Supporting Information

for

Distributed Representation of Chemical Fragments

Suman K. Chakravarti[*,†]

[†]MultiCASE Inc., 23811 Chagrin Blvd., Suite 305, Beachwood, OH 44122, USA
[*]E-mail: chakravarti@gmail.com
Phone: +1-216-831-3740

1. Query Ligands for the 26 Kinase Targets

Obtained from: http://dude.docking.org.

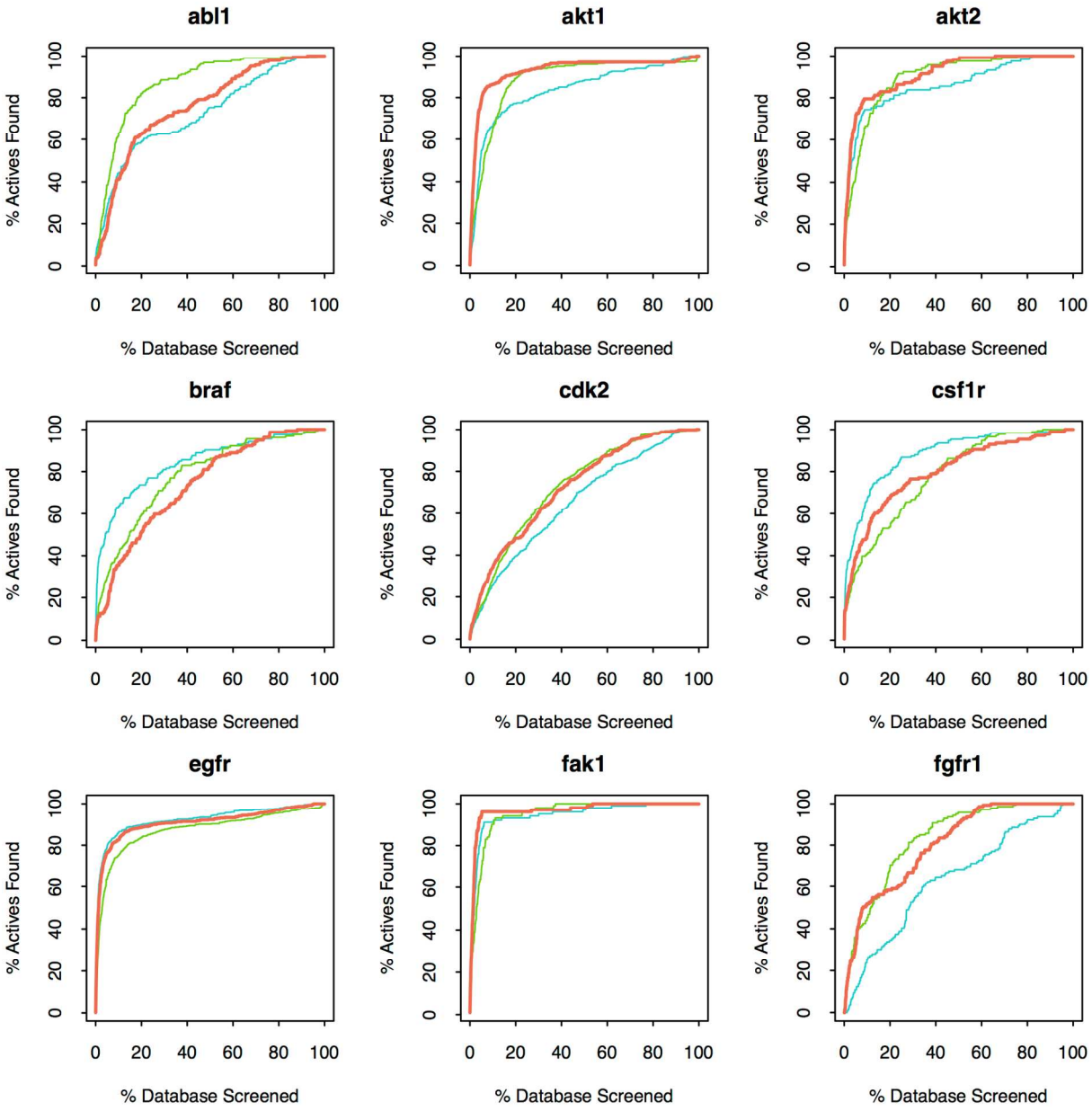
Reference:

