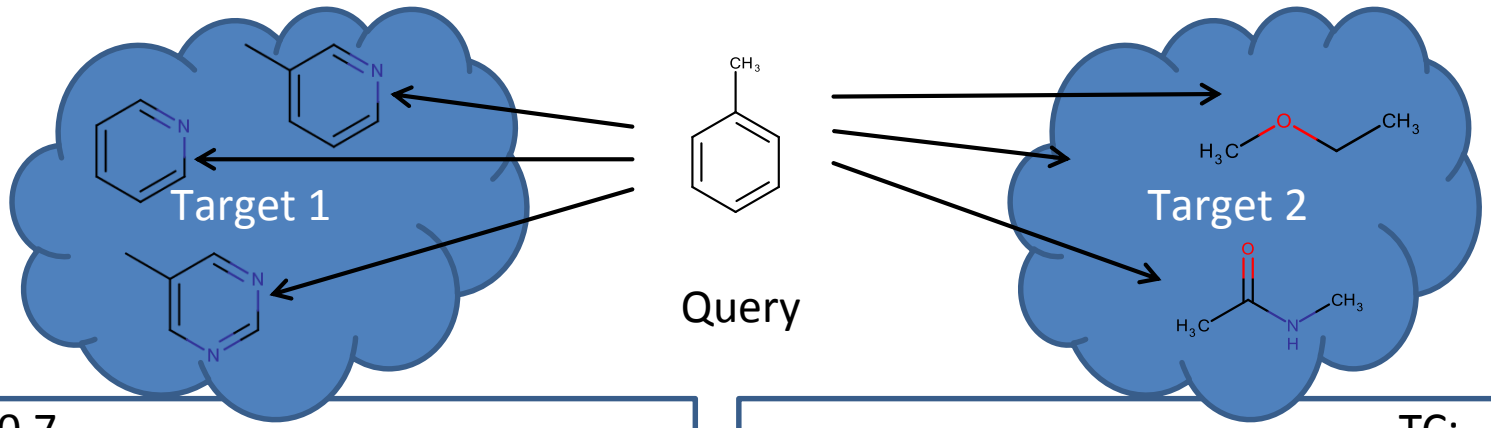
OCEAN
Optimized Cross rEActivity estimatioN

Wolf-Guido Bolick, Paul Czodrowski / RDKit UserGroupMeeting

OCEAN?

Re-Implentation of SEA (similarity ensemble approach) by Keiser et al.



TC:	0.7			
	0.6			T1:
	0.4	z-Score	p-Value	e-Value
Σ	1.7 →	2.5 →	0.01 →	0.02

				TC: 0.6
T2:				0.4
e-Value	p-Value	z-Score		0.3
0.22 ←	0.11 ←	1.2 ←		Σ 1.3

How to get to the e-value

$$Score_{Raw} = \sum_{i=0}^{n} TC[i] > Threshold \quad (1)$$

$$Score_{Z} = \frac{\mu(Score_{Raw}) - Score_{Raw,expected}}{\sigma(Score_{Raw,expected})} \quad (2)$$

$$x = -exp(\frac{-Score_{Z} * \pi}{\sqrt{6} - 0{,}57721}) \quad (3)$$

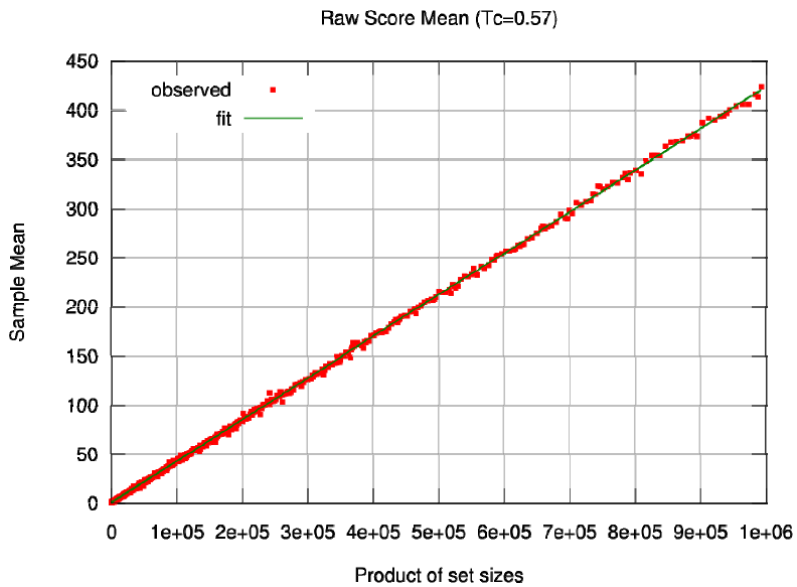
$$Value_{P} = -exp(\frac{x + x^2}{2} + \frac{x^3}{6}) \quad (4)$$

