p> <p>Prostanoid DP receptor</p> <p>Prostanoid EP4 receptor</p> <p>Prostanoid EP2 receptor</p>
AC9	<p>Sodium channel protein type 10 subunit alpha</p> <p>NAD-dependent protein deacetylase sirtuin-3, mitochondrial</p> <p>Potassium channel subfamily K member 9</p> <p>Tyrosine-protein kinase ABL1</p> <p>NAD-dependent protein deacetylase sirtuin-2</p>	<p>Sodium channel protein type X alpha subunit</p> <p>Histone deacetylase 1</p> <p>Melanin-concentrating hormone receptor 1</p> <p>Tyrosine-protein kinase SRC</p> <p>Anandamide amidohydrolase</p>
AC10	<p>Sodium channel protein type 10 subunit alpha</p> <p>Tyrosine-protein kinase ABL1</p> <p>Melanin-concentrating hormone receptor 1</p> <p>NAD-dependent protein deacetylase sirtuin-3, mitochondrial</p> <p>5-hydroxytryptamine receptor 1D</p>	<p>Tyrosine-protein kinase SYK</p> <p>Protein kinase C mu</p> <p>Serine/threonine-protein kinase D2</p> <p>Tyrosine-protein kinase ZAP-70</p> <p>Cyclin T1</p>
AC11	<p>C-X-C chemokine receptor type 4</p> <p>Envelope glycoprotein gp160</p> <p>P2X purinoceptor 7</p> <p>Nicotinamide phosphoribosyltransferase</p> <p>Serine/threonine-protein kinase TA03</p>	<p>Fatty acid synthase</p> <p>Vasopressin V2 receptor</p> <p>Oxytocin receptor</p> <p>Cathepsin K</p> <p>Phosphodiesterase 7A</p>
AC12	<p>NAD-dependent protein deacetylase sirtuin-3, mitochondrial</p> <p>NAD-dependent protein deacetylase sirtuin-2</p> <p>Platelet-derived growth factor receptor alpha</p> <p>Nucleosome-remodeling factor subunit BPTF</p> <p>Arachidonate 15-lipoxygenase</p>	<p>Cyclooxygenase-2</p> <p>15-hydroxyprostaglandin dehydrogenase [NAD+]</p> <p>Urokinase-type plasminogen activator</p> <p>Serine/threonine-protein kinase Chk2</p> <p>Heat shock protein HSP 90-alpha</p>
AC13	<p>CCR4-NOT transcription complex subunit 7</p> <p>Endothelial lipase</p> <p>5-hydroxytryptamine receptor 5A</p> <p>G-protein coupled receptor 183</p> <p>Enoyl-[acyl-carrier-protein] reductase [NADH] FabI</p>	<p>Melatonin receptor 1A</p> <p>Epoxide hydratase</p> <p>Proteinase activated receptor 4</p> <p>ATP-binding cassette sub-family G member 2</p> <p>MAP kinase p38 alpha</p>
AC14	<p>High affinity nerve growth factor receptor</p> <p>Metabotropic glutamate receptor 4</p> <p>Signal transducer and activator of transcription 6</p> <p>Ketohexokinase</p> <p>Metabotropic glutamate receptor 4</p>	<p>Neuropeptide Y receptor type 5</p> <p>Thrombin and coagulation factor X</p> <p>Tyrosine-protein kinase JAK3</p> <p>Dihydroorotate dehydrogenase</p> <p>Thrombin</p>

