Supporting Information

LIT-PCBA: An Unbiased Dataset for Machine Learning and Virtual Screening.

Viet-Khoa Tran-Nguyen,[†] Célien Jacquemard* and Didier Rognan[†]*

[†] Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, 67400 Illkirch, France.

* To whom correspondence should be addressed (phone: +33 3 68 85 42 35, fax: +33 3 68 85 43 10, email: rognan@unistra.fr)

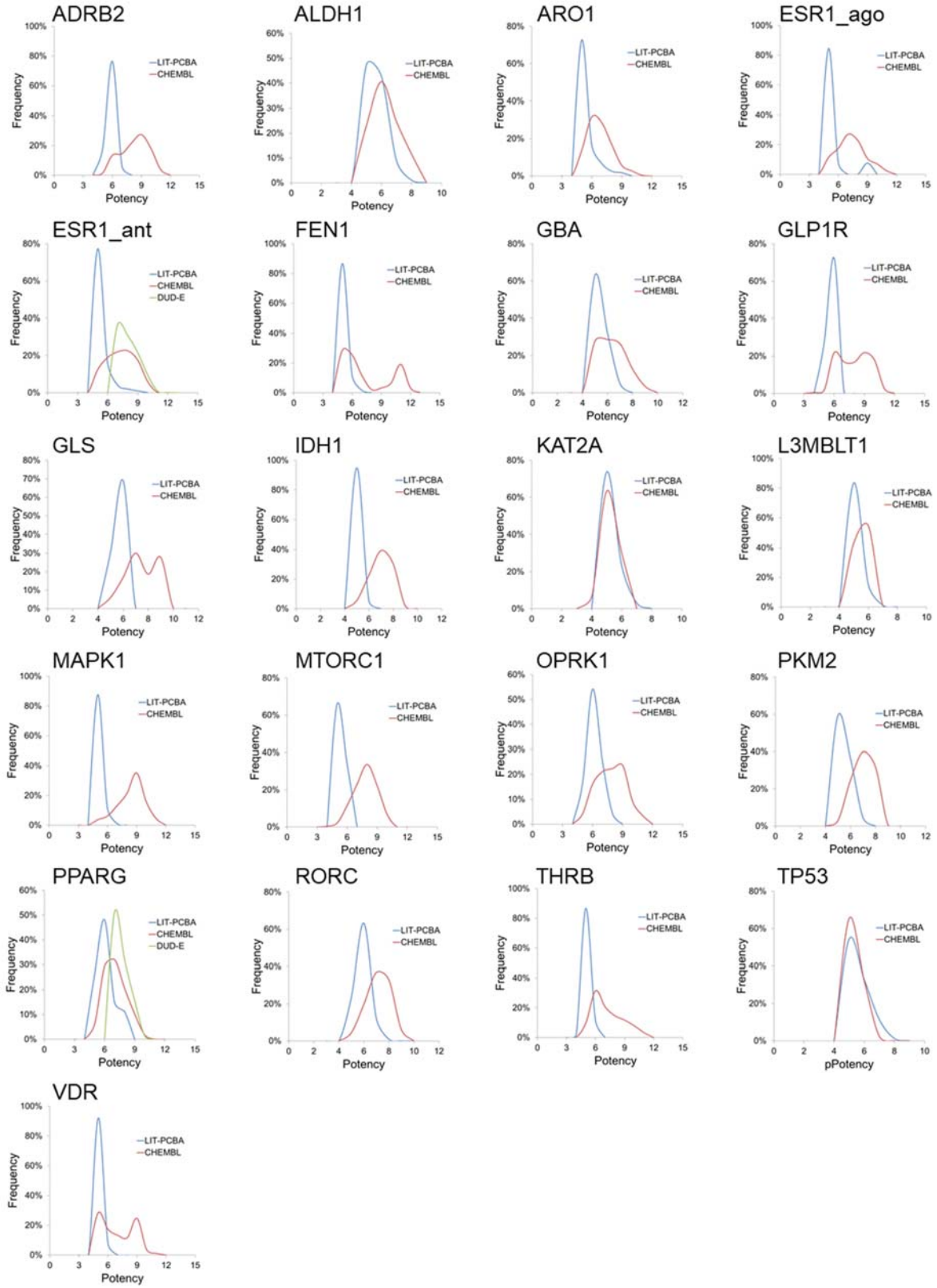

Figure S1. Comparison of potencies (in pIC_{50}, pEC_{50}, pK_i, pK_d) for confirmed actives of the LIT-PCBA, DUD-E and ChEMBL ligands.