$$Value_{E} = Value_{P} * N_{Targets} \quad (5)$$

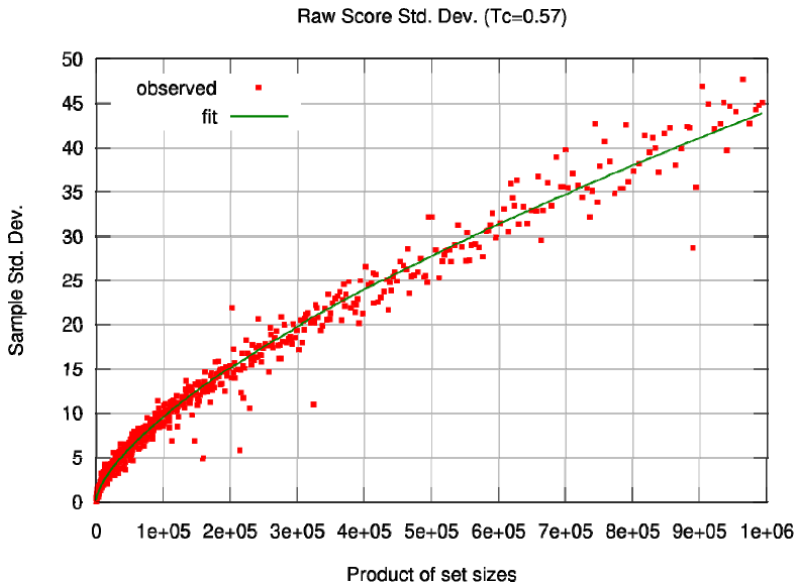
$$Score_{Raw} = \sum_{i=0}^{n} TC[i] > Threshold \qquad (1)$$

Background of SEA

Determination of the expected mean average by means of large-scale random molecular comparisons



Standard deviations of the molecular comparison



Figures taken from: http://dx.doi.org/10.1038/nbt1284

$$Score_Z = \frac{\mu(Score_{Raw}) - Score_{Raw,expected}}{\sigma(Score_{Raw,expected})} \quad (2)$$

Background of SEA

Calculation of z-Scores of all comparisons

z-Score distribution (Tanimoto = 0.57)