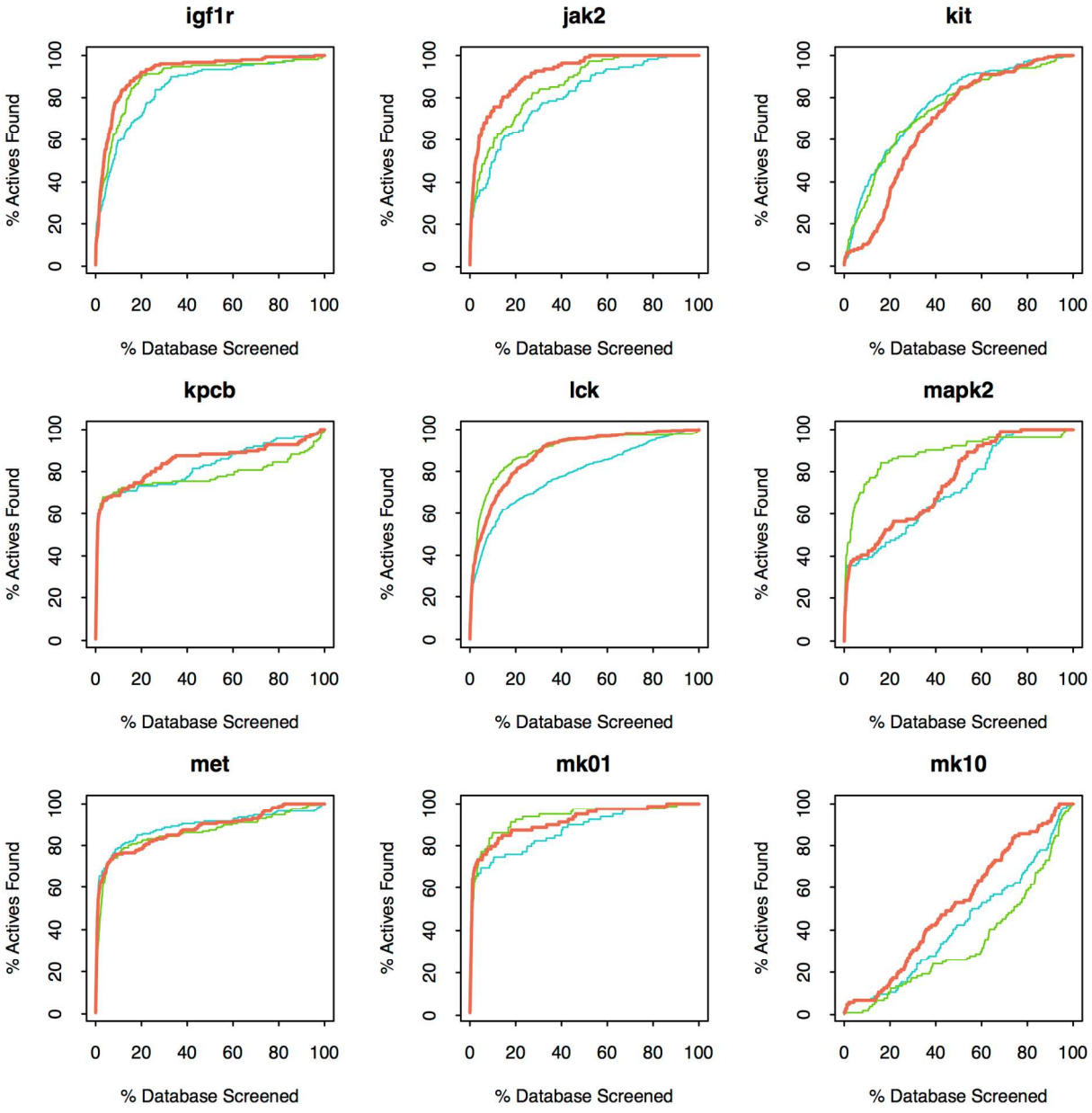
Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594.