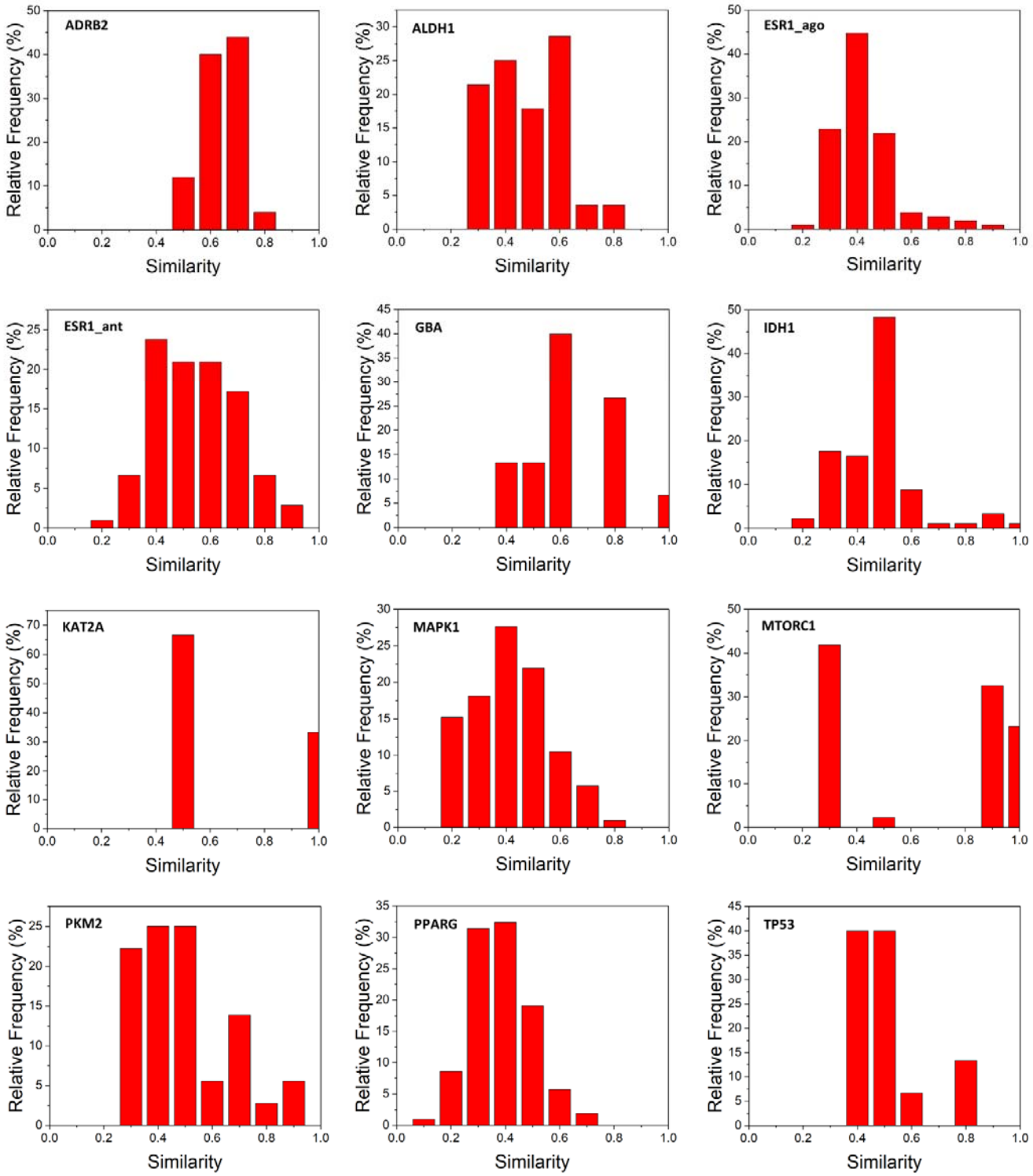

Figure S2. Self-similarity matrix of PDB template ligands. Pairwise similarity between template ligands is expressed by a Tanimoto coefficient calculated from MDL public keys implemented in PipelinePilot 2019.[40] No analysis is provided for three target sets (FEN1, OPRK1, VDR) for which a single PDB template ligand is available.