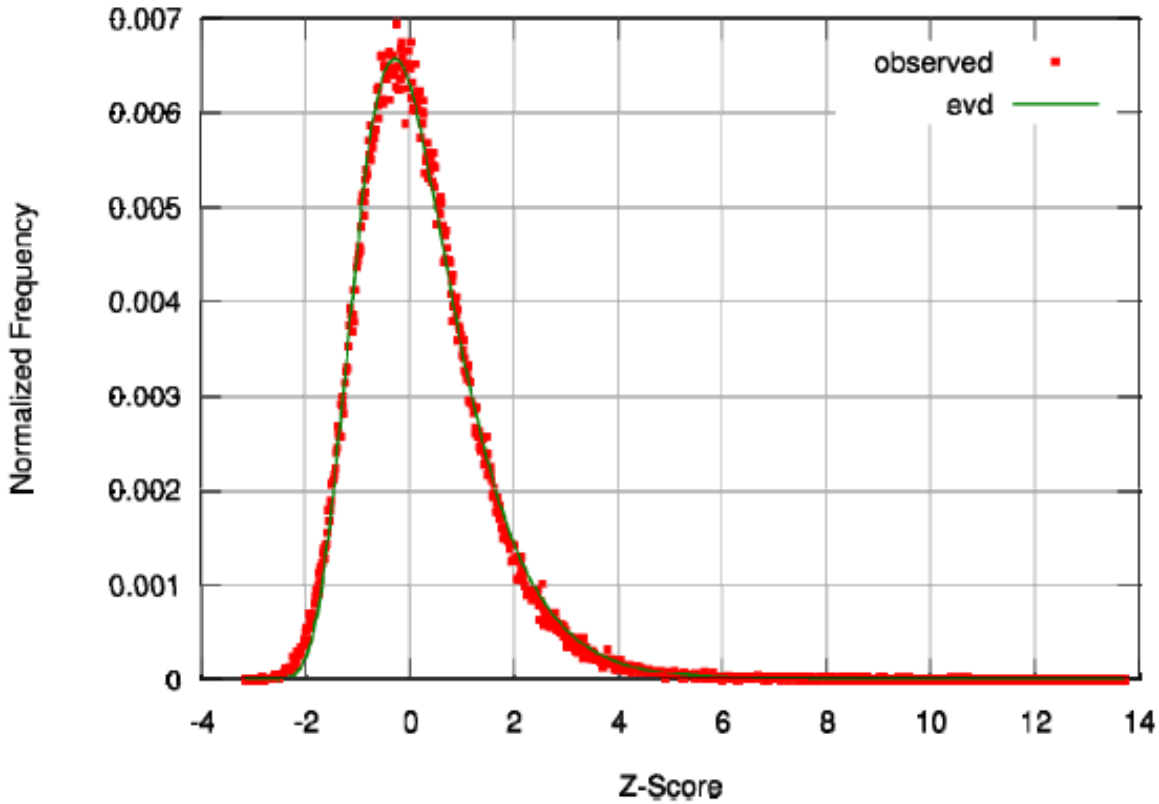


Figure taken from: http://dx.doi.org/10.1038/nbt1284

Choice of the optimal Tanimoto coefficient

Fitting of normal and extreme value distribution of the z-scores

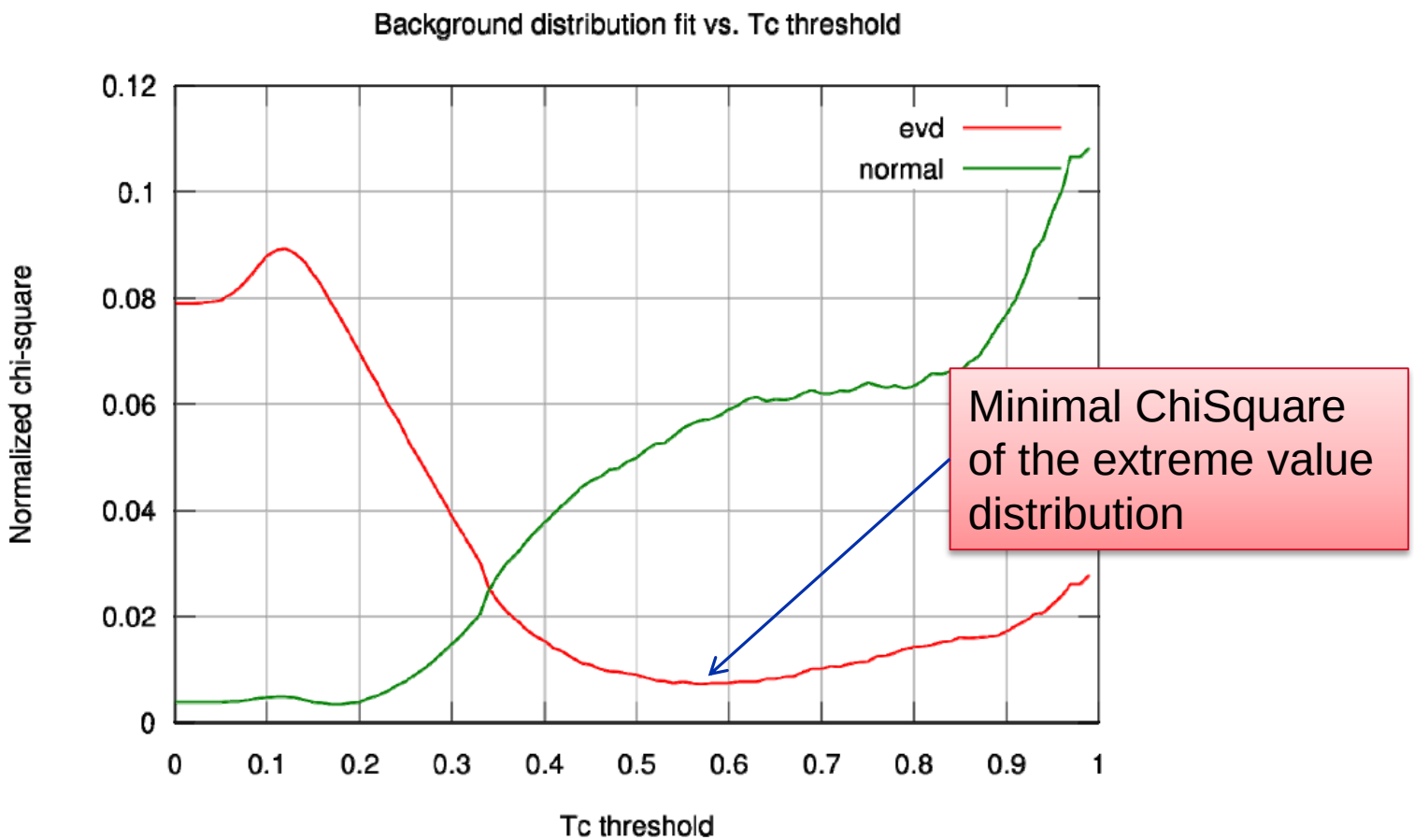
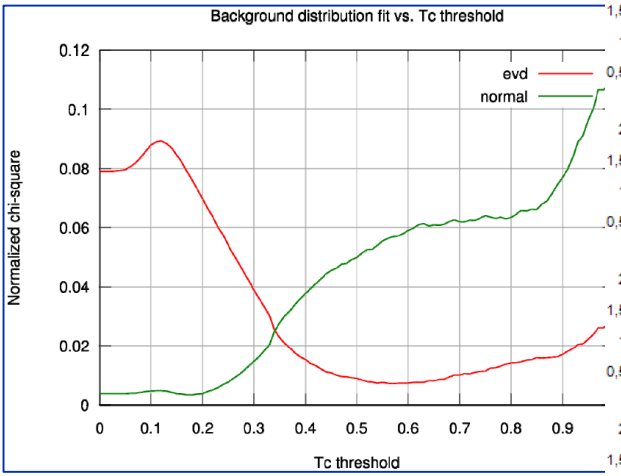