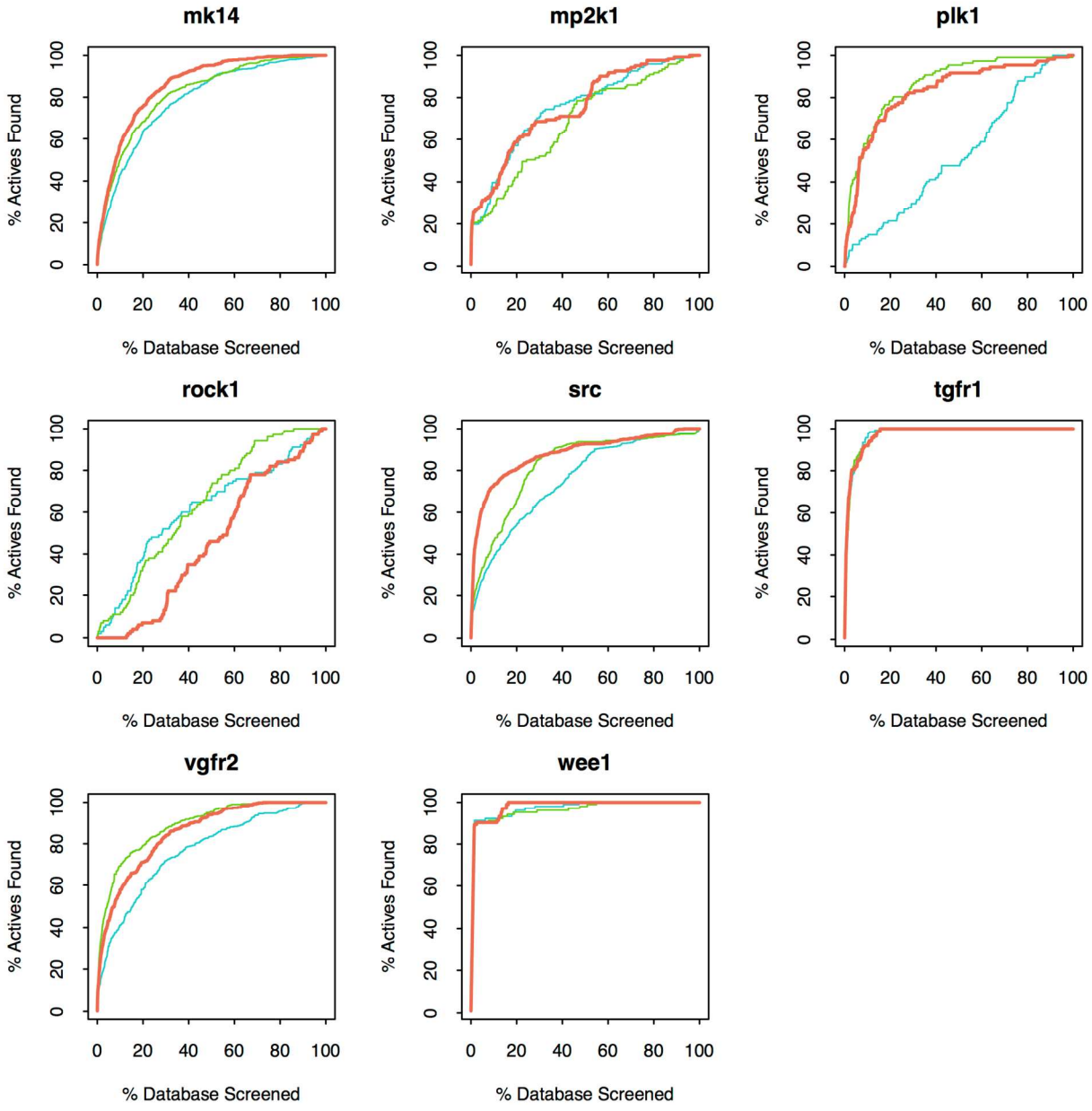
Kinase Target	Query Ligand SMILES
abl1	CN1C(=O)C(=Cc2cnc(Nc3ccc(F)c(C)c3)nc12)c4c(Cl)cccc4Cl
akt1	[H][n]1cnc2CCN(Cc21)c3ncnc4c3c(Cl)c[n]4[H]
akt2	CC[n]1c(nc2c(ncc(OCC3CCCNC3)c21)C#CC(C)(C)O)c4n[o]nc4N
braf	ONC1C=CC2C=C(C=CC2=1)c3c[n](nc3c4ccncc4)C5CCNCC5
cdk2	CN(C)CC(O)COc1ccc(Nc2cc(Nc3c(F)cccc3F)ncn2)cc1
csf1r	[H][n]1cc(nc1C(=O)Nc2ccc(cc2C3CCCCC=3)C4CCNCC=4)C#N
egfr	Nc1ncnc(Nc2ccc3c(cn[n]3Cc4cccc(F)c4)c2)c1CNN5CCCCC5
fak1	CN(c1ncccc1CNc2nc(Nc3cc[c]4=NC(=O)C=[c]4c3)ncc2C(F)(F)F)S(C)(=O)=O
fgfr1	[H][n]1cc(Cc2cccc(OC)c2)c3ccccnc31
igf1r	[H][n]1c(nc2cc(cc(C)c21)[n]3ccnc3)C4C(=O)NC=CC=4NCc5ccccn5
jak2	CNS(=O)(=O)c1ccc(cc1)c2cccc3ncc(nc32)c4cc(OC)c(OC)c(OC)c4
kit	[H][n]1c(C)c(C(N)=O)c(C)c1C=C2C(=O)Nc3ccc(F)cc32
kpcb	[H][n]1c(C)c(c2ccccc21)C3C(=O)NC(=O)C=3c4c(C)[n](CCCN(C)C)c5ccccc54