Table S1. Description of 21 selected PubChem BioAssays.

Target set	AID	Assay description	Readout	Format	PDB templates
ADRB2	492947	qHTS assay of beta-arrestin-biased ligands of beta2-adrenergic receptor	Lumi	CBA	3P0G, 3PDS, 3SN6, 4LDE, 4LDL, 4LDO, 4QKX, 6MXT
ALDH1	1030	qHTS Assay for Inhibitors of Aldehyde Dehydrogenase 1	Fluo	EAE	4WP7, 4WPN, 4X4L, 5AC2, 5L2M, 5L2N, 5L2O, 5TEI
ARO1	743083	qHTS assay to identify aromatase inhibitors	Fluo	CBA	3S7S, 4GL5, 4GL7
ESR1_ago	743075	qHTS assay to identify small molecule agonists of the estrogen receptor alpha (ER-alpha) signaling pathway	Fluo	CBA	1L2I,2B1V, 2B1Z, 2P15, 2Q70, 2QR9, 2QSE, 2QZO, 4IVW, 4PPS, 5DRJ, 5DU5, 5DUE, 5DZI, 5E1C
ESR1_ant	743080	qHTS assay to identify small molecule antagonists of the estrogen receptor alpha (ER-alpha) signaling pathway using the BG1 cell line	Lumi	CBA	1XP1, 1XQC, 2YAR, 2IOG, 2IOK, 2OUZ, 2POG, 2R6W, 3DT3, 5AAU, 5FQV, 5T92, 5UFX, 6B0F, 6CHW
FEN1	588795	qHTS Assay for the Inhibitors of Human Flap endonuclease 1	Fluo	EAE	5FV7
GBA	2101	qHTS Assay for Inhibitors and Activators of N370S glucocerebro-sidase as a Potential Chaperone Treatment of Gaucher Disease	Fluo	EAE	2V3D, 2V3E, 2XWD, 2XWE, 3RIK, 3RIL
GLP1R	624417	qHTS of GLP-1 Receptor Inverse Agonists	Lumi	CBA	5VEW, 5VEX
GLS	624170	qHTS for Inhibitors of Glutaminase	Fluo[a]	EAE[b]	3UO9, 3VOZ, 3VP1, 5FI2, 5FI6, 5FI7, 5HL1, 5I94, 5JYO, 5WJ6, 5JYP
IDH1	602179	qHTS for Inhibitors of mutant isocitrate dehydrogenase 1	Fluo	EAE	4I3K, 4I3L, 4UMX*, 4XRX, 4XS3, 5DE1*, 5L57*, 5L58*, 5LGE*, 5SUN*, 5SVF*, 5TQH*, 6ADG*, 6B0Z*
KAT2A	504327	qHTS Assay for Inhibitors of GCN5L2	Fluo	PPI	5H84, 5H86, 5MLJ
L3MBTL1	485360	qHTS Assay for the Inhibitors of L3MBTL1	Alpha	PPI	3P8H
MAPK1	995	qHTS Assay for Inhibitors of the ERK Signaling Pathway using a Homogeneous Screening Assay	Alpha[e]	CBA	1PME, 2OJG, 3SA0, 3W55, 4QP3, 4QP4, 4QP9, 4QTA, 4QTE, 4WJ0, 4ZZN, 5AX3, 5BUJ, 5V62, 6G9H
MTROC1	493208	Acumen qHTS Assay for Inhibitors of the mTORC1 Signaling Pathway in MEF (Tsc2-/-, p53-/-) Cells: Sytravon	Fluo	CBA	1FAP, 1NSG, 2FAP, 3FAP, 4DRH, 4DRI, 4DRJ, 4FAP, 4JSX*, 4JT5*, 5GPG
OPRK1	1777	uHTS identification of small molecule agonists of the kappa opioid receptor via a luminescent beta-arrestin assay	Lumi[c]	CBA[d]	6B73