Figure taken from: http://dx.doi.org/10.1038/nbt1284

Optimal Tanimoto coefficient



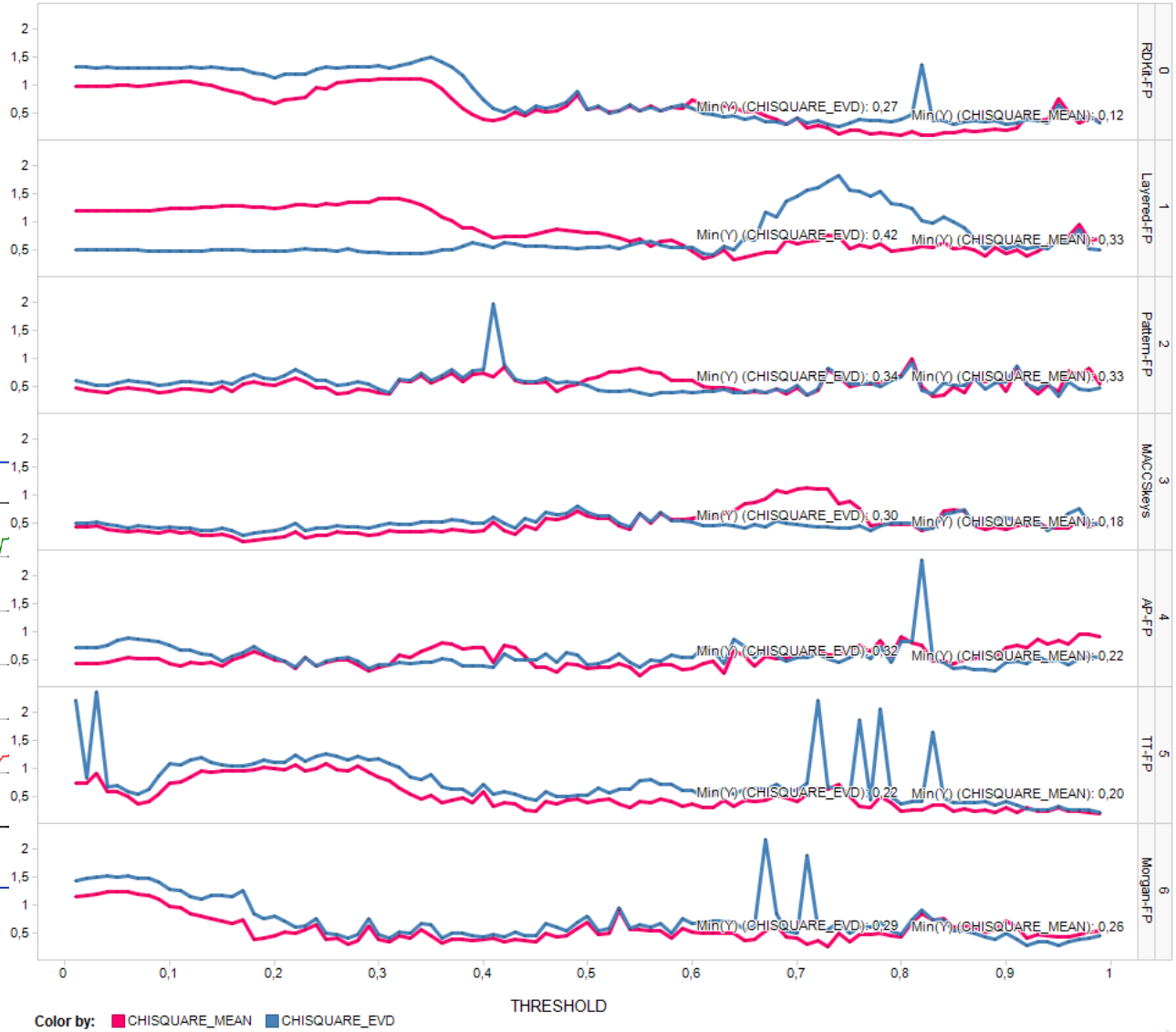
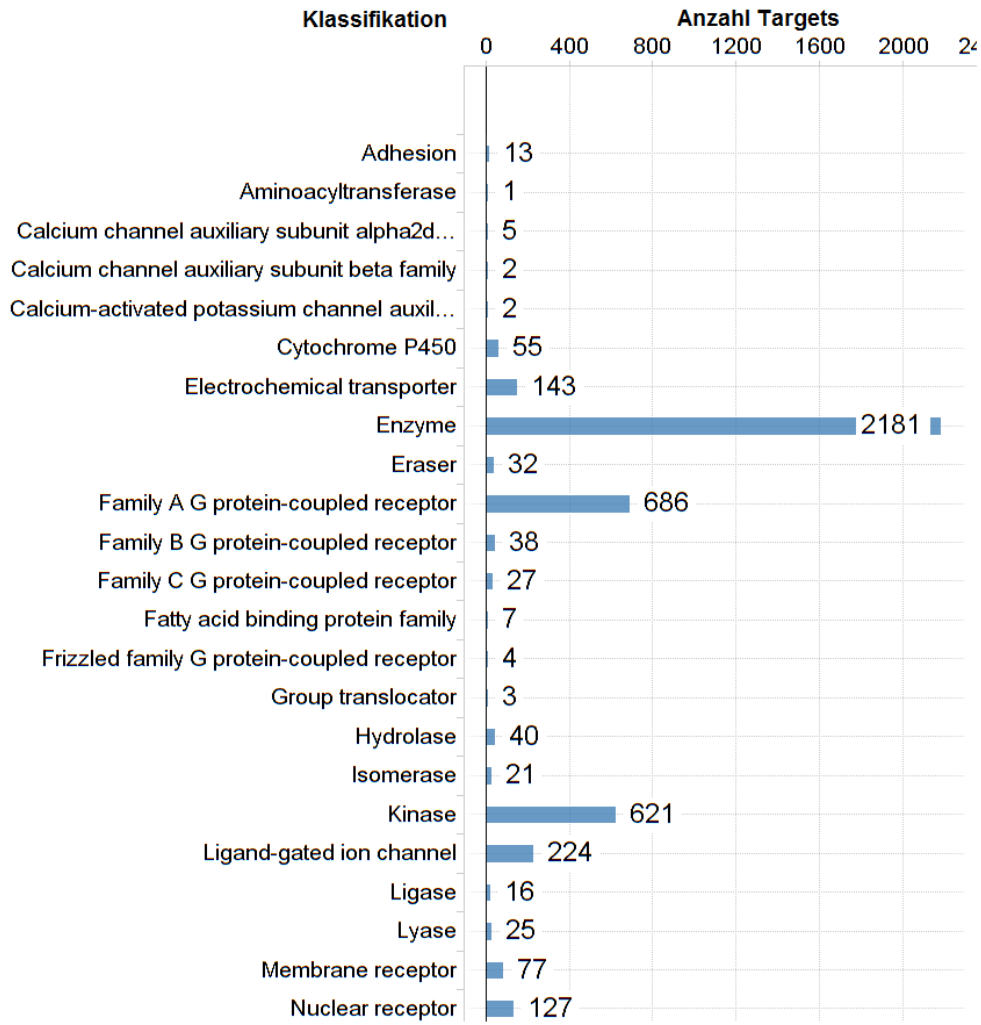


