Supplementary Information

LIDeB's Useful decoys

ChemBL30 database curation

The complete ChEMBL30 database, containing approximately 2.3 million chemical structures, was downloaded and curated. The compounds were standardized using the MolVS package [9] returning the largest organic covalent unit in the molecule, has all atoms replaced with the most abundant isotope for that element and their charge removed. Next, the protonation state was calculated at pH 7.4, with the appropriate charges, using the Openbabel module [37]. Finally, duplicate molecules were removed.

For each molecule, the following physicochemical descriptors were calculated, employing the package Rdkit.Chem: molecular weight (MW), log P (LogP), number of rotatable bonds (nRotB), number of H-bond acceptors (nHAcc), number of H-bond donors (nHDon) and formal charge (Charge). The compounds that presented extreme values of these descriptors were removed, retaining those that presented: 100 < MW < 1000, -5 < logP < 10, nRotB <20, nHAcc <20, nHDon < 20 and 10 < Charge < 10. In this way, only 1.9 million compounds remained, that were distributed along 13 subsets to be easily accessible for online purposes.

Decoys obtained from each subset

Table S1. Number of decoy compounds obtained from a given number of query compounds on each DUD-E [30] subset.

Subset	Number of queries	Number of decoys
aa2ar	482	20006
abl1	182	9095
ace	282	13278
aces	453	19822
ada	93	4605
ada17	532	20274
adrb1	247	11624
adrb2	231	11419
akt1	293	13871
akt2	117	5794

aldr	159	7661
ampc	48	2400
andr	269	11686
aofb	122	5940
bace1	283	12256
braf	152	7523
cah2	492	23845
casp3	199	9647
cdk2	474	15372
comt	41	2050
cp2c9	120	5953
cp3a4	170	8275
csf1r	166	8050
cxcr4	40	2000
def	102	5100
dhi1	330	15060
dpp4	533	22493
drd3	480	18887
dyr	231	10926
egfr	542	23605
esr1	383	17922
esr2	367	17174
fa10	537	20652
fa7	114	5700
fabp4	47	2111
fak1	100	5000
fgfr1	139	6910
fkb1a	111	5550
fnta	592	24778
fpps	85	4238
gcr	258	11681
glcm	54	2650
gria2	158	7862

grik1	101	4794
hdac2	185	8972
hdac8	170	8431
hivint	100	5000
hivpr	536	22738
hivrt	338	14785
hmdh	170	8340
hs90a	88	4400
hxk4	92	4600
igf1r	148	7400
inha	43	2010
ital	138	6679
jak2	107	5350
kif11	116	5659
kit	166	8271
kith	57	2850
kpcb	135	6750
lck	420	18309
lkha4	171	8438
mapk2	101	5050
mcr	94	4700
met	166	8207
mk01	79	3950
mk10	104	5200
mk14	578	21633
mmp13	572	19258
mp2k1	121	6050
nos1	100	5000
nram	98	4575
pa2ga	99	4885
parp1	508	24545
pde5a	398	17438
pgh1	195	9641
pgh2	435	15554

plk1	107	5300
pnph	103	5150
ppara	373	7893
ppard	240	7966
pparg	484	9099
prgr	293	11200
ptn1	130	6500
pur2	50	2128
pygm	77	3799
pyrd	111	5400
reni	104	5138
rock1	100	4953
rxra	131	6070
sahh	63	2507
src	524	21509
tgfr1	133	6354
thb	103	5150
thrb	461	17499
try1	449	17501
tryb1	148	7283
tysy	109	5387
urok	162	8091
vgfr2	409	18335
wee1	102	4999
xiap	100	4927

Fast Druggability Assessment (FaDrA)

Best models’ equations and definitions of the descriptors.

Model 260

Y=-12.232+***SolventAccessibilityD1100****0.160 + ***SecondaryStrD1025****0.038 - ***ChargeD1100****0.027 - ***SolventAccessibilityD1050****0.037 + ***HydrophobicityD3025****0.025

Model 361

Y=8.400+***SecondaryStrD3001****(-0.109) - ***ChargeD2075****0.096 - ***PolarityD1001****0.089 + ***PolarizabilityD2050****0.028 - ***ChargeD3100****0.019

Model 424

Y= -12.245+***SolventAccessibilityD1100****0.174 - ***ChargeD2075****0.091 + ***NormalizedVDWVD3025****0.026 - ***NormalizedVDWVD3075****0.034 + ***PolarizabilityD3100****0.044

Model 763

Y=-8.019+***SolventAccessibilityD1100****0.160 - ***ChargeD2050****0.054 - ***NormalizedVDWVD3075****0.027 - ***SolventAccessibilityD1075****0.040 + ***PolarizabilityD3025****0.016

ChargeD1100:Charge distribution descriptor of the group 1 amino acids in the 100% of the protein sequence.

ChargeD2050:Charge distribution descriptor of the group 2 amino acids in the 50% of the protein sequence.

ChargeD2075:Charge distribution descriptor of the group 2 amino acids in the 75% of the protein sequence.

ChargeD3100:Charge distribution descriptor of the group 3 amino acids in the 100% of the protein sequence.

HydrophobicityD3025:Hydrophobicity distribution descriptor of the group 3 amino acids in the 25% of the protein sequence.

NormalizedVDWVD3025:Normalized Van der Waals volume distribution descriptor of the group 3 amino acids in the 25% of the protein sequence.

NormalizedVDWVD3075:Normalized Van der Waals volume distribution descriptor of the group 3 amino acids in the 75% of the protein sequence.

PolarityD1001:Polarity distribution descriptor of the group 1 amino acids for the first residue of the protein sequence.

PolarizabilityD2050:Polarizability distribution descriptor of the group 2 amino acids in the 50% of the protein sequence.

PolarizabilityD3025:Polarizability distribution descriptor of the group 3 amino acids in the 25% of the protein sequence.

PolarizabilityD3100:Polarizability distribution descriptor of the group 3 amino acids in the 100% of the protein sequence.

SecondaryStrD1025:Secondary structure distribution descriptor of the group 1 amino acids in the 25% of the protein sequence.

SecondaryStrD3001:Secondary structure distribution descriptor of the group 3 amino acids for the first residue of the protein sequence.

SolventAccessibilityD1050:Solvent accessibility distribution descriptor of the group 1 amino acids in the 50% of the protein sequence.

SolventAccessibilityD1075:Solvent accessibility distribution descriptor of the group 1 amino acids in the 75% of the protein sequence.

SolventAccessibilityD1100:Solvent accessibility distribution descriptor of the group 1 amino acids in the 100% of the protein sequence.