AC15	Beta-1,4-mannosyl-glycoprotein 4-beta-N-acetylglucosaminyltransferase Ryanodine receptor 2 Melatonin receptor type 1A Melatonin receptor type 1B G-protein coupled receptor 183	Melatonin receptor 1A Melatonin receptor 1B Cathepsin K Cathepsin S MAP kinase-activated protein kinase 2
AC16	Sodium channel protein type 10 subunit alpha Tyrosine-protein kinase ABL1 NAD-dependent protein deacetylase sirtuin-3, mitochondrial Sodium channel protein type 10 subunit alpha 5-hydroxytryptamine receptor 1D	Phospholipase A2 group IIA Dual specificity phosphatase Cdc25B Phospholipase D1 Adenosine A1 receptor Acyl coenzyme A:cholesterol acyltransferase 1
AC17	Sodium channel protein type 10 subunit alpha Sodium channel protein type 2 subunit alpha Ubiquitin carboxyl-terminal hydrolase BAP1 Luciferin 4-monooxygenase Melanin-concentrating hormone receptor 1	Potassium channel subfamily K member 3 Cannabinoid receptor 1 Fibroblast growth factor receptor 1 Dopamine D4 receptor Epoxide hydratase
AC18	Sodium channel protein type 10 subunit alpha Cell division protein FtsZ Sodium channel protein type 10 subunit alpha Atypical chemokine receptor 3 ATPase family AAA domain-containing protein 2	Serine/threonine-protein kinase PIM1 Inhibitor of nuclear factor kappa B kinase beta subunit Prokineticin receptor 1 WD repeat-containing protein 5 Protein kinase C delta
AC19	Pantothenate synthetase DNA repair protein RAD51 homolog 1 Apoptosis regulator BAX Luciferin 4-monooxygenase	Histone deacetylase 1 Beta-secretase 1 Hepatocyte growth factor receptor Cathepsin D Histone deacetylase 2
AC20	Sodium channel protein type 10 subunit alpha Melanin-concentrating hormone receptor 1 Cytochrome P450 11B2, mitochondrial 5-hydroxytryptamine receptor 1D Follicle-stimulating hormone receptor	TGF-beta receptor type I Protein tyrosine kinase 2 beta Sterol regulatory element-binding protein 2 Nuclear receptor ROR-gamma Cathepsin K
AC21	5-hydroxytryptamine receptor 3A Histamine N-methyltransferase Histidine-rich protein PFHRP-II 5-hydroxytryptamine receptor 2A Transforming protein RhoA	Vascular endothelial growth factor receptor 2 Proteasome Macropain subunit MB1 Prolyl endopeptidase Fibroblast activation protein alpha Proto-oncogene protein Wnt-3

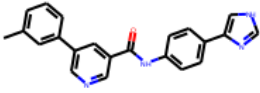
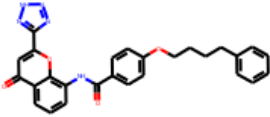
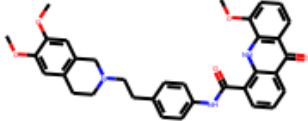
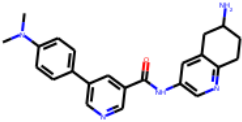

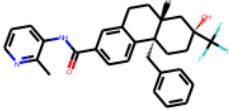
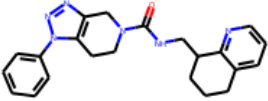
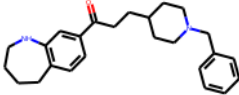
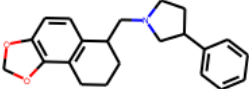
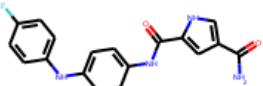
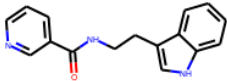
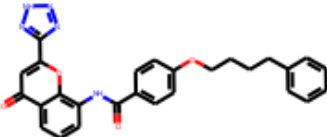
AC22	3 beta-hydroxysteroid dehydrogenase/Delta 5-->4- isomerase type 2 Histone lysine demethylase PHF8 Lysine-specific demethylase 2B Taste receptor type 1 member 2 Taste receptor type 1 member 3	Cannabinoid receptor 1 Cannabinoid receptor 2 Mu opioid receptor Kappa Opioid receptor Opioid growth factor receptor-like protein 1
AC23	Substance-P receptor Peptidyl-prolyl cis-trans isomerase B Kinesin-like protein KIF20A 5-hydroxytryptamine receptor 1B 5-hydroxytryptamine receptor 1D	Mu opioid receptor Advanced glycosylation end product-specific receptor Inhibitor of apoptosis protein 3 Purinergic receptor P2Y1 Serotonin 3a (5-HT3a) receptor