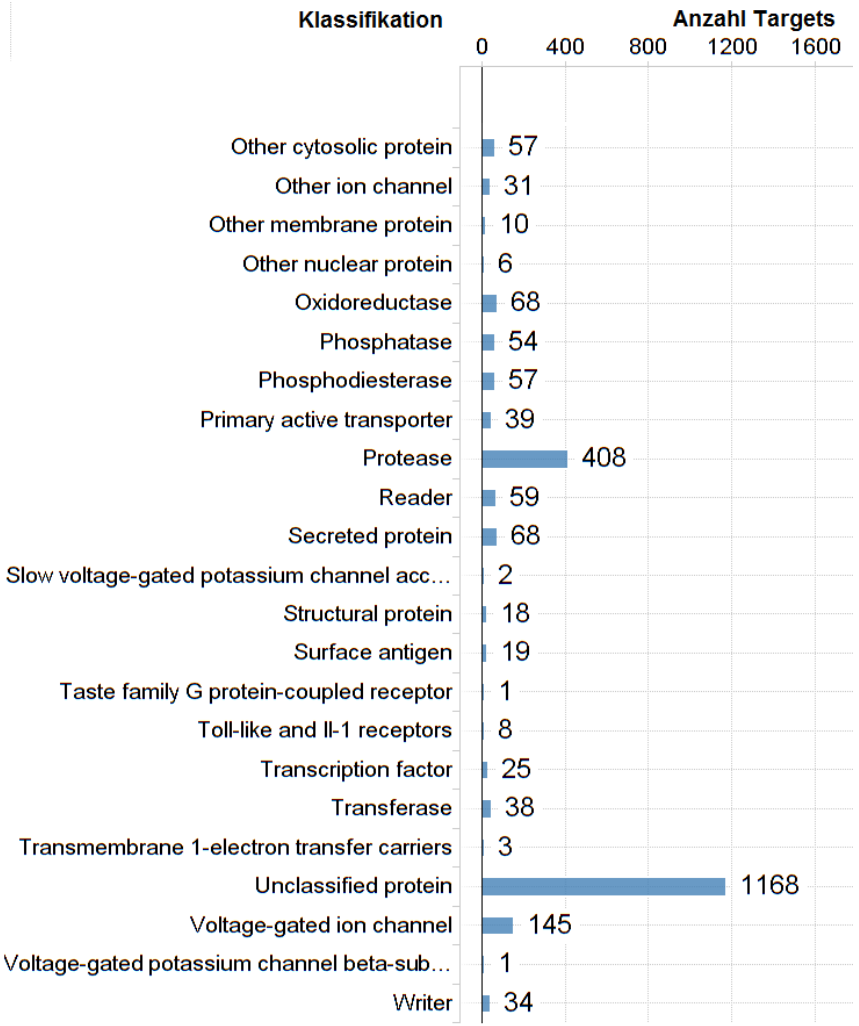
Figure taken from: http://dx.doi.org/10.1038/nbt1284

ChEMBL20 filtering

- IC_{50} and K_i
- Activity ≤10 μM
- ≥ 10 Compounds per Target
- Target Type: Single Protein or Protein Complex
- Homo sapiens
- If multiple measurements: only if single IC_{50} within mean ± 3 std_dev