PKM2	1631	qHTS Assay for Activators of Human Muscle isoform 2 Pyruvate Kinase	Lumi	EAE	3GQY, 3GR4, 3H6O, 3ME3, 3U2Z, 4G1N, 4JPG, 5X1V, 5X1W
PPARG	743094	qHTS assay to identify small molecule agonists of the peroxisome proliferator-activated receptor gamma (PPARg) signaling pathway	Fluo	CBA	1ZGY, 2I4J, 2P4Y, 2Q5S, 2YFE, 3B1M, 3HOD, 3R8A, 4CI5, 4FGY, 4PRG, 5TTO, 5TWO, 5Y2T, 5Z5S
RORC	2551	qHTS for inhibitors of ROR gamma transcriptional activity	Lumi	CBA	4WPQ, 4YMQ, 5APH, 5C4T, 5NTK, 5NTN, 5NTP, 5NTQ, 5NTW, 5UFR, 5VB6, 5X8Q, 6A22, 6B33, 6CVH
THRB	1469	qHTS for Inhibitors of the Interaction of Thyroid Hormone Receptor and Steroid Receptor Coregulator 2	FP[f]	PPI[g]	2PIN
TP53	651631	qHTS assay for small molecule agonists of the p53 signaling pathway	Fluo	CBA	2VUK, 3ZME, 4AGO, 4AGQ, 5G4O, 5O1I
VDR	504847	Inhibitors of the vitamin D receptor (VDR): qHTS	FP	PPI	3A2J, 3A2I

[a] fluorescence intensity

[b] enzyme activity assay

[c] luminescence

[d] cell-based assay

[e] alpha screen

[f] fluorescence polarization

[g] soluble protein-protein interaction assay

Table S2. Number of confirmed active compounds remaining after each filtering step.

Target set	PubChem AID	Start	Filtering steps					
			Step 1	Step 2a	Step 2b	Step 2c	Step 3	Step4
ADRB2	492947	80	80	19	19	19	17	17
ALDH1	1030	16117	16070	8052	8023	7716	7170	7,168
ARO1	743083	905	852	298	150	150	121	121
ESR1_ago	743075	105	89	20	18	18	15	13
ESR1_ant	743080	473	453	217	145	145	103	102
FEN1	588795	1368	1353	502	448	425	370	369
GBA	2101	299	298	240	236	233	166	166
GLP1R	624417	6432	6431	3000	2997	2942	2180	2,180
GLS	624170	846	842	255	251	236	224	224
IDH1	602179	365	364	57	56	54	39	39
KAT2A	504327	817	794	297	268	234	194	194
L3MBTL1	485360	1495	1492	587	583	541	501	501
MAPK1	995	711	707	414	402	322	308	308
MTORC1	493208	342	342	137	136	136	97	97
OPRK1	1777	35	35	30	30	29	24	24
PKM2	1631	892	892	578	578	557	546	546
PPARG	743094	78	75	46	41	41	27	27
RORC	2551	16824	16805	8397	8355	8053	6874	6,874
THRB	1469	183	179	92	78	64	53	53
TP53	651631	602	571	181	111	111	81	79
VDR	504847	3735	3685	1099	1067	1041	886	884
Unique compounds		45771	45294	23058	22653	21819	18939	18930
% remaining		100.00	98.96	50.38	49.49	47.67	41.38	41.36

Table S3. Number of inactive compounds remaining after each filtering step.

Target set	PubChem AID	Start	Filtering steps			Final Actives/Inactives ratio
			Step 1	Step 3	Step 4	
ADRB2	492947	329716	329642	312493	312483	1/18381
ALDH1	1030	148322	148166	137980	137965	1/19
ARO1	743083	8846	8661	5440	5381	1/44
ESR1_ago	743075	9089	8897	5640	5583	1/429
ESR1_ant	743080	8297	8121	5003	4948	1/49
FEN1	588795	382244	382117	355420	355402	1/963
GBA	2101	314877	314654	296080	296052	1/1783
GLP1R	624417	321735	321657	304879	304866	1/140
GLS	624170	401810	401672	371883	371860	1/1660
IDH1	602179	388463	388376	362063	362049	1/9283
KAT2A	504327	376634	376467	348571	348548	1/1797
L3MBTL1	485360	217165	217107	204490	204480	1/408
MAPK1	995	66078	65908	62652	62629	1/203
MTORC1	493208	41294	41294	32972	32972	1/340
OPRK1	1777	284169	284120	269818	269816	1/11242
PKM2	1631	259866	259782	245525	245523	1/450
PPARG	743094	8532	8357	5267	5211	1/193
RORC	2551	256777	256580	243311	243284	1/35
THRB	1469	281374	281090	254491	254442	1/4801
TP53	651631	6973	6836	4215	4168	1/53
VDR	504847	384189	383989	355415	355388	1/402
Unique compounds		464805	464047	422400	422256	
% remaining		100.00	99.84	90.88	90.85	

Table S4. Virtual screening results obtained by 2D ECFP4 similarity search on 21 fully processed selected target sets.

