Enrichments and Rescoring

Trent Balius

AMS 535 / CHE 535

Enrichments and Rescoring

Trent Balius

AMS 535 / CHE 535

Directory of Useful Decoys

Benchmarking Sets for Molecular Docking

Niu Huang, Brian K. Shoichet, and John J. Irwin

J. Med. Chem., 2006, 49 (23), 6789-6801

Outline

- Introduction
  - Docking Introduction
  - Docking Validations
  - Enrichment
- DUD Background
- DUD Enrichments
- DUD Cross-Enrichments
- Binding pose predictions
- Conclusions

Introduction

Docking Introduction

- Objectives of Docking programs
  - generate binding modes (or poses)
  - select the true pose out all poses generated with a scoring function
- Uses of Docking programs
  - pose reproduction

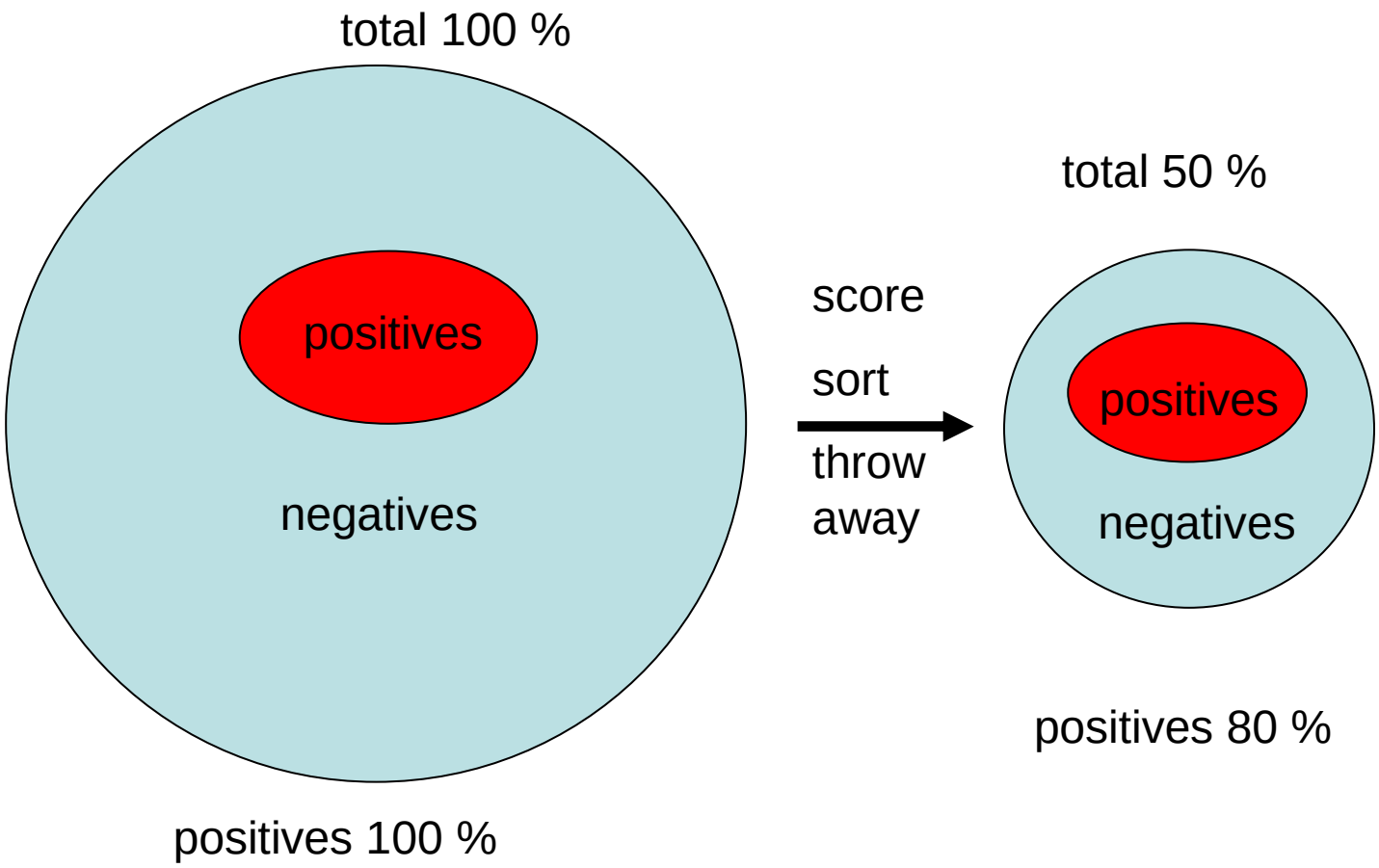
    (pdb of receptor but not of complex)
  - virtual screening

    find a new drug lead by screening virtual databank (e.g. ZINC)

Docking Validations Studies

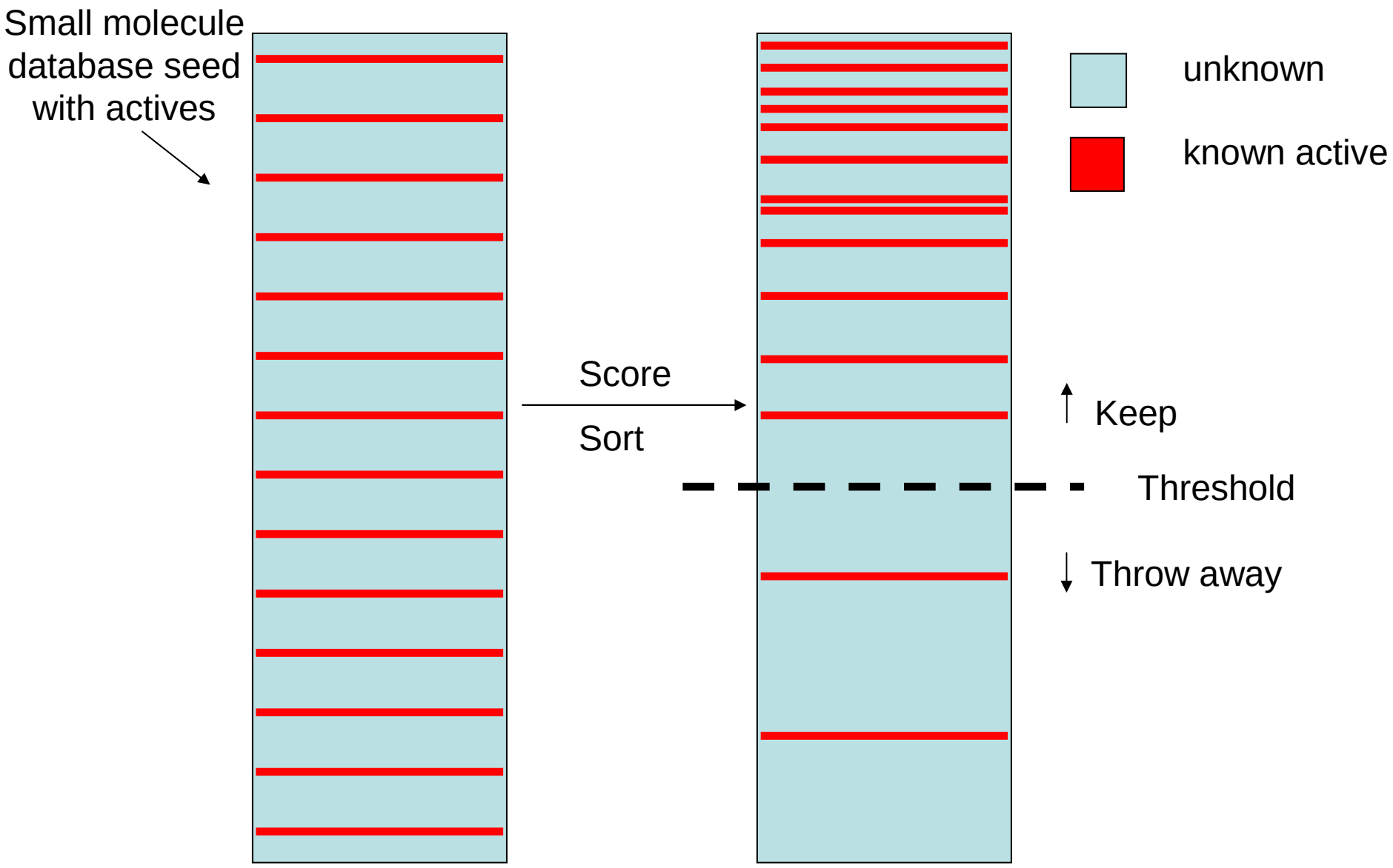
- Pose reproduction:
  - regenerating the known binding mode of a ligand in the context of the protein with a docking program
  - protein-ligand complex needed
- Enrichments:
  - after docking a database of known actives and decoys the actives are top scoring
  - protein structure needed

Enrichment


total 100 %
positives
negatives
positives 100 %
score
sort
throw
away
total 50 %
positives
negatives
positives 80 %

Enrichment Studies

unknowns may have activity



Enrichment Studies

- Active and inactive is not known
  - Why not run an assay on all the small molecules?
    - expensive
    - takes time
    - multiple levels of experiments (needs to compare several assays. e.g. HTPS)
  - Seed the population with known active compounds
  - See how many bubble to the top.
- Enrichment curves
- Receiver operating characteristic (ROC) curves

DUD Background

Overview of DUD

- Directory of Useful Decoys (DUD)
  - used for enrichment studies
  - # of systems = 40 targets (proteins)
  - # of ligands = 2,950 molecules (actives)
  - # of decoys = 98,266 (presumed non-binders)
  - every active molecule has 36 decoys
    - 36*2950 = 106,200 ≠ 98,266 because there are some decoys shared among ligands.
    - decoys are physically similar
    - topologically distinct