Target Classes





Validation set

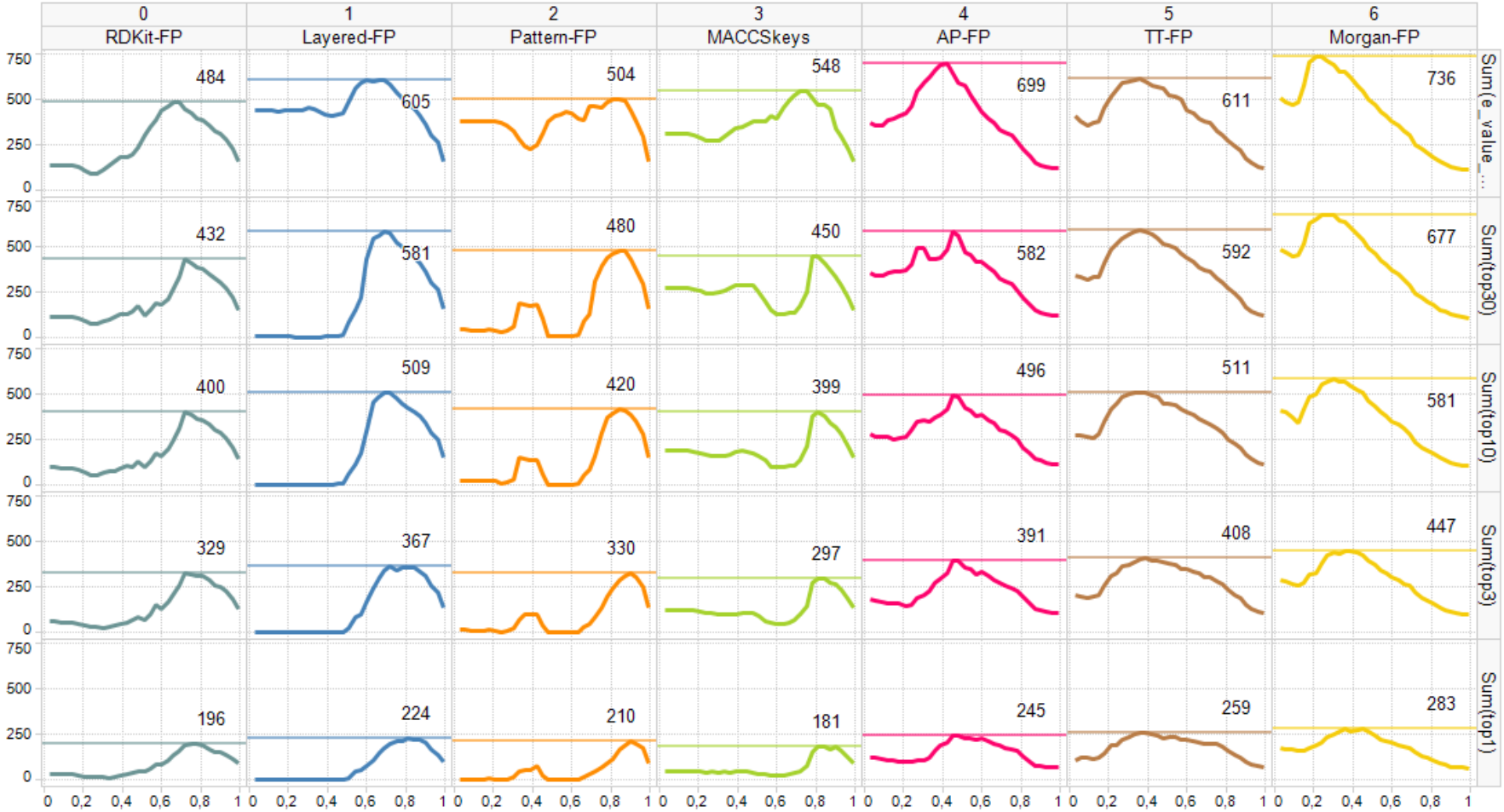
1,885 drugs on the market extracted from ChEMBL17

- For 928 drugs, no drug annotated in OCEAN-DB
  ➔ 957 drugs with annotated target in OCEAN-DB

How does OCEAN perform in terms of target prediction for this dataset?

[The drug itself was removed for the OCEAN search]