Target set	PubChem AID	ROC				EF1%			
		Min	Max	Mean ± SD	Fused	Min	Max	Mean ± SD	Fused
ADRB2	492947	0.53	0.70	0.63 ± 0.06	0.68	5.88	23.53	18.38 ± 6.62	17.65
ALDH1	1030	0.49	0.52	0.51 ± 0.01	0.52	0.99	2.30	1.40 ± 0.50	2.61
ARO1	743083	0.50	0.52	0.51 ± 0.01	0.52	1.65	1.65	1.65	1.65
ESR1_ago	743075	0.56	0.72	0.65 ± 0.05	0.72	0.00	20.51	7.52 ± 6.09	7.69
ESR1_ant	743080	0.42	0.54	0.50 ± 0.03	0.50	0.00	3.19	1.39 ± 1.06	0.98
FEN1	588795	0.44	0.44	0.44	0.44	1.08	1.08	1.08	1.08
GBA	2101	0.45	0.53	0.50 ± 0.03	0.48	0.00	4.22	2.01 ± 1.41	3.61
GLP1R	624417	0.48	0.50	0.49 ± 0.01	0.50	1.28	1.79	1.54 ± 0.36	1.61
GLS	624170	0.32	0.37	0.34 ± 0.02	0.33	0.00	0.45	0.04 ± 0.14	0.00
IDH1	602179	0.28	0.52	0.42 ± 0.07	0.38	0.00	10.26	1.83 ± 3.09	5.13
KAT2A	504327	0.36	0.37	0.40 ± 0.06	0.44	0.00	0.52	0.17 ± 0.30	0.52
L3MBTL1	485360	0.41	0.41	0.41	0.41	0.98	0.98	0.98	0.98
MAPK1	995	0.45	0.58	0.52 ± 0.04	0.53	0.00	2.92	1.20 ± 0.74	1.95
MTORC1	493208	0.47	0.52	0.48 ± 0.02	0.45	0.00	1.03	0.28 ± 0.48	0.00
OPRK1	1777	0.69	0.69	0.69	0.69	12.50	12.50	12.50	12.50
PKM2	1631	0.41	0.64	0.55 ± 0.08	0.64	0.18	7.68	3.30 ± 2.73	7.95
PPARG	743094	0.58	0.80	0.68 ± 0.07	0.78	0.00	14.81	3.65 ± 5.14	7.41
RORC	2551	0.36	0.56	0.43 ± 0.05	0.44	0.20	1.55	0.55 ± 0.38	0.46
THRB	1469	0.37	0.37	0.37	0.37	0.00	0.00	0.00	0.00
TP53	651631	0.38	0.56	0.47 ± 0.06	0.42	0.00	2.53	1.06 ± 0.95	0.00
VDR	504847	0.44	0.44	0.44	0.44	3.37	3.37	3.37	3.37
ALL		0.45	0.54	0.50 ± 0.05	0.51	1.34	5.57	3.06 ± 1.43	3.67

Table S5. Virtual screening results obtained by 3D shape similarity search on 21 fully processed selected target sets.