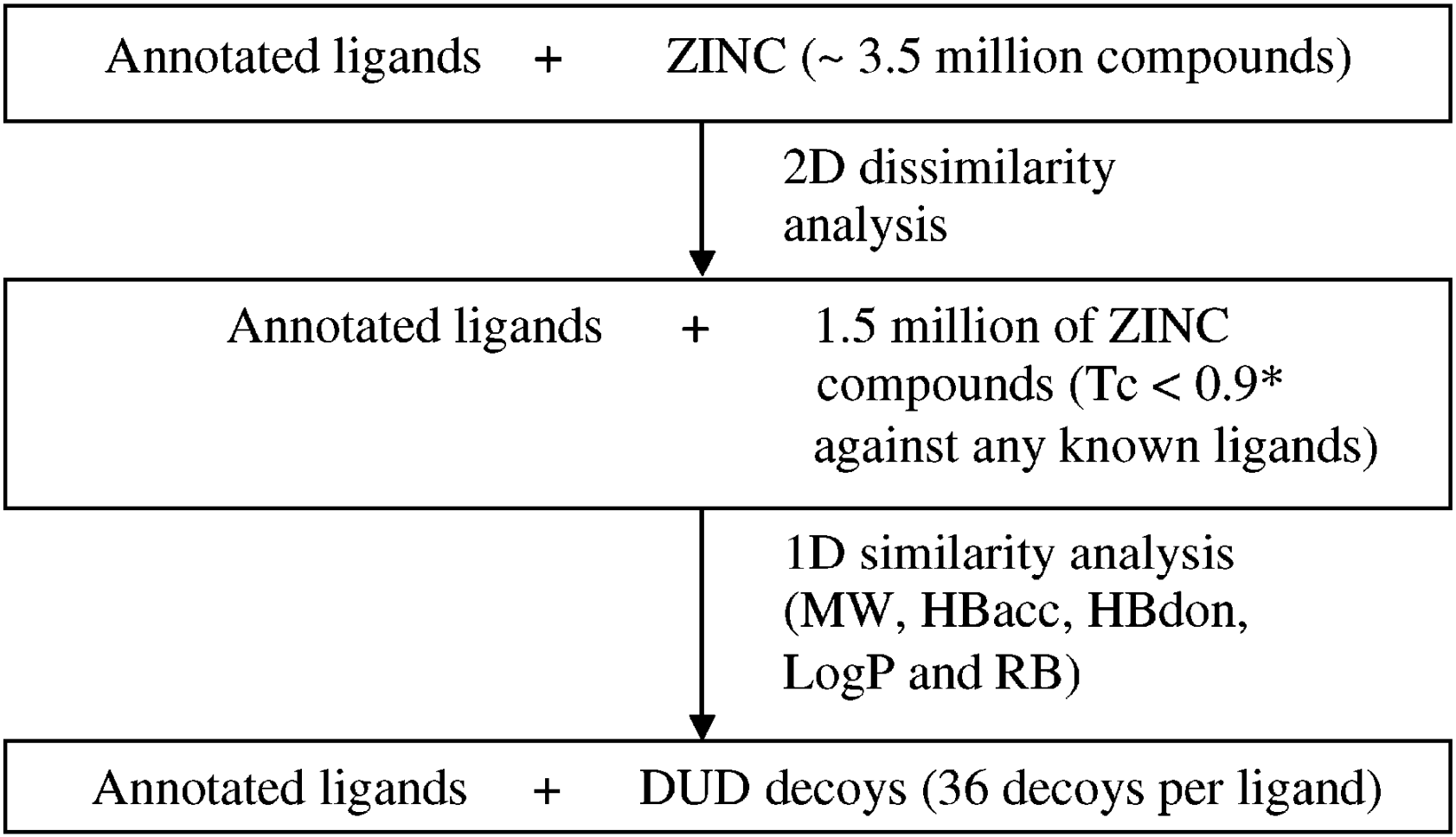
J. Med. Chem. **2006,** *49,* 6789-6801

Overview of DUD (cont'd)

- Directory of Useful Decoys (DUD)
  - Systems chosen for the following reasons:
    availability of annotated ligands
    crystal structures
    previous docking studies
  - Designed to remove sorting bias on "gross features"
  - Decoys are "chemically distinct" from active ligand "unlikely binders"

J. Med. Chem. **2006,** *49,* 6789-6801

Protocol for DUD prep



J. Med. Chem. **2006,** *49,* 6789-6801

Tanimoto Coefficient

$$Tc = \frac{|A \cap B|}{|A \cup B|}$$

- intersection is # of ON bits common in both A and B
- union is # of ON bits present in either A or B
- Examples of Daylight Fingerprint descriptors:
  - ring systems
  - common functional groups
  - which elements are present
  - unusual electronic configurations.

DUD systems

Table 1. Enrichments of the Annotated Ligands Using the Decoys in DUD for Forty Targets by Docking[a]

	protein	PDB code	resolution (Å)	no. of ligands[b]	no. of decoys	EF_{max}	EF_1	EF_{20}
				Nuclear Hormone Receptors				
1.	AR	1xq2	1.9	74 (a,b)	2630	60.2	33.5	3.8
2.	$ER_{agonist}$	1l2i	1.9	67 (a−c)	2361	29.6	19.2	4.5
**3.	**$ER_{antagonist}$**	**3ert**	**1.9**	**39 (a−d)**	**1399**	**101.6**	**12.7**	**1.3**
4.	GR	1m2z	2.5	78 (a)	2804	31.7	8.9	1.4
5.	MR	2aa2	1.9	15 (a)	535	330.0	46.2	3.7
6.	PPARg	1fm9	2.1	81 (a)	2910	1.0	0.0	0.0
7.	PR	1sr7	1.9	27 (a)	967	2.9	0.0	2.0
8.	RXRa	1mvc	1.9	20 (a)	708	148.5	24.8	2.2
				Kinases				
9.	CDK2	1ckp	2.1	50 (e,f)	1780	19.8	13.9	1.4
10.	EGFr	1m17	2.6	416 (g)	14914	3.8	2.1	2.4
11.	FGFr1	1agw	2.4	118 (g)	4216	1.0	0.0	0.2
12.	HSP90	1uy6	1.9	24 (h)	861	10.8	8.6	2.0
**13.	**P38 MAP**	**1kv2**	**2.8**	**234 (g)**	**8399**	**4.1**	**2.1**	**2.4**
14.	PDGFrb	model	n/a	157 (g)	5625	1.2	0.0	0.6
15.	SRC	2src	1.5	162 (g)	5801	3.1	1.2	1.5
**16.	**TK**	**1kim**	**2.1**	**22 (a,d,i)**	**785**	**63.0**	**54.0**	**5.0**
17.	VEGFr2	1vr2	2.4	74 (j)	2647	2.2	1.3	1.4
				Serine Proteases				
18.	FXa	1f0r	2.7	142 (e,f,k)	5102	34.9	14.6	3.8
19.	thrombin	1ba8	1.8	65 (e,l,m)	2294	18.3	13.7	2.9
20.	trypsin	1bju	1.8	43 (e,l)	1545	22.5	22.5	2.6

J. Med. Chem. **2006,** *49,* 6789-6801

DUD systems (cont'd)

1NDW →

1Q4G →

	protein	PDB code	resolution (Å)	no. of ligands[b]	no. of decoys	EF_{max}	EF_1	EF_{20}
				Metalloenzymes				
21.	ACE	1o86	2.0	49 (a,m)	1728	141.4	40.4	3.7
**22.	**ADA**	**1stw**	**2.0**	**23 (a,e)**	**822**	**21.5**	**12.9**	**2.4**
23.	COMT	1h1d	2.0	12 (a)	430	11.8	0.0	3.3
24.	PDE5	1xp0	1.8	51 (f)	1810	29.1	11.8	2.3
				Folate Enzymes				
25.	DHFR	3dfr	1.7	201 (m)	7150	28.7	21.7	3.5
26.	GART	1c2t	2.1	21 (n)	753	70.7	42.4	3.3
				Other Enzymes				
27.	AChE	1eve	2.5	105 (a,e,m)	3732	3.1	1.9	2.0
**28.	**ALR2**	**1ah3**	**2.3**	**26 (o)**	**920**	**76.2**	**38.1**	**2.3**
29.	AmpC	1xgj	2.0	21 (p)	734	23.6	17.1	4.7
30.	COX-1	1p4g	2.1	25 (i)	850	9.9	4.0	1.6
31.	COX-2	1cx2	3.0	349 (c,f,m)	12491	29.1	20.1	3.3
32.	GPB	1a8i	1.8	52 (e,m)	1851	28.6	22.8	4.1
33.	HIVPR	1hpx	2.0	53 (a,e)	1888	9.3	3.7	2.2
34.	HIVRT	1rt1	2.6	40 (q)	1439	49.5	5.0	3.0
35.	HMGR	1hw8	2.1	35 (a,i)	1242	198.0	33.9	2.1
**36.	**InhA**	**1p44**	**2.7**	**85 (r)**	**3043**	**1.0**	**0.0**	**0.3**
37.	NA	1a4g	2.2	49 (c,e,i)	1745	60.6	20.2	3.3
38.	PARP	1efy	2.2	33 (s)	1178	6.3	6.0	3.6
39.	PNP	1b8o	1.5	25 (e,t)	884	158.4	31.7	4.4
40.	SAHH	1a7a	2.8	33 (i)	1159	120.0	78.0	5.0

J. Med. Chem. **2006,** *49,* 6789-6801

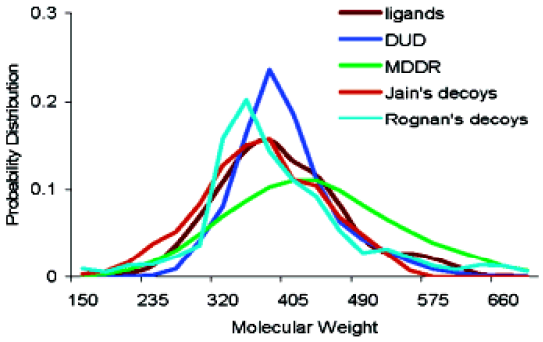
Six DUD systems

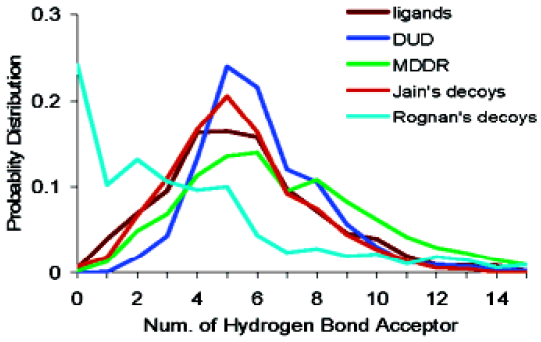
protein	PDB code	resolution (Å)	no. of ligands	no. of decoys	EFmax	EF1	EF20
ERantagonist	3ert	1.9	39	1399	101.6	12.7	1.3
P38 MAP	1kv2	2.8	234	8399	4.1	2.1	2.4
TK	1kim	2.1	22	785	63.0	54.0	5.0
ADA	1stw	2.0	23	822	21.5	12.9	2.4
ALR2	1ah3	2.3	26	920	76.2	38.1	2.3
InhA	1p44	2.7	85	3043	1.0	0.0	0.3