Tanimoto scan based on the validation set



Validation

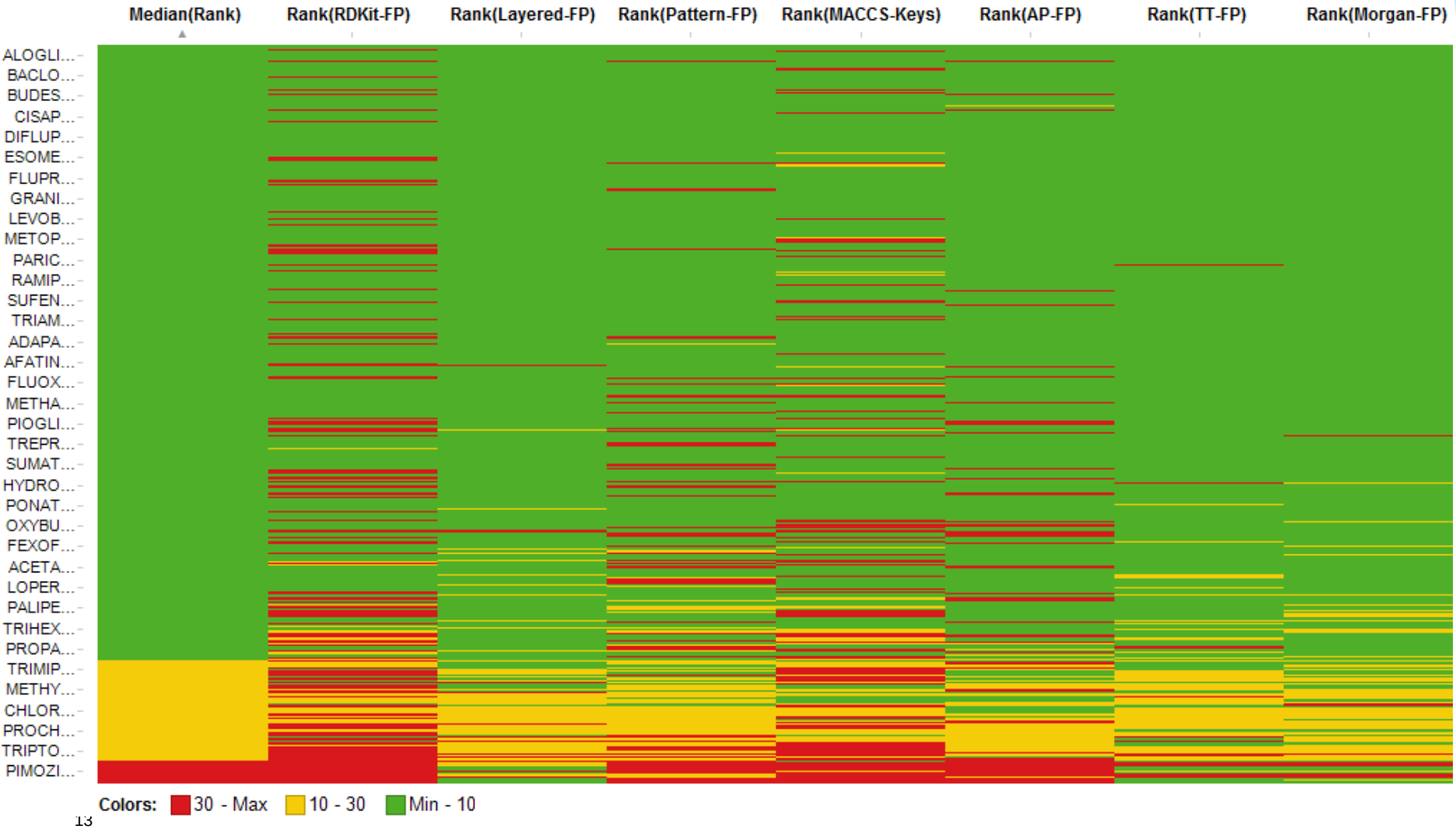
SEA

Fingerprint	Threshold	Sum(e_value<1)	Sum(Top30)	Sum(Top10)	Sum(Top3)	Sum(Top1)
RDKit-FP	0,75	427	412	385	318	196
Layered-FP	0,33	444	3	1	0	0
Pattern-FP	0,60	423	8	1	1	1
MACCSkeys	0,75	548	250	213	141	74
AP-FP	0,39	693	442	388	298	189
TT-FP	0,99	123	120	115	105	68
Morgan-FP	0,96	115	112	107	100	65

OCEAN

Fingerprint	Threshold	Sum(e_value<1)	Sum(Top30)	Sum(Top10)	Sum(Top3)	Sum(Top1)
RDKit-FP	0,72	448	432	400	329	187
Layered-FP	0,69	604	581	509	339	177
Pattern-FP	0,84	499	480	420	291	169
MACCSkeys	0,81	469	450	399	297	181
AP-FP	0,45	647	582	496	391	245
TT-FP	0,36	611	592	511	404	259
Morgan-FP	0,30	690	674	581	438	251

Validation for the drugs



Validation for the target classes

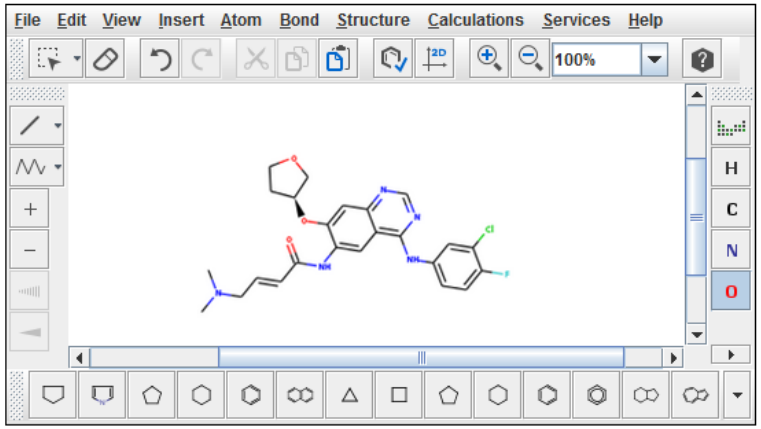
Front-End

Optimized Cross rEActivity estimatioN

neighbourhood-search for proteins
based on their compounds

smiles:

or sketch:



Start OCEAN-Search | Check Toxic-Targets

- ocean Home
- About ocean

- © Wolf-Guido Bolick, Paul Czodrowski - Global Computational Chemistry - 2015-02-26 (v 0.5)

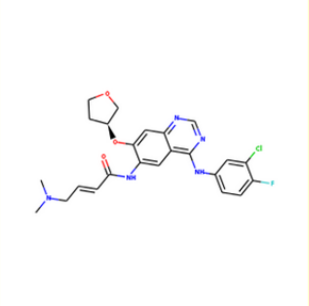
- Questions? Comments? Please send an email to Paul.Czodrowski@merckgroup.com or wolf-guido.bolick@external.merckgroup.com