Target set	PubChem AID	ROC				EF1%			
		Min	Max	Mean ± SD	Fused	Min	Max	Mean ± SD	Fused
ADRB2	492947	0.47	0.66	0.53 ± 0.06	0.67	6.67	21.43	9.76 ± 5.35	20.00
ALDH1	1030	0.46	0.53	0.50 ± 0.03	0.49	0.81	1.96	1.20 ± 0.39	1.85
ARO1	743083	0.60	0.69	0.65 ± 0.05	0.62	1.69	1.69	1.69	1.69
ESR1_ago	743075	0.46	0.65	0.56 ± 005	0.65	0.00	15.38	4.61 ± 4.86	7.69
ESR1_ant	743080	0.58	0.64	0.60 ± 0.02	0.61	0.00	4.90	1.66 ± 1.55	3.92
FEN1	588795	0.45	0.45	0.45	0.45	0.27	0.27	0.27	0.27
GBA	2101	0.33	0.40	0.38 ± 0.03	0.34	0.00	0.86	0.72 ± 0.35	0.86
GLP1R	624417	0.51	0.52	0.52 ± 0.01	0.52	0.47	1.38	0.93 ± 0.64	0.38
GLS	624170	0.37	0.45	0.40 ± 0.03	0.44	0.00	2.07	0.50 ± 0.88	1.38
IDH1	602179	0.35	0.50	0.41 ± 0.05	0.39	0.00	8.00	0.86 ± 2.32	0.00
KAT2A	504327	0.38	0.44	0.39 ± 0.03	0.43	1.44	1.55	1.48 ± 0.07	1.55
L3MBTL1	485360	0.50	0.50	0.50	0.50	1.18	1.18	1.18	1.18
MAPK1	995	0.45	0.62	0.53 ± 0.05	0.55	0.32	4.09	1.96 ± 1.01	5.19
MTORC1	493208	0.44	0.52	0.47 ± 0.03	0.52	0.00	3.26	1.38 ± 1.29	3.23
OPRK1	1777	0.55	0.55	0.55	0.55	0.00	0.00	0.00	0.00
PKM2	1631	0.48	0.67	0.60 ± 0.07	0.59	0.00	11.32	4.86 ± 4.30	6.22
PPARG	743094	0.59	0.76	0.72 ± 0.05	0.73	0.00	18.52	8.79 ± 5.24	18.52
RORC	2551	0.38	0.51	0.45 ± 0.03	0.44	0.25	0.82	0.56 ± 0.18	0.33
THRB	1469	0.57	0.57	0.57	0.57	0.00	0.00	0.00	0.00
TP53	651631	0.54	0.62	0.58 ± 0.04	0.56	0.00	2.63	0.88 ± 1.36	2.53
VDR	504847	0.37	0.37	0.37	0.37	0.35	0.35	0.35	0.35
ALL		0.47	0.55	0.51 ± 0.03	0.52	0.64	4.84	2.08 ± 1.42	3.67

Table S6. Virtual screening results obtained by molecular docking on 21 fully processed selected target sets.

Target set	PubChem AID	ROC				EF1%			
		Min	Max	Mean ± SD	Fused	Min	Max	Mean ± SD	Fused
ADRB2	492947	0.41	0.52	0.46 ± 0.04	0.44	0.00	5.88	2.94 ± 3.14	5.88
ALDH1	1030	0.51	0.53	0.52 ± 0.01	0.53	1.09	1.38	1.19 ± 0.10	0.96
ARO1	743083	0.47	0.53	0.51 ± 0.03	0.5	0.00	0.83	0.28 ± 0.48	0.00
ESR1_ago	743075	0.26	0.51	0.36 ± 0.07	0.48	0.00	0.00	0.00	0.00
ESR1_ant	743080	0.43	0.54	0.50 ± 0.03	0.53	0.00	1.96	0.91 ± 0.78	0.98
FEN1	588795	0.47	0.47	0.47	0.47	4.34	4.34	4.34	4.34
GBA	2101	0.48	0.72	0.64 ± 0.08	0.69	2.41	7.83	5.72 ± 1.97	8.43
GLP1R	624417	0.50	0.51	0.51 ± 0.01	0.51	1.06	1.19	1.13 ± 0.09	1.01
GLS	624170	0.34	0.43	0.38 ± 0.03	0.35	0.00	2.68	0.93 ± 1.05	0.00
IDH1	602179	0.30	0.48	0.37 ± 0.05	0.38	0.00	2.56	0.73 ± 1.20	0.00
KAT2A	504327	0.35	0.40	0.38 ± 0.03	0.39	1.55	2.58	1.89 ± 0.59	2.58
L3MBTL1	485360	0.45	0.45	0.45	0.45	0.80	0.80	0.80	0.80
MAPK1	995	0.48	0.55	0.52 ± 0.02	0.54	0.65	2.60	1.32 ± 0.53	1.62
MTORC1	493208	0.49	0.54	0.52 ± 0.02	0.51	0.00	2.06	1.03 ± 0.65	1.03
OPRK1	1777	0.58	0.58	0.58	0.58	4.17	4.17	4.17	4.17
PKM2	1631	0.49	0.56	0.54 ± 0.02	0.62	0.18	1.28	0.57 ± 0.37	0.18
PPARG	743094	0.65	0.74	0.69 ± 0.02	0.71	0.00	14.81	5.92 ± 5.20	7.41
RORC	2551	0.36	0.42	0.37 ± 0.01	0.36	0.22	0.45	0.31 ± 0.07	0.39
THRB	1469	0.36	0.36	0.36	0.36	1.89	1.89	1.89	1.89
TP53	651631	0.47	0.57	0.51 ± 0.04	0.51	0.00	0.00	0.00	0.00
VDR	504847	0.34	0.36	0.35 ± 0.01	0.34	0.00	0.68	0.34 ± 0.48	0.00
ALL		0.44	0.51	0.48 ± 0.02	0.49	0.87	2.86	1.73 ± 0.80	1.98