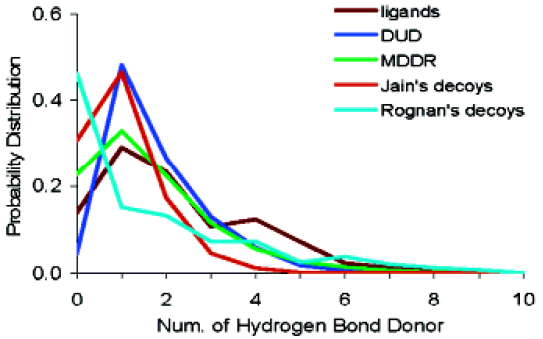
- Families chosen for the following reasons:
  - ER and TK -- strong ligand enrichment and substantial number of published docking studies
  - P38 MAP kinase -- poorly performing protein kinases
  - ADA -- failed with the fully automated docking engine and rescued by the semiautomated procedure
  - ALR2 -- intermediate enrichment.
  - InhA -- failure of the docking method

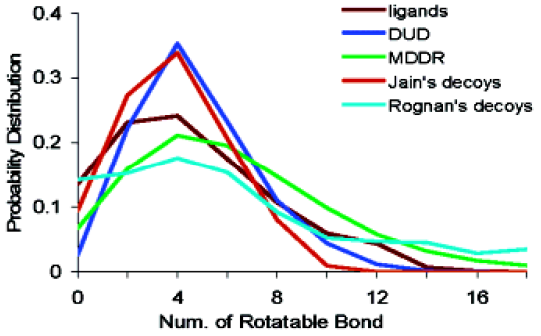
J. Med. Chem. **2006,** *49,* 6789-6801

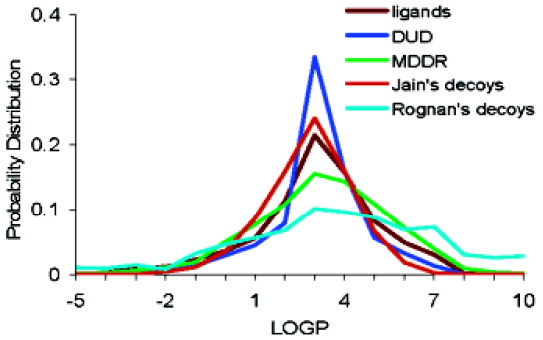
Molecule Properties







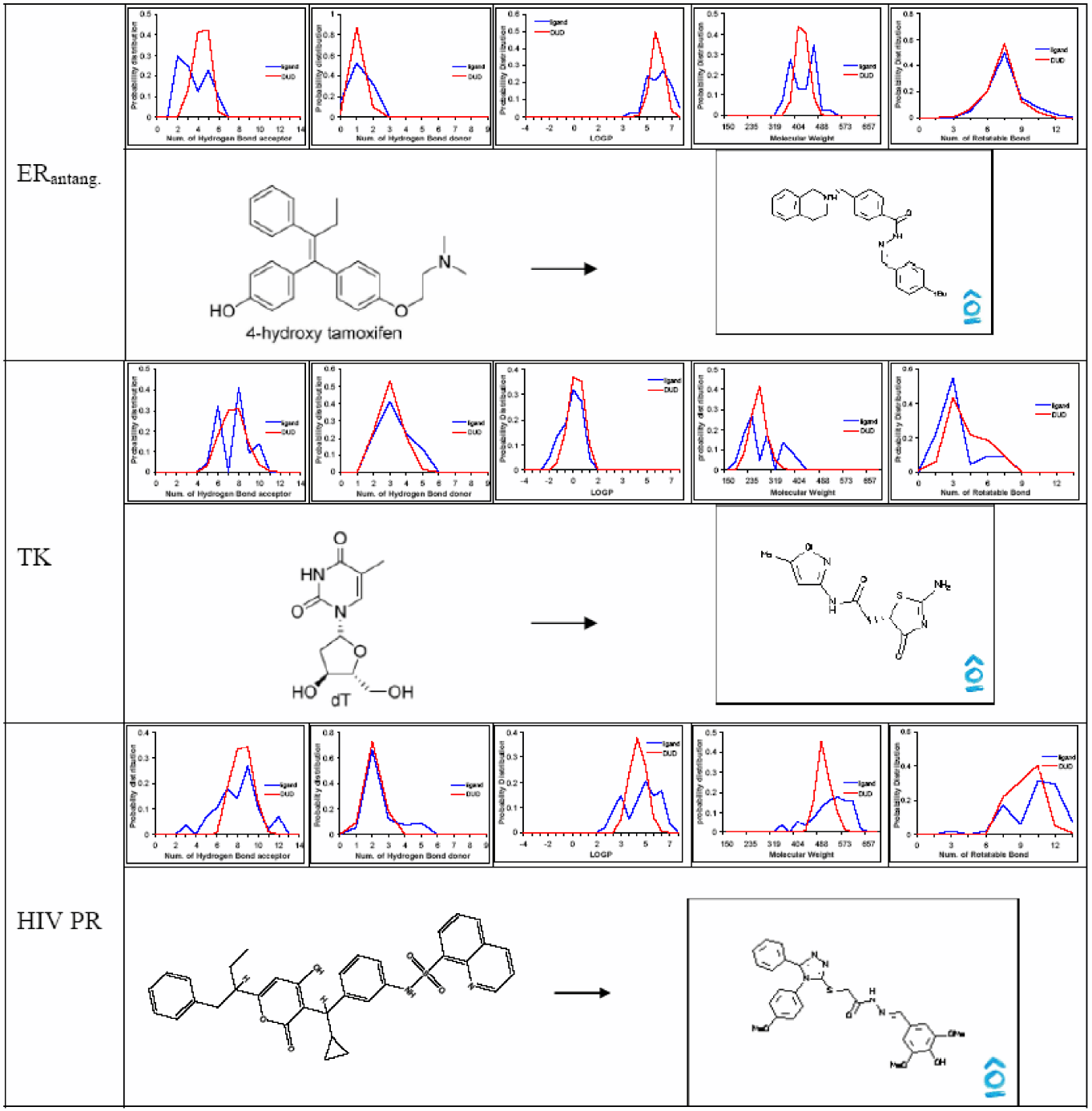




The physical property distributions

- brown -- annotated ligands (2950 compounds)
- blue -- the DUD decoys (95 316 compounds)
- green -- properties of the MDDR database (98 000 compounds)
- orange -- Jain's decoys (randomly selected 1000 ZINC druglike compounds)
- cyan -- Rognan's decoys (randomly selected 990 ACD compounds).

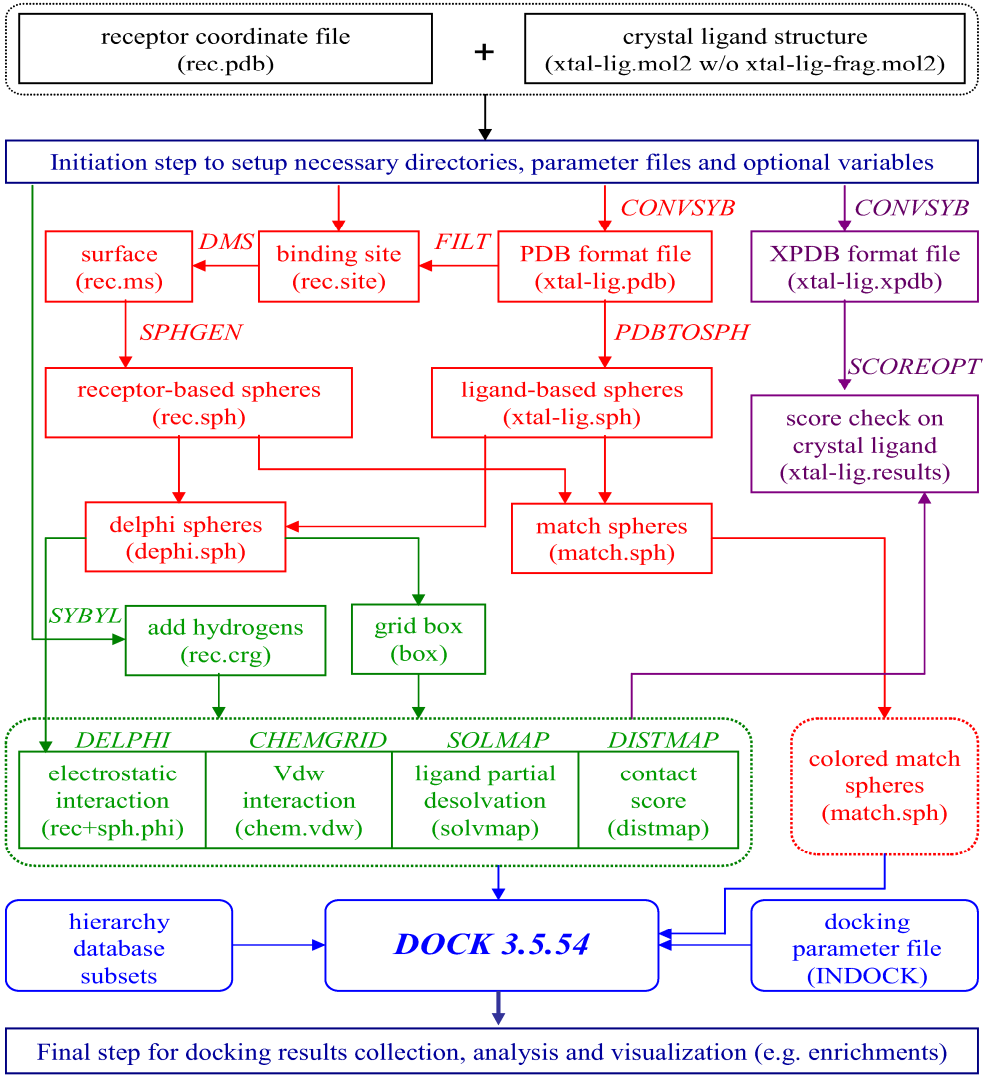
J. Med. Chem. **2006,** *49,* 6789-6801



- The physical property
  - # of HB acceptors
  - # of HB donors
  - xlogp
  - Molecular Weight
  - #of rotatable bonds