Front-End: Output

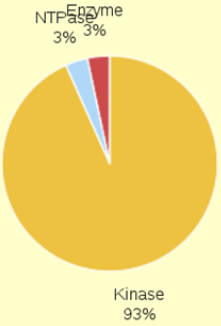
	#	Target_ID	Target_Name	Classification	#Ligands	e-Value	mean-TC
+	0	CHEMBL203	Epidermal growth factor receptor erbB1	Kinase	2834	8.25e-218	0.448
+	1	CHEMBL1824	Receptor protein-tyrosine kinase erbB-2	Kinase	1178	2.60e-184	0.464
+	2	CHEMBL3009	Receptor protein-tyrosine kinase erbB-4	Kinase	131	6.18e-64	0.461
+	3	CHEMBL279	Vascular endothelial growth factor receptor 2	Kinase	3953	1.79e-52	0.400
+	4	CHEMBL2634	Tyrosine-protein kinase CSK	Kinase	38	8.19e-43	0.354
+	5	CHEMBL1868	Vascular endothelial growth factor receptor 1	Kinase	858	1.98e-41	0.388
+	6	CHEMBL3650	Fibroblast growth factor receptor 1	Kinase	870	3.07e-30	0.403
+	7	CHEMBL3935	Serine/threonine-protein kinase Aurora-C	Kinase	23	1.03e-23	0.431
+	8	CHEMBL267	Tyrosine-protein kinase SRC	Kinase	1908	6.25e-23	0.402
+	9	CHEMBL5251	Tyrosine-protein kinase BTK	Kinase	216	1.04e-20	0.424
+	10	CHEMBL4722	Serine/threonine-protein kinase Aurora-A	Kinase	1497	1.62e-20	0.400
+	11	CHEMBL3290	Ephrin type-B receptor 2	Kinase	12	2.31e-17	0.410
+	12	CHEMBL1913	Platelet-derived growth factor receptor beta	Kinase	677	2.79e-17	0.440
+	13	CHEMBL4142	Fibroblast growth factor receptor 2	Kinase	49	4.60e-17	0.407
+	14	CHEMBL1955	Vascular endothelial growth factor receptor 3	Kinase	508	1.06e-16	0.392
+	15	CHEMBL2073	Tyrosine-protein kinase YES	Kinase	44	3.59e-15	0.430
+	16	CHEMBL5699	Serine/threonine-protein kinase SIK2	Kinase	20	1.21e-13	0.401
+	17	CHEMBL3975	Fructose-1,6-bisphosphatase	Enzyme	288	1.55e-11	0.374
+	18	CHEMBL2185	Serine/threonine-protein kinase Aurora-B	Kinase	1329	1.57e-10	0.390
+	19	CHEMBL2250	Tyrosine-protein kinase BLK	Kinase	230	6.46e-10	0.392
+	20	CHEMBL2007	Platelet-derived growth factor receptor alpha	Kinase	302	9.28e-09	0.389
+	21	CHEMBL2041	Tyrosine-protein kinase receptor RET	Kinase	452	4.06e-08	0.388
+	22	CHEMBL3905	Tyrosine-protein kinase Lyn	Kinase	369	3.79e-07	0.415

Query:

Smiles: **CN(C)C\C=C\C(=O)Nc1cc2c(Nc3ccc(F)c(Cl)c3)ncnc2cc1O[C@H]4CCOC4.OC(=O)\C=C/C(=O)O.OC(=O)\C=C/C(=O)O**

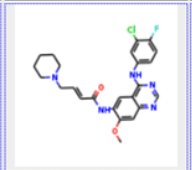
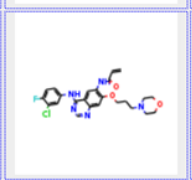
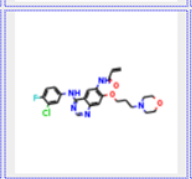
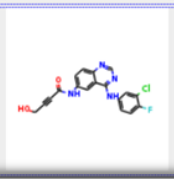


Class-Distribution (first 30)



Front-End: Output

	#	Target_ID	Target_Name	Classification	#Ligands	e-Value	mean-TC	threshold	activ
+	0	CHEMBL203	Epidermal growth factor receptor erbB1	Kinase	2834	8.25e-218	0.448		
+	1	CHEMBL1824	Receptor protein-tyrosine kinase erbB-2	Kinase	1178	2.60e-184	0.464		
+	2	CHEMBL3009	Receptor protein-tyrosine kinase erbB-4	Kinase	131	6.18e-64	0.461		

#	molregno	Molecule ID	Smiles	Molecule	Standard Value	Tanimoto Distance to Query ▲
87	1377989	CHEMBL2105719	O.COc1cc2ncnc(Nc3ccc(F)c(Cl)c3)c2cc1NC(=O)\C=C\CN4CCCCC4 ↻ Kinase: 87 % NTPase: 3 % Enzyme: 3 % Hydrolase: 3 %		74	0.890681003584
69	41764	CHEMBL31965	Fc1ccc(Nc2ncnc3cc(OCCCN4CCOCC4)c(NC(=O)C=C)cc23)cc1Cl ↻ Kinase: 90 % Enzyme: 3 % Phosphodiesterase: 3 % 7TM1: 3 %		12	0.890070921986
17	547496	CHEMBL545315	Cl.Cl.Fc1ccc(Nc2ncnc3cc(OCCCN4CCOCC4)c(NC(=O)C=C)cc23)cc1Cl ↻ Get Hit Profile		31	0.890070921986
103	338755	CHEMBL202411	OCC#CC(=O)Nc1ccc2ncnc(Nc3ccc(F)c(Cl)c3)c2c1 ↻ Get Hit Profile		7.3	0.672855879752

Next steps

- Still looking for publically available validation data not contained in ChEMBL
- Validation on internal data
- PCA plot (how well is the query compound represented by the target data)