Table S7. Chemical diversity of PDB template ligands, assessed by the number of unique Bemis-Murcko frameworks.[51]

Target set	Number of PDB templates	Number of Bemis-Murcko scaffolds
ADRB2	8	5
ALDH1	8	8
ESR1_ago	15	14
ESR1_ant	15	15
FEN1	1	1
GBA	6	4
IDH1	14	14
KAT2A	3	2
MAPK1	15	15
MTORC1	11	5
OPRK1	1	1
PKM2	9	8
PPARG	15	15
TP53	6	5
VDR	2	1

Bemis-Murcko frameworks were computed from mol2 files, with the 'Generate Fragments' component of Pipeline Pilot 2019.

Table S8. Virtual screening results (EF1%) obtained by 2D ECFP4 similarity search, 3D shape similarity search and molecular docking on 15 validation sets after debiasing with AVE.

Target set	PubChem AID	2D ECFP4 similarity search				3D shape similarity search				Molecular docking			
		Min	Max	Mean ± SD	Fused	Min	Max	Mean ± SD	Fused	Min	Max	Mean ± SD	Fused
ADRB2	492947	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ALDH1	1030	0.82	2.75	1.58 ± 0.62	2.68	0.67	1.64	1.08 ± 0.35	1.64	0.89	1.56	1.25 ± 0.23	0.82
ESR1_ago	743075	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ESR1_ant	743080	0.00	12.00	2.67 ± 3.60	0.00	0.00	4.00	1.07 ± 1.83	4.00	0.00	4.00	1.60 ± 2.03	4.00
FEN1	588795	1.09	1.09	1.09	1.09	0.00	0.00	0.00	0.00	3.26	3.26	3.26	3.26
GBA	2101	0.00	2.44	1.63 ± 1.26	2.44	0.00	4.88	0.81 ± 1.99	0.00	0.00	9.76	4.47 ± 3.59	4.88
IDH1	602179	0.00	11.11	1.59 ± 4.03	0.00	0.00	11.11	0.79 ± 2.97	0.00	0.00	11.11	0.79 ± 2.97	0.00
KAT2A	504327	0.00	2.08	0.69 ± 1.20	0.00	0.00	2.08	0.69 ± 1.20	0.00	2.08	6.25	4.17 ± 2.09	2.08
MAPK1	995	0.00	5.19	0.95 ± 1.43	1.30	0.00	5.19	1.39 ± 1.59	2.60	0.00	5.19	1.99 ± 1.38	1.30
MTORC1	493208	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.17	1.52 ± 2.10	4.17
OPRK1	1777	16.67	16.67	16.67	16.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PKM2	1631	0.00	4.41	1.31 ± 1.79	0.74	0.00	5.15	2.13 ± 1.93	2.21	0.00	1.47	0.90 ± 0.61	0.74
PPARG	743094	0.00	16.67	5.56 ± 8.13	16.67	0.00	16.67	5.56 ± 8.13	16.67	0.00	16.67	5.56 ± 8.13	0.00
TP53	651631	0.00	0.00	0.00	0.00	0.00	5.26	0.88 ± 2.15	0.00	0.00	0.00	0.00	0.00
VDR	504847	3.64	3.64	3.64	3.64	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ALL		1.48	5.20	2.49 ± 1.47	3.02	0.04	3.73	0.96 ± 1.48	1.81	0.42	4.23	1.70 ± 1.54	1.42