supporting material
J. Med. Chem. **2006,** *49,* 6789-6801

Automated Docking Pipeline



red - sphere generation,
green - scoring grids computation and scoring
purple - crystallographic ligand

supporting material
J. Med. Chem. **2006,** *49,* 6789-6801

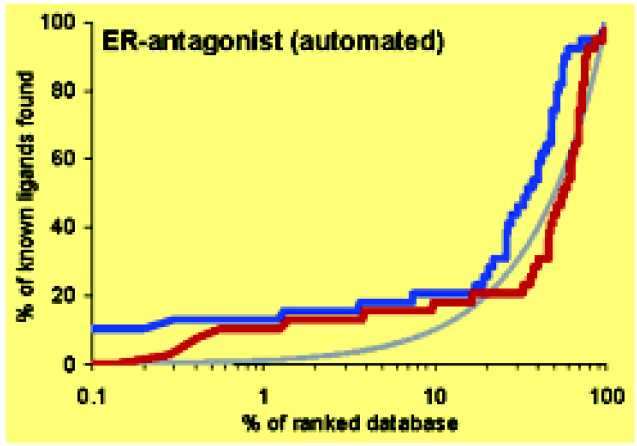
DUD Enrichments

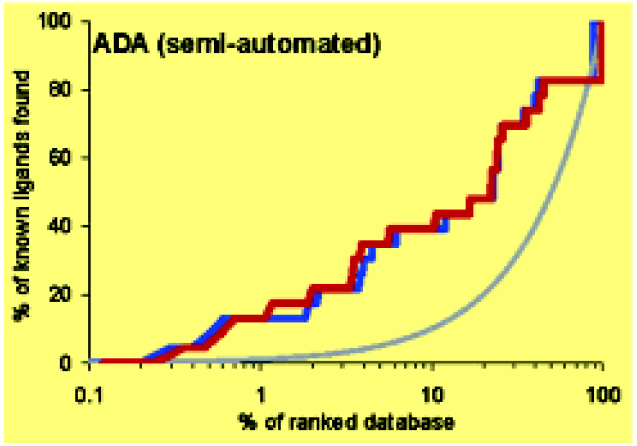
Enrichment Factors

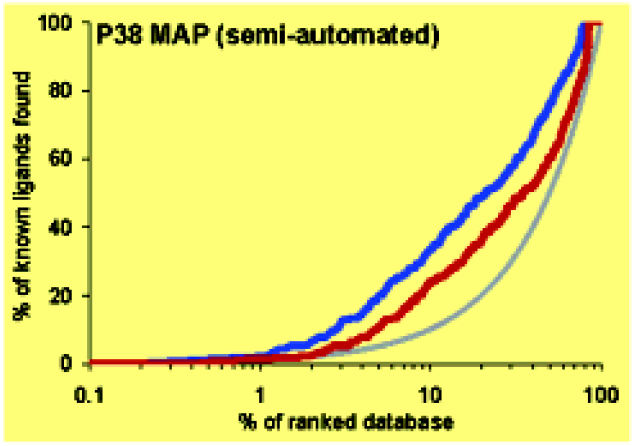
$$EF_{\text{subset}} = \frac{ligands_{\text{selected}} / N_{\text{subset}}}{ligands_{\text{total}} / N_{\text{total}}}$$

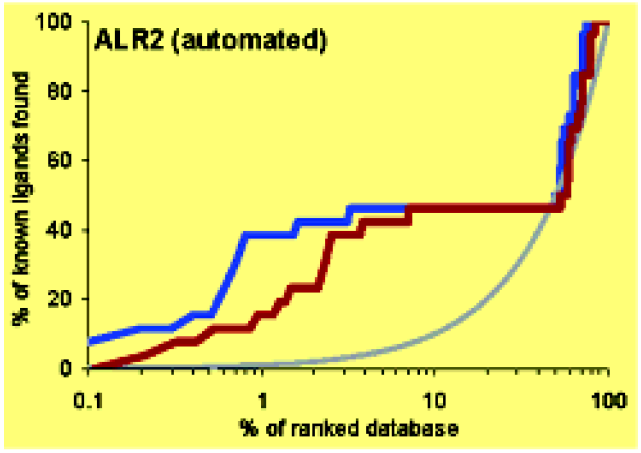
$$= \frac{ligands_{\text{selected}}}{ligands_{\text{total}}} \frac{N_{\text{total}}}{N_{\text{subset}}}$$

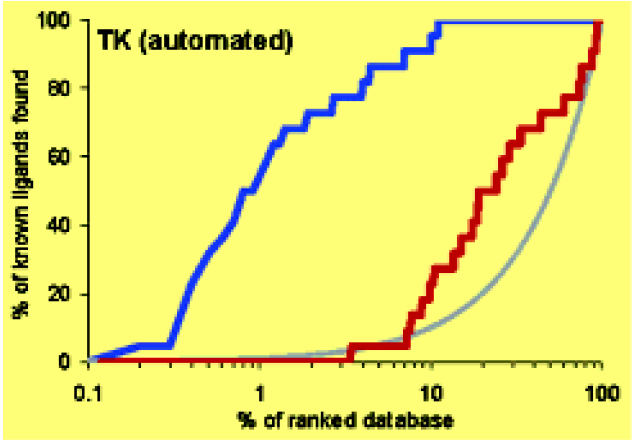
$$EF_1 = \frac{ligands_{\text{selected,top1\%database}}}{ligands_{\text{total}}} * \frac{100}{1} \quad EF_{20} = \frac{ligands_{\text{selected,top20\%database}}}{ligands_{\text{total}}} * \frac{100}{20}$$

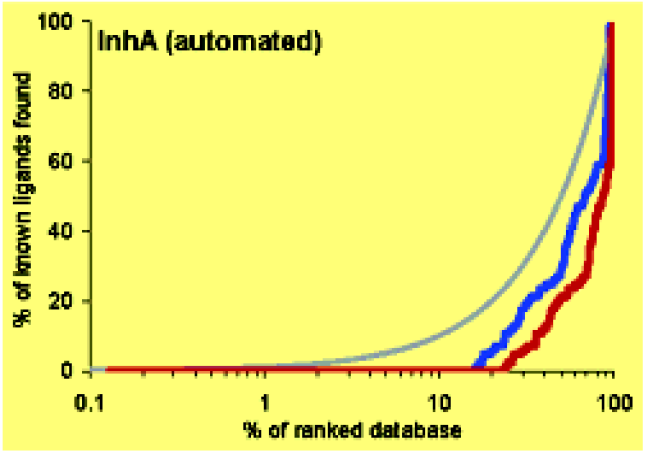












The docking ranked database

the percentage of known ligands found

six representative systems are highlighted in light yellow.

gray -- random

blue DUD database (98 266 compounds)

red target subset decoy

J. Med. Chem. **2006,** *49,* 6789-6801

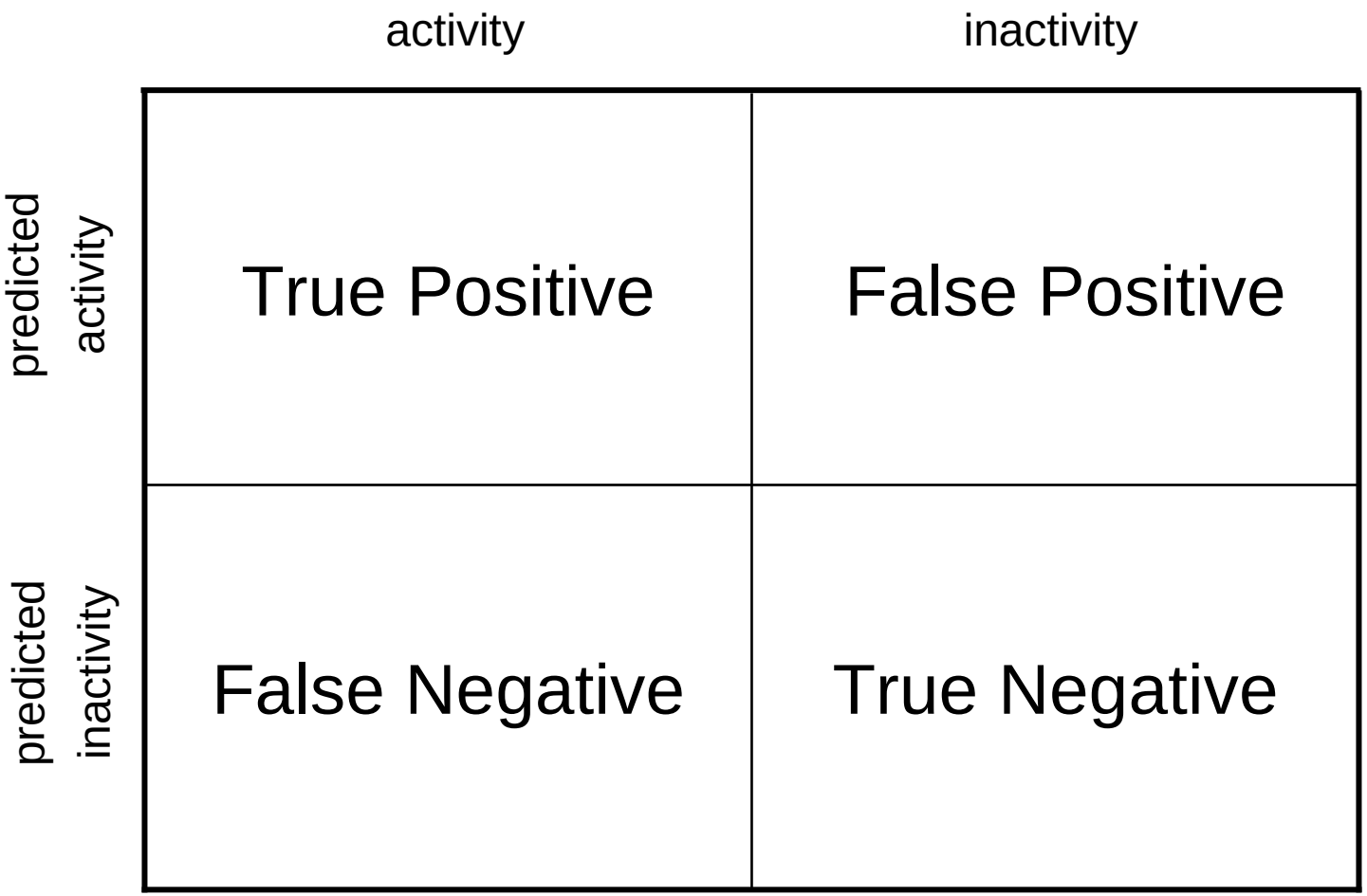
ROC curves

$$TP_{Rate} = Se_{\text{subset}} = \frac{ligands_{\text{selected}}}{ligands_{\text{total}}}$$

$$FP_{Rate} = (1 - Sp)_{\text{subset}} = \frac{decoys_{\text{selected}}}{decoys_{\text{total}}}$$