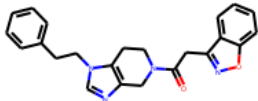
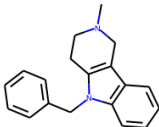
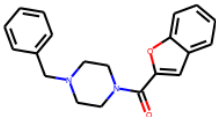
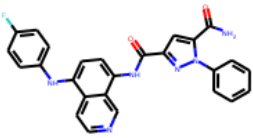
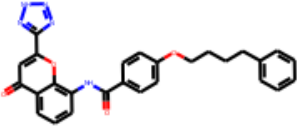
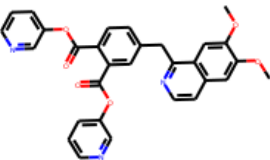
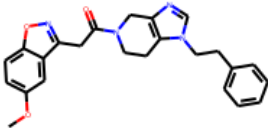
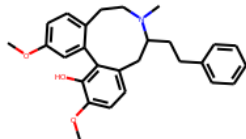
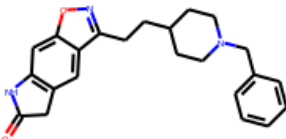
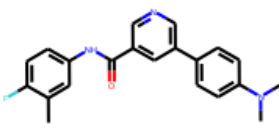
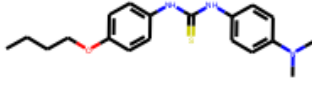
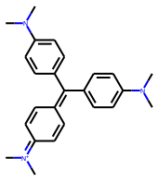
**Table S9: Activity predictions for candidate AChE inhibitors.** Predicted activity of the compounds selected by vScreenML, from SEA and SwissTarget. Neither recognizes these compounds as AChE inhibitors, suggesting that indeed these are new scaffolds for this target.

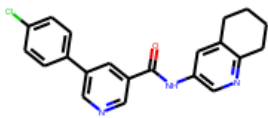
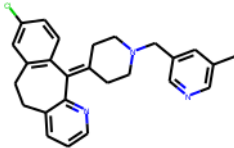
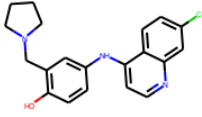
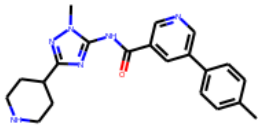
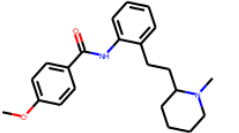
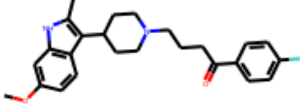
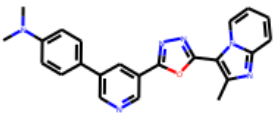
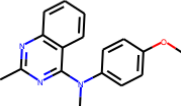
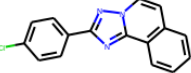
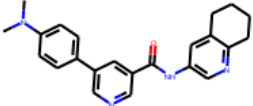
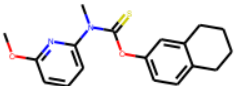
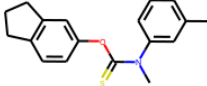
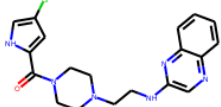
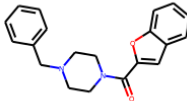
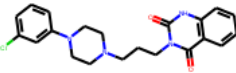
vScreenML compound	Closest matches amongst annotated AChE inhibitors in ChEMBL
 <p data-bbox="550 533 597 558">AC1</p>	 <p data-bbox="906 447 1096 472">CHEMBL1491339</p>  <p data-bbox="943 590 1058 615">CHEMBL45</p>
 <p data-bbox="550 842 597 867">AC2</p>	 <p data-bbox="922 737 1089 762">CHEMBL94113</p>  <p data-bbox="906 888 1096 913">CHEMBL2104675</p>
 <p data-bbox="550 1161 597 1186">AC3</p>	 <p data-bbox="906 1083 1096 1108">CHEMBL2105485</p>  <p data-bbox="906 1230 1096 1255">CHEMBL2107681</p>
 <p data-bbox="550 1503 597 1528">AC4</p>	 <p data-bbox="906 1409 1096 1434">CHEMBL1213257</p>  <p data-bbox="922 1581 1084 1606">CHEMBL21333</p>