lck	C1CN(CCN1)CCNc2ccnc3[o]c(c4ccccc4)c(c5ccccc5)c23
mapk2	[H][n]1c2c(CCc3cnc(cc32)c4ccccc4F)c5C(=O)N=CC6(CN(C)C6)c51
met	COc1cc2c(Oc3ccc(NC(=O)C4(CC4)C(=O)Nc5ccc(F)cc5)cc3F)ccnc2cc1OCCCN6CCOCC6
mk01	[H][n]1cc(cc1C(=O)N(C)C)c2c(cn[n]2[H])c3ccccc3
mk10	CCC(=O)C1N(Cc2ccc(cc2)C(O)=O)C(=O)c3ccc(Cl)cc3C=1c4ccccc4
mk14	[H][n]1cc(C(=O)C(=O)N2CCC(O)C2)c3cc(C(=O)N4CC[n]5c(C4)cnc5c6ccc(F)cc6F)c(OC)nc31
mp2k1	NC(Sc1ccccc1N)=C(C#N)C(C#N)=C(N)Sc2ccccc2N
plk1	[H][n]1nc(NC(=O)c2ccc(cc2)N3CCN(C)CC3)c4c[n](cc41)C(=O)Cc5ccc[s]5
rock1	CC(N)C1CCC(CC1)C(=O)Nc2ccncc2
src	CC(C)[n]1nc(c2ccc(NC(=O)Nc3cccc(c3)C(F)(F)F)cc2)c4c(N)ncnc41
tgfr1	Cc1cccc(n1)c2nc(Nc3ccncc3)c4ccccc4n2
vgfr2	O=C(Nc1ccc(Oc2ccccc2)cc1)c3cccnc3NCc4ccncc4
wee1	CN(C)CCCOc1cc2c(cc1O)c3c(cc(c4ccccc4Cl)c5C(=O)NC(=O)c53)[n]2C