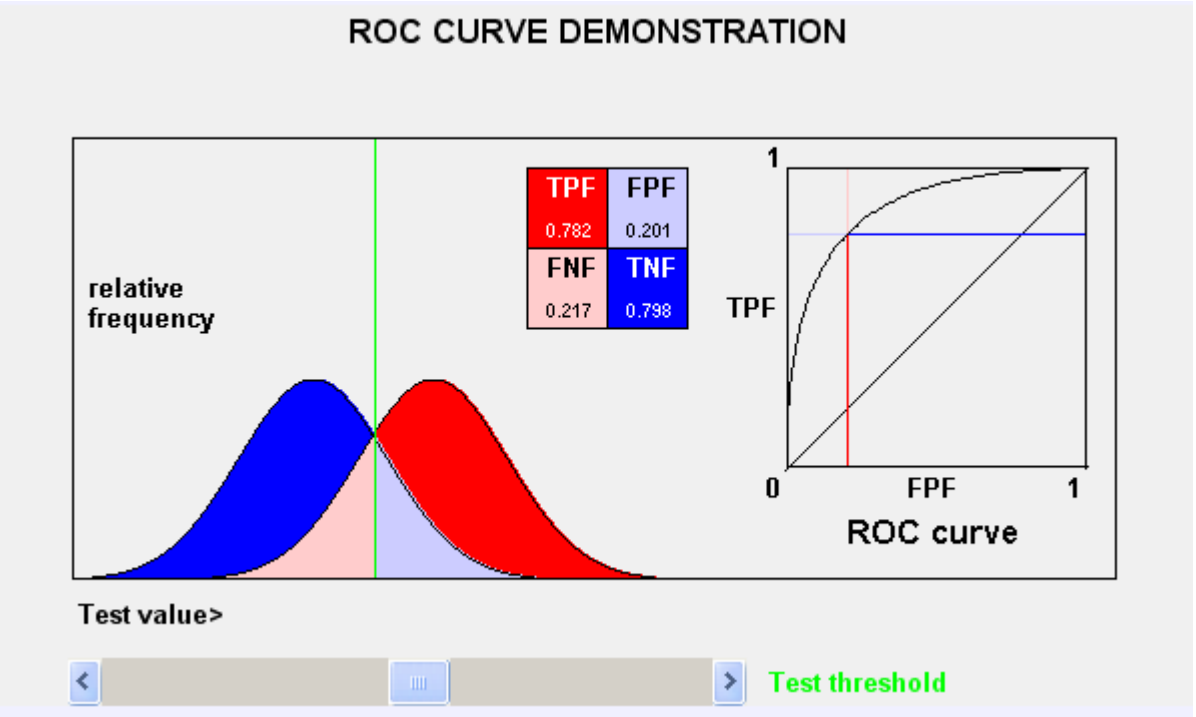
Se - Sensitivity, Sp - Specificity

Computational Prediction vs. Experimental Evidenced



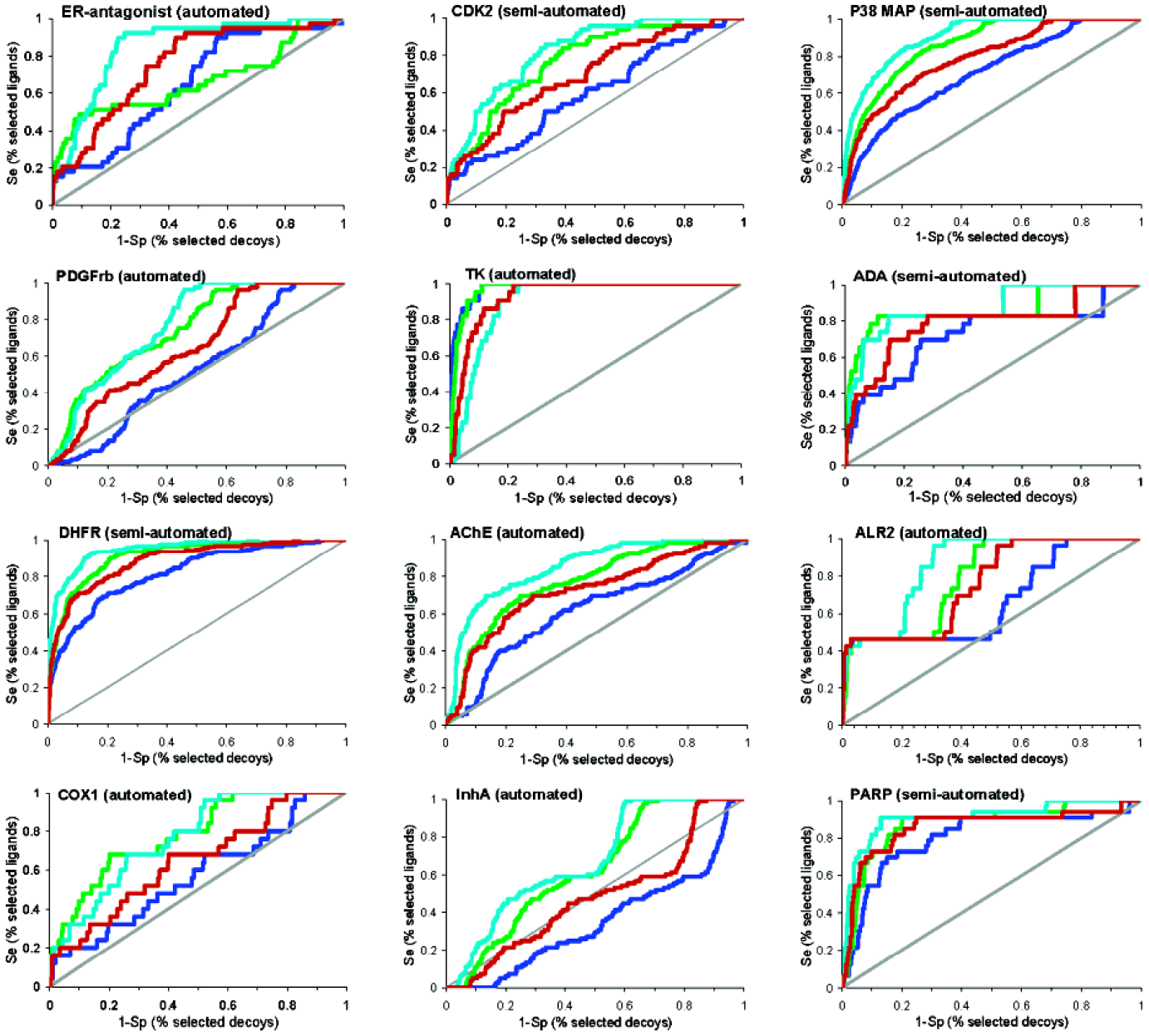
ROC curves

- ROC -- Receiver Operating Characteristic



http://www.anaesthetist.com/mnm/stats/roc/Findex.htm

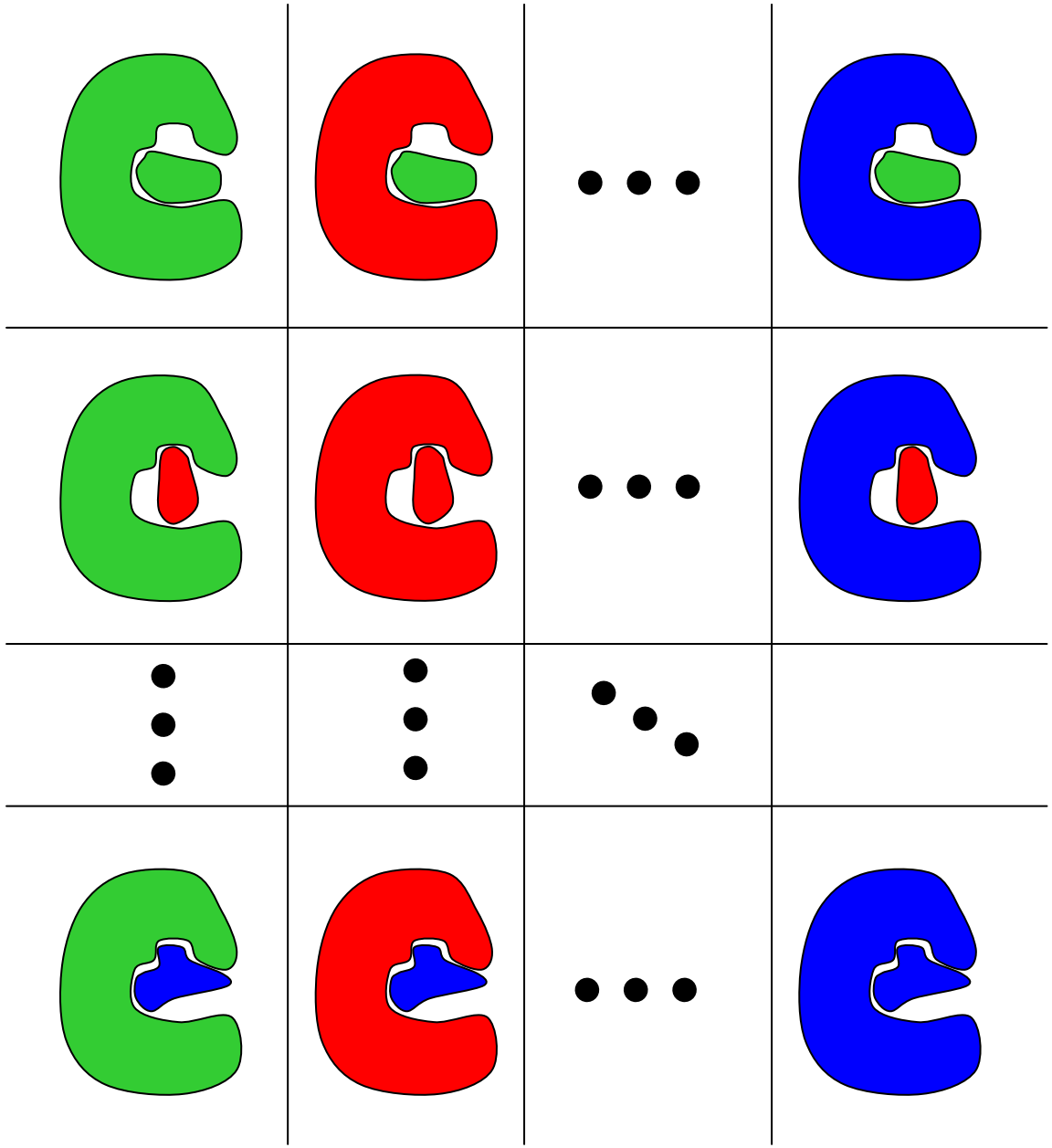
ROC curves



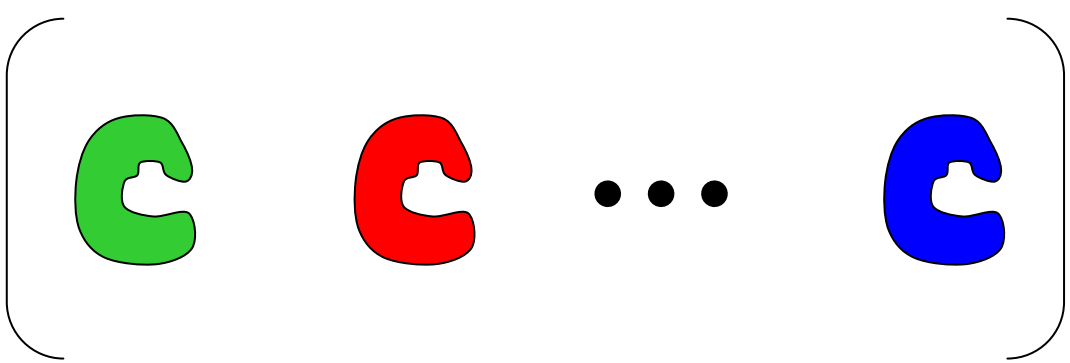
ROC curves for 12 targets
DUD (blue)
MDDR (green)
Jain's decoys (orange)
Rognan's decoys (cyan)
random (gray)