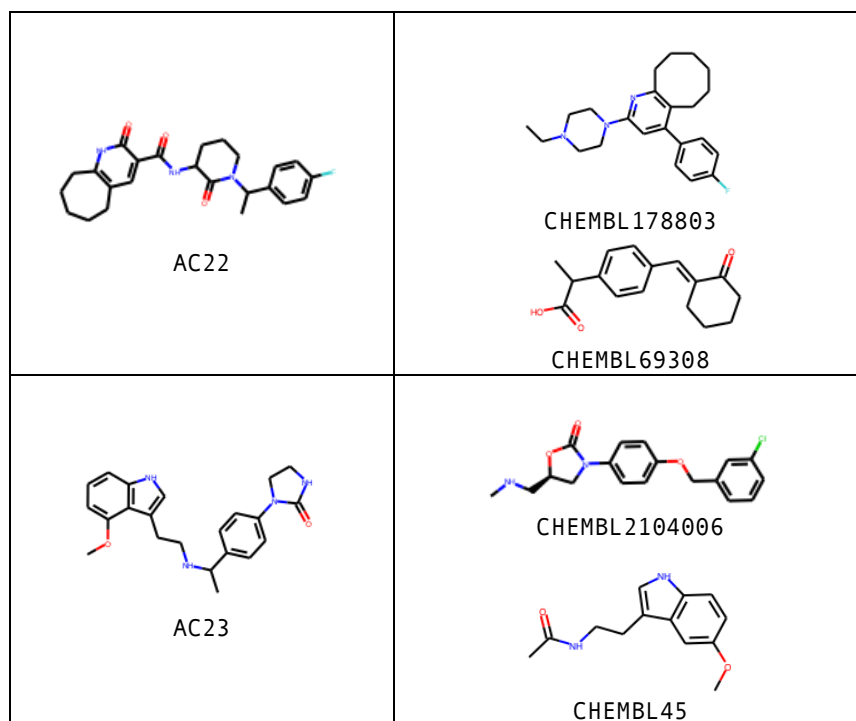
 <p>AC5</p>	 <p>CHEMBL66648</p>  <p>CHEMBL29404</p>
 <p>AC6</p>	 <p>CHEMBL2104597</p>  <p>CHEMBL2104675</p>
 <p>AC7</p>	 <p>CHEMBL1491339</p>  <p>CHEMBL218166</p>
 <p>AC8</p>	 <p>CHEMBL2106056</p>  <p>CHEMBL348763</p>



 <p>AC9</p>	 <p>CHEMBL21333</p>  <p>CHEMBL569424</p>
 <p>AC10</p>	 <p>CHEMBL61872</p>  <p>CHEMBL3137304</p>
 <p>AC11</p>	 <p>CHEMBL75013</p>  <p>CHEMBL27926</p>
 <p>AC12</p>	 <p>CHEMBL1491339</p>  <p>CHEMBL21333</p>

 <p>AC13</p>	 <p>CHEMBL1625607</p>  <p>CHEMBL1076256</p>
 <p>AC14</p>	 <p>CHEMBL21333</p>  <p>CHEMBL2105169</p>
 <p>AC15</p>	 <p>CHEMBL2104050</p>  <p>CHEMBL359570</p>
 <p>AC16</p>	 <p>CHEMBL2107052</p>  <p>CHEMBL459265</p>

 <p>AC17</p>	 <p>CHEMBL91397</p>  <p>CHEMBL1213257</p>
 <p>AC18</p>	 <p>CHEMBL315838</p>  <p>CHEMBL2104902</p>
 <p>AC19</p>	 <p>CHEMBL492399</p>  <p>CHEMBL94113</p>
 <p>AC20</p>	 <p>CHEMBL1591365</p>  <p>CHEMBL2105583</p>
 <p>AC21</p>	 <p>CHEMBL1076256</p>  <p>CHEMBL2110792</p>



**Table S10: Comparison of vScreenML's candidate AChE inhibitors to known AChE inhibitors.** For each compound selected by vScreenML as a candidate AChE inhibitor, we present the two closest compounds (as measured by 2D fingerprint similarity) that are annotated in ChEMBL as AChE inhibitors. The lack of similarity implies that these compounds would not have been identified as AChE inhibitors on the basis of 2D fingerprint similarity.

Compound	Identified as a top-scoring hit
AC6	1 <sup>st</sup> round
AC3	1 <sup>st</sup> round
AC10	Both rounds
AC11	1 <sup>st</sup> round
AC15	2 <sup>nd</sup> round
AC5	2 <sup>nd</sup> round
AC13	2 <sup>nd</sup> round
AC19	Both rounds
AC23	1 <sup>st</sup> round
AC9	Both rounds

**Table S11: Provenance of AChE inhibitors.** For each of the 10 AChE inhibitors that provided more than 50% inhibition at a concentration of 50  $\mu$ M, we determined at what stage this compound was prioritized for testing. Our strategy included two stages of screening: first we screened only 15 million diverse compounds from the Enamine collection, then we expanded our search by collecting analogs for each of these hits. We note that 7 of these 10 compounds were identified in the first round of screening; after re-refinement in the second round, 3 of these were still highly-ranked whereas 4 had been surpassed by analogs (or received lower scores upon re-refinement). Only 3 of these 10 compounds would have been missed if our screening had been limited to a single round of 15 million compounds.