2. Recall Plots for Kinase Ligands

Recall plots for kinase ligands from decoys. The plots cover 26 kinase targets and three different fingerprints – orange lines for DISTRIB_FP_300, blue for FRAG_FP_1024 and green for CACTVS_881. DISTRIB_FP_300 were computed by averaging fragment vectors obtained from skip-gram architecture.




igf1r
jak2
kit
kpcb
lck
mapk2
met
mk01
mk10
% Actives Found
% Database Screened
0
20
40
60
80
100


mk14
mp2k1
plk1
rock1
src
tgfr1
vgfr2
wee1
% Actives Found
% Database Screened
0
20
40
60
80
100

3. SMILES of chemicals not covered by the distributed vectors

These chemicals are from different datasets employed in the binary classification and the QSAR section of the manuscript.

Mutagenicity dataset	**Anti-HIV dataset**	**Tetrahymena Dataset**	**LogP dataset**
[C-]#[O+]	[Li]F	CO	C=C
C#N	[Li]Cl		CBr
ClCl			C#C
C=O			CF
CBr			CC
C=C			CI
CCl			CO
CN			
CO			

4. Mutagenicity and anti-HIV Prediction performance as a function of different threshold values (in contrast to p-values)

4.1 Performance of the three fingerprints in 10-fold cross-validation exercise for predicting mutagenicity using k-nearest neighborhood method.

Similarity Threshold ->	**0.0**	**0.6**	**0.7**	**0.8**	**0.9**
	p-Value, Sensitivity%, Specificity%, Coverage%				
DISTRIB_FP_300*	1.000, 80, 74, 100	0.845, 80, 74, 100	0.676, 80, 74, 98	0.373, 81, 75, 96	0.053, 84, 75, 86
FRAG_FP_1024	1.000, 79, 73, 100	0.000, 86, 74, 71	0.000, 88, 76, 56	0.000, 90, 76, 37	0.000, 90, 80, 16
CACTVS_881	1.000, 82, 74, 100	0.006, 82, 75, 98	0.001, 84, 75, 91	0.000, 87, 76, 78	0.000, 89, 77, 52

*DISTRIB_FP_300 were computed by averaging fragment vectors obtained from skip-gram architecture.

4.2 Performance of the three fingerprints in 10-fold cross-validation exercise for predicting anti-HIV using k-nearest neighborhood method.

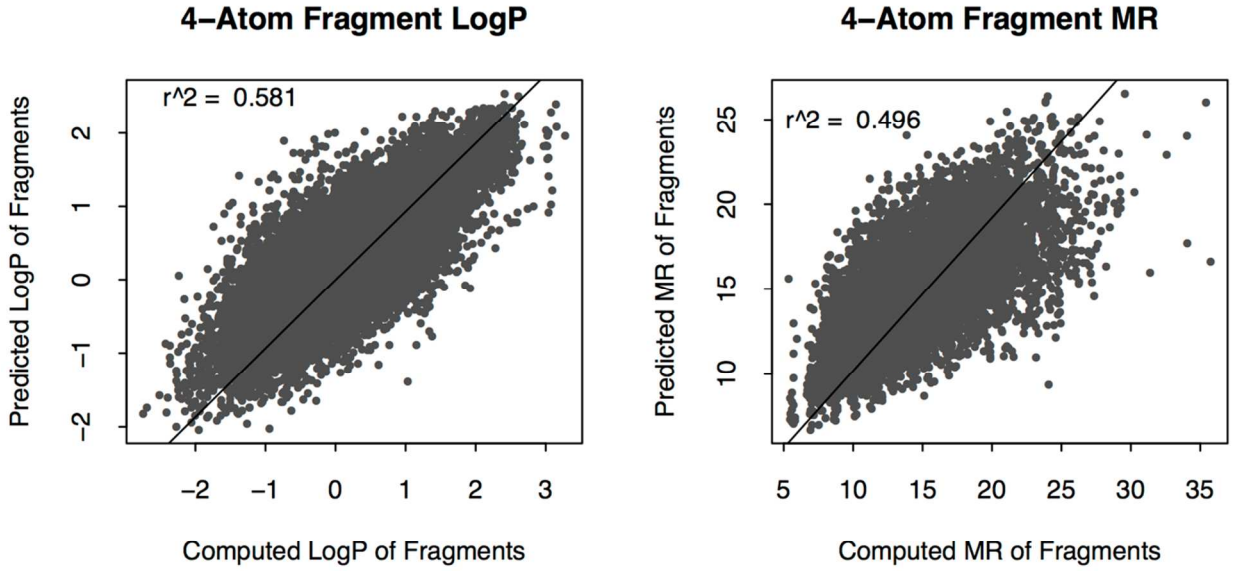
Similarity Threshold ->	**0.0**	**0.6**	**0.7**	**0.8**	**0.9**
	p-Value, Sensitivity%, Specificity%, Coverage%				