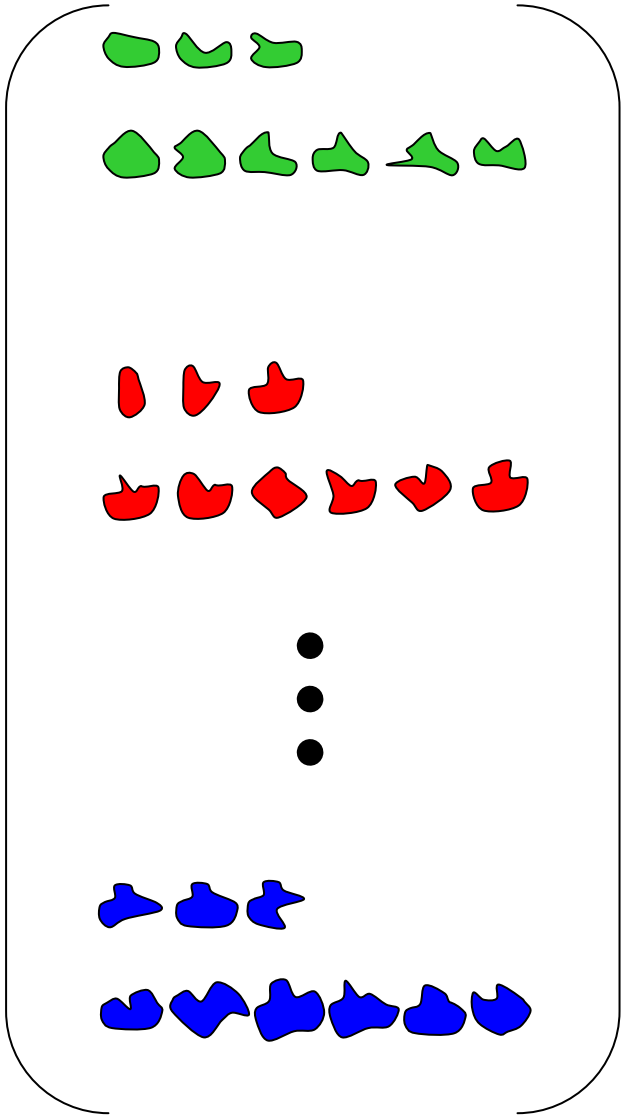
DUD Cross-Enrichments

Cross-Docking

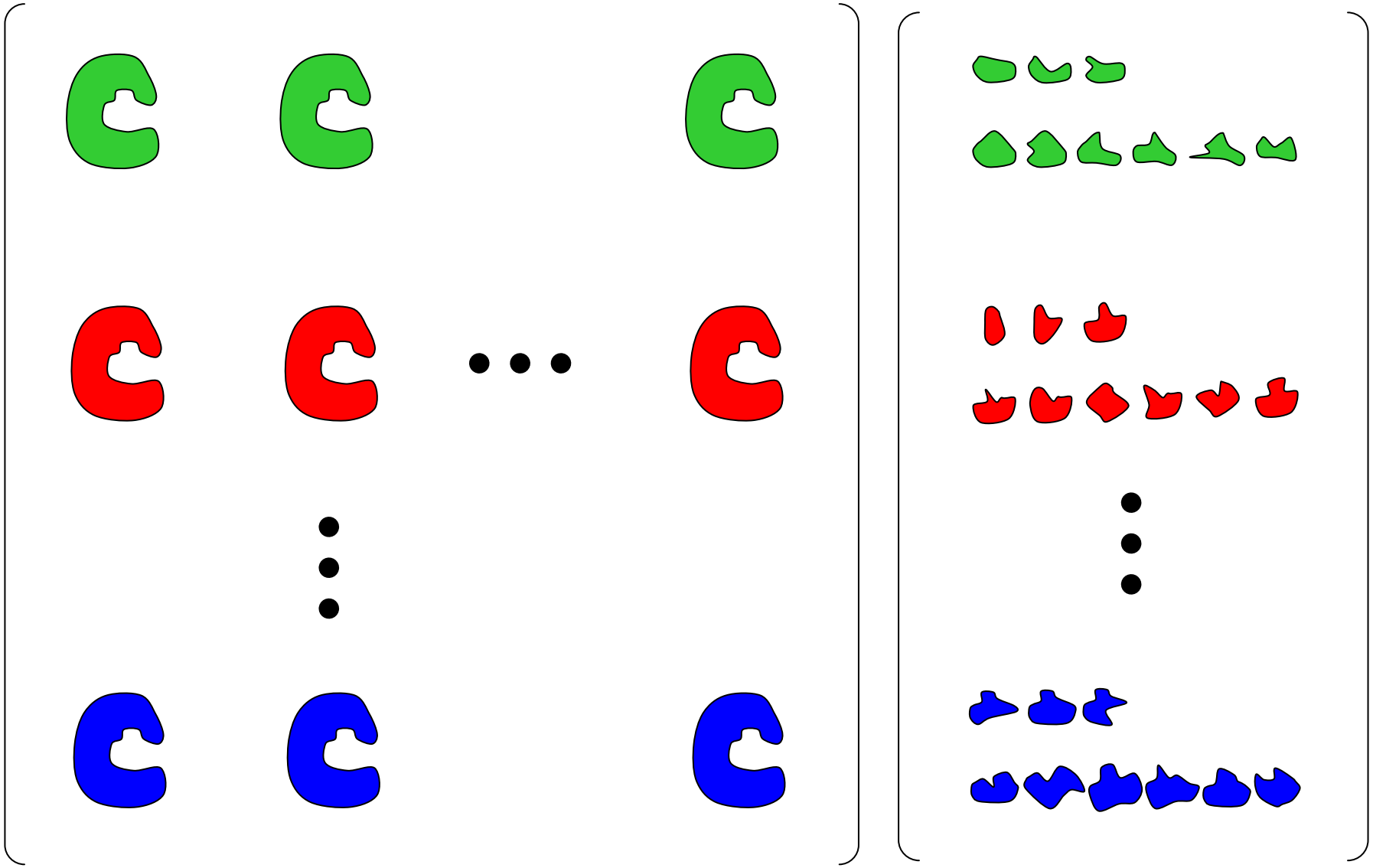


"Cognate" Enrichment Study





Cross-Enrichments



Cross-Enrichments

Matrix of Cross-Enrichments

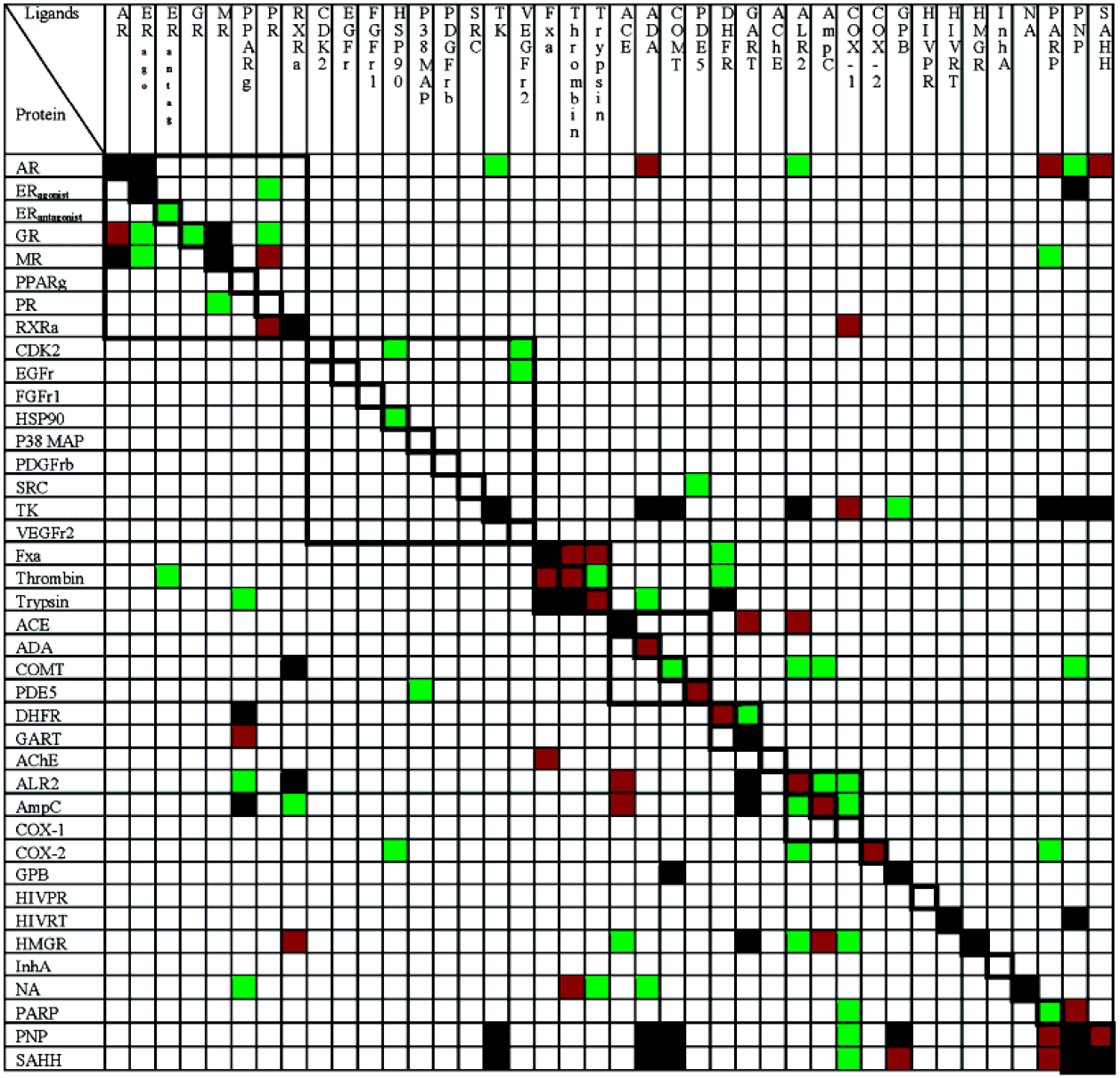
Very good (black):
ETmax ≥ 30 and
ET20 ≥ 3

good (red):
30 > ETmax ≥ 20 and
3 > ET20 ≥ 2.5

medium (green):
20 > ETmax ≥ 10 and
2.5 > ET20 ≥ 2

poor (white):
ETmax < 10 and ET20 < 2

boxes are drawn around related targets.



J. Med. Chem. **2006,** *49,* 6789-6801

Cross-Enrichments

Matrix of Cross-Enrichments

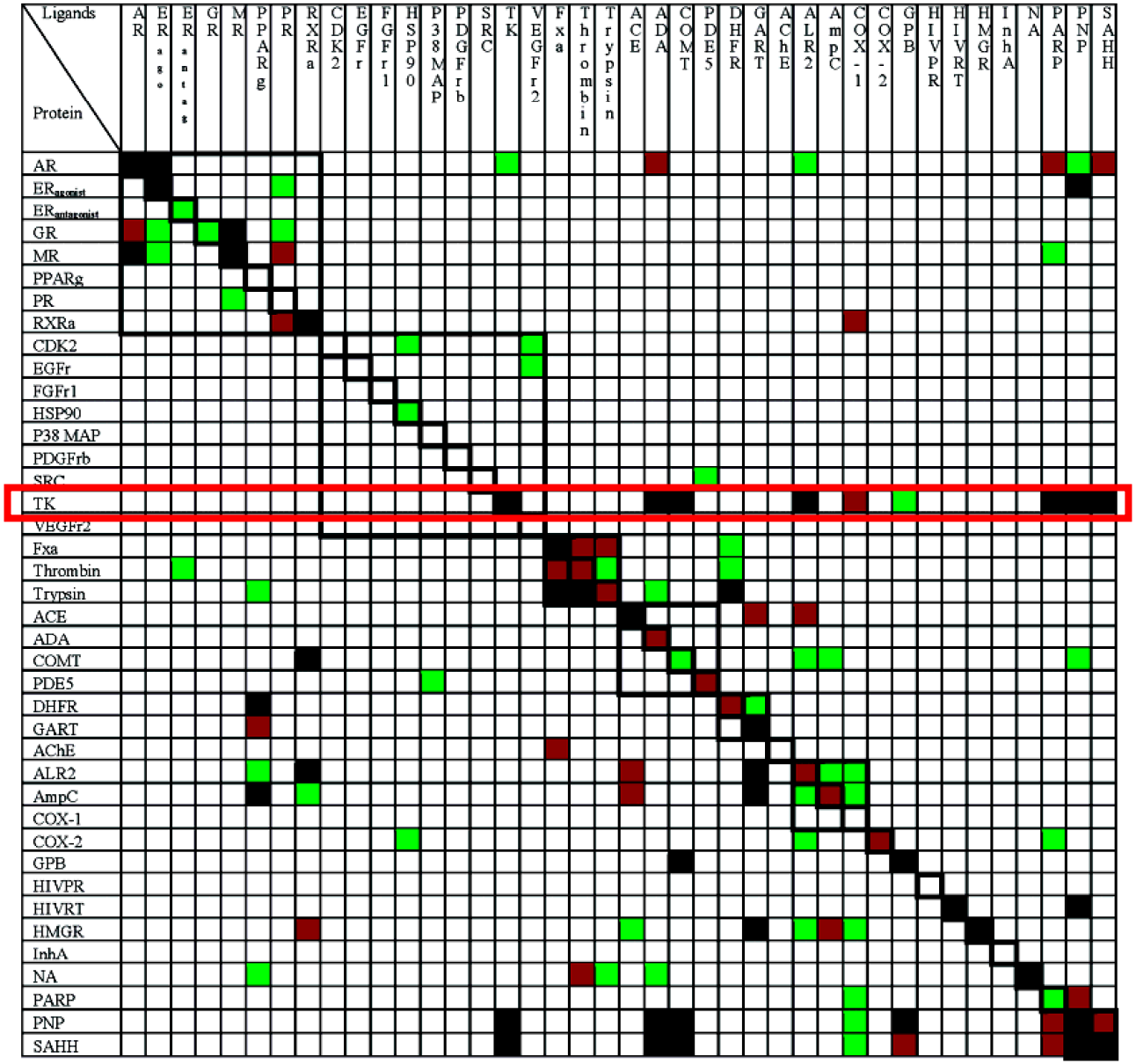
Very good (black):
ETmax ≥ 30 and
ET20 ≥ 3

good (red):
30 > ETmax ≥ 20 and
3 > ET20 ≥ 2.5

medium (green):
20 > ETmax ≥ 10 and
2.5 > ET20 ≥ 2

poor (white):
ETmax < 10 and ET20 < 2

boxes are drawn around related targets.



J. Med. Chem. **2006,** *49,* 6789-6801

Statistics and Timings

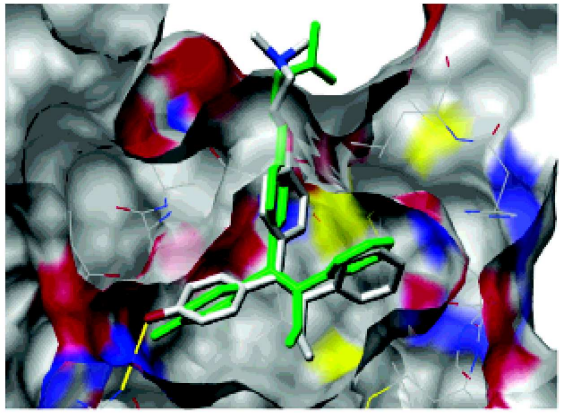
Table 3. Docking Statistics on Six Representative Targets

receptor	unique molecules scored[a]	total molecules scored[b]	orientations sampled per molecule	conformations sampled per molecule	total configurations scored[b]	total time (h)[c]
ER	97 427	416 990	1 895	6 543	2.69×10^{10}	54.4
P38 MAP	93 887	294 917	592	7 875	8.97×10^{9}	20.1
TK	37 240	180 451	3 437	4 302	2.67×10^{9}	21.9
ADE	85 053	297 400	14 632	5 308	2.19×10^{10}	65.5
ALR2	98 724	430 313	4 272	10 109	1.44×10^{11}	296.4
InhA	97 668	429 579	2 325	6 809	5.87×10^{10}	123.5

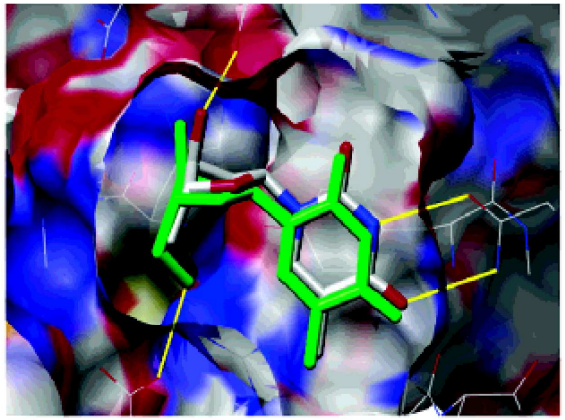
[a] Only orientations and configurations passing the steric filter were scored. [b] Some molecules were represented in the database in multiple rigid fragment, protonation, and tautomeric forms. [c] Scaled to reflect time on a 2800-MHz Pentium IV.

J. Med. Chem. **2006,** *49,* 6789-6801

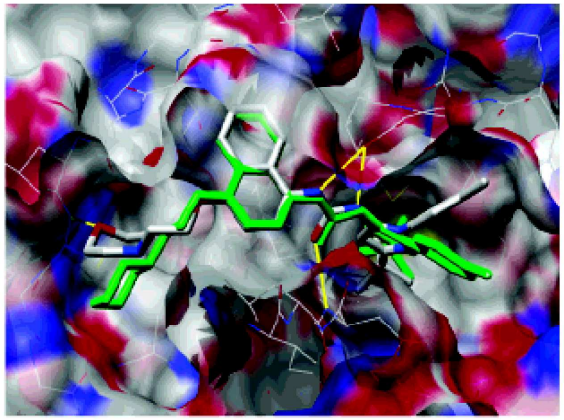
Binding pose predictions



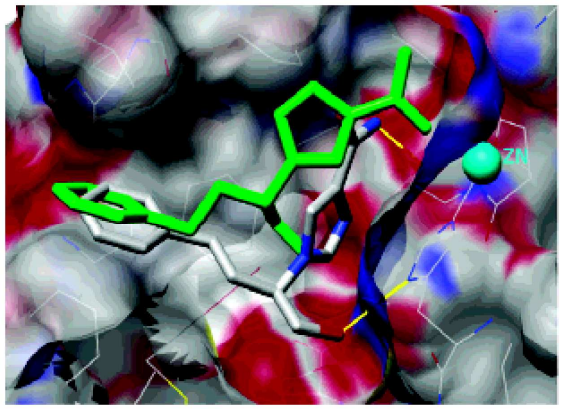
5A. ER



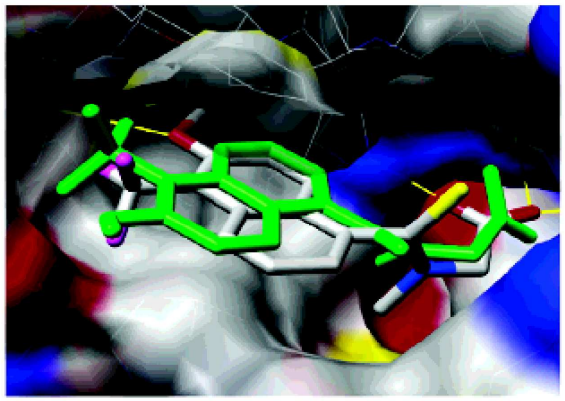
5B. TK



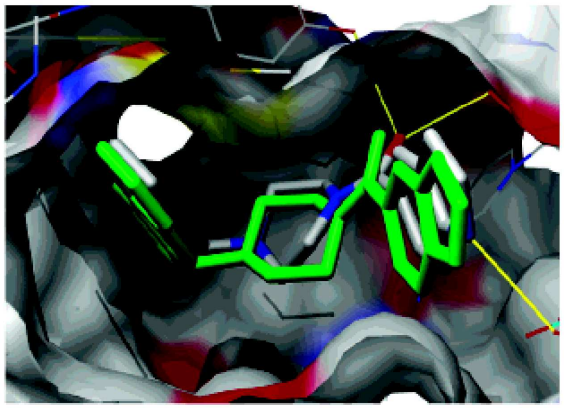
5C. P38 MAP



5D. ADA



5E. ALR2



5F. InhA

- six representative targets
- docked ligands (green)
- crystallographic structures (colored by atom type)
- Key hydrogen bonds (yellow)

J. Med. Chem. **2006,** *49,* 6789-6801

Conclusions

- DUD is designed to match physical properties of active ligands
- Other databases used in enrichment studies are more physically dissimilar from the actives
- DUD gives poorer enrichment over other databases
  - better to gauge a docking program's abilities
- Most systems have no cross-enrichment with notable exceptions including TK

Rescores

http://dud.docking.org/

DUD Release 2: http://dud.docking.org/r2/

Notes accompanying release 2 as found on http://dud.docking.org/r2/

"Why is the ratio of decoys to annotated ligands described as 36 to 1 in the paper, yet there are on average only 33 to 1 in DUD? This is due to overlap, as the same decoy could be used for multiple targets, particularly in the kinase class where there was so much overlap.

Two DUD decoy compounds (ZINC154632 for RXR decoys and ZINC608655 for ER decoys) were structurally identical/similar to the crystal ligands of RXR and ER, individually. This problem was caused by failing to include the crystallographic ligands in our annotated ligands set, and will be fixed in the next version of DUD. Thanks to Paul Hawkins of OpenEye for bringing this to our attention.

Also: PDB code for COX-1 structure in given as 1P4G but should be 1Q4G. We regret this error, and thank alert reader Paul Hawkins of OpenEye for this information Also, Hao Li of UCSF Pharm Chem points out that the PDB id of ADA in the paper is wrong. It should be 1ndw."

SIFp

Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein-Ligand Binding Interactions

Zhan Deng, Claudio Chuaqui, and Juswinder Singh

J. Med. Chem. **2004,** *47,* 337-344

SIFt Introduction

- Structural Interaction Fingerprints (SIFt)
- Identification of Ligand Binding Site Residues
  - non-hydrogen protein atoms solvent accessibility loss upon ligand binding
  - protein atoms h-bonding with the ligands
- Extraction and Classification of Binding Interactions

J. Med. Chem. **2004,** *47,* 337-344

SIFt Introduction

- Seven different types of interactions
  (1) residue is in contact with the ligand
  (2) backbone is in contact
  (3) sidechain is in contact
  (4) polar interaction
  (5) non-polar interaction
  (6) h-bond acceptor
  (7) h-bond donor
- Concatenating all figure prints together

J. Med. Chem. **2004,** *47,* 337-344

SIFt Introduction

- Three applications of SIFt in Drug Discovery :
  - sorting, clustering, and organizing docking poses (identifying like binding poses)
  - organizing and clustering 90 crystal complexes
  - filtering virtual screening results to find ligands with certain binding mode and interaction patterns

J. Med. Chem. **2004,** *47,* 337-344

Tanimoto Coefficient

$$Tc = \frac{|A \cap B|}{|A \cup B|}$$

- intersection is # of ON bits common in both A and B
- union is # of ON bits present in either A or B

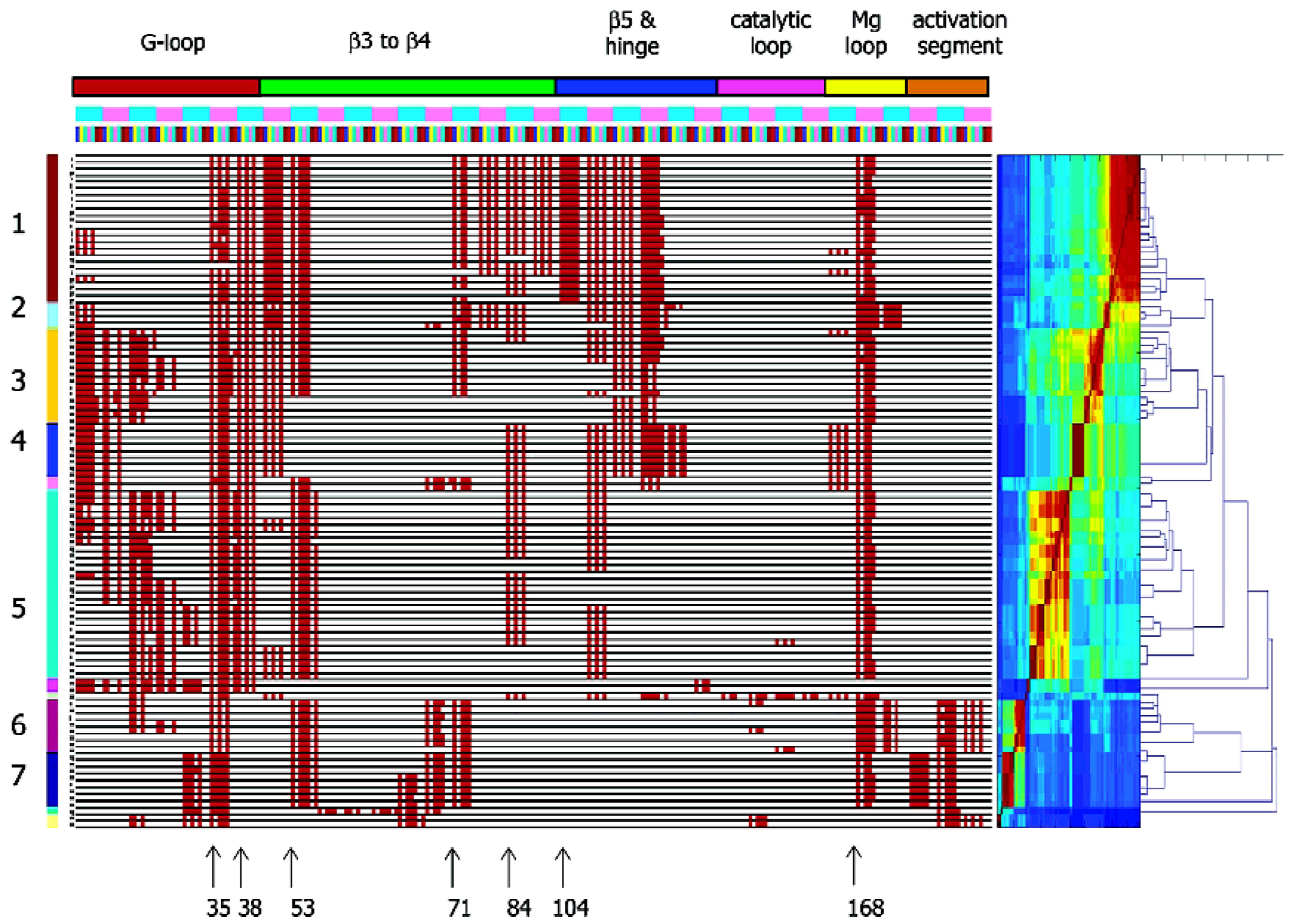
J. Med. Chem. **2004,** *47,* 337-344

Docking studies

- Study #1 (single ligand)
  - ligand SB203580 docked to p38 (pdb code 1a9u)
  - poses generated with FlexX in Sybyl
  - 100 poses generated
- Study #2 (enrichment study)
  - 16 known p38 inhibitors
  - 1000 with diverse chemical structures
  - docked database to p38 (pdb code 1a9u)
  - 30 480 (30 1016) poses generated

J. Med. Chem. **2004,** *47,* 337-344

SB203580 Clusters in P38



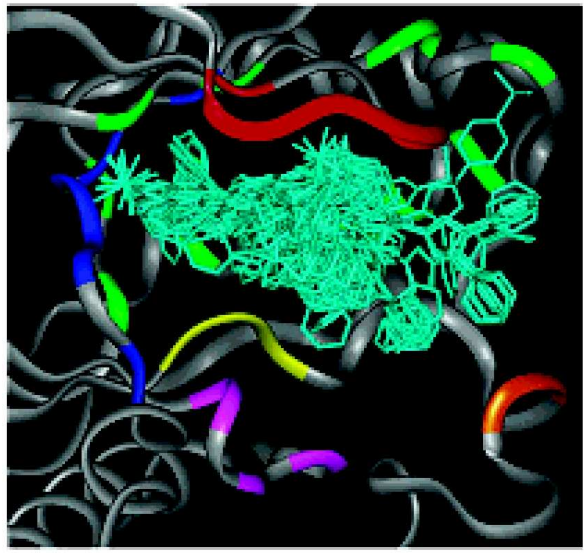