DISTRIB_FP_300*	1.000, 57, 91, 100	0.845, 57, 91, 100	0.676, 57, 91, 100	0.373, 58, 91, 98	0.053, 64, 90, 81
FRAG_FP_1024	1.000, 60, 92, 100	0.000, 83, 86, 44	0.000, 86, 86, 33	0.000, 87, 86, 21	0.000, 89, 88, 8
CACTVS_881	1.000, 58, 92, 100	0.006, 60, 91, 92	0.001, 71, 88, 68	0.000, 79, 85, 43	0.000, 86, 84, 21

*DISTRIB_FP_300 were computed by averaging fragment vectors obtained from skip-gram architecture.

5. Results obtained by using fragment vectors obtained from Continuous bag-of-words (CBOW) architecture

5.1 Using CBOW produced vectors for prediction of LogP and MR contribution (vertical axis) of 4-atom linear fragments using 5 closest fragments in the high dimensional vector space.



5.2 Percentage of recalled kinase ligands using distributed fingerprints computed by averaging fragment vectors obtained from CBOW architecture.

Kinase target	# Ligands	# Decoys	Distributed FP (CBOW)		
			$PR_{1\%}$	$PR_{5\%}$	$PR_{10\%}$
abl1	182	10749	5.5	19.2	32.4
akt1	293	16439	15.4	68.6	82.3
akt2	117	6900	29.1	65.8	77.8
braf	152	9950	9.2	32.9	47.4
cdk2	474	27846	4.0	11.4	20.5

csf1r	166	12149	15.1	31.9	47.6
egfr	542	35049	45.6	77.5	82.5
fak1	100	5350	34.0	69.0	91.0
fgfr1	139	8700	7.2	25.2	41.7
igf1r	148	9300	17.6	49.3	68.9
jak2	107	6498	20.6	43.9	61.7
kit	166	10450	3.0	6.6	12.7
kpcb	135	8697	61.5	68.1	71.1
lck	420	27396	24.5	47.4	61.0
mapk2	101	6145	13.9	30.7	34.7
met	166	11250	28.3	56.0	64.5
mk01	79	4547	49.4	70.9	75.9
mk10	104	6600	2.9	6.7	6.7
mk14	578	35850	8.8	36.5	53.3
mp2k1	121	8148	21.5	28.1	33.9
plk1	107	6800	3.7	11.2	24.3
rock1	100	6300	0.0	0.0	0.0
src	524	34494	21.4	46.0	59.7
tgfr1	133	8500	45.1	82.7	94.0
vgfr2	409	24949	11.7	30.8	41.3
wee1	102	6149	61.8	90.2	90.2

5.3. Performance of distributed fingerprints obtained from CBOW in 10-fold cross-validation exercise for predicting mutagenicity using k-nearest neighborhood method.

Similarity Threshold ->	**0.0**	**0.6**	**0.7**	**0.8**	**0.9**
	Sensitivity%, Specificity%, Coverage%				
Distributed vector FP, 300 elements (**CBOW**)	78, 73, 100	78, 73, 100	78, 73, 98	80, 74, 93	84, 73, 77

5.4. Performance of the fingerprints obtained from CBOW in 10-fold cross-validation exercise for predicting mutagenicity using k-nearest neighborhood method.

Similarity Threshold ->	**0.0**	**0.6**	**0.7**	**0.8**	**0.9**

	Sensitivity%, Specificity%, Coverage%				
Distributed vector FP, 300 elements (**CBOW**)	56, 92, 100	57, 92, 100	58, 92, 97	62, 91, 85	77, 88, 53

5.5 Observed vs predicted plots for toxicity against *T. pyriformis* and LogP using QSARs built using distributed fingerprints computed using CBOW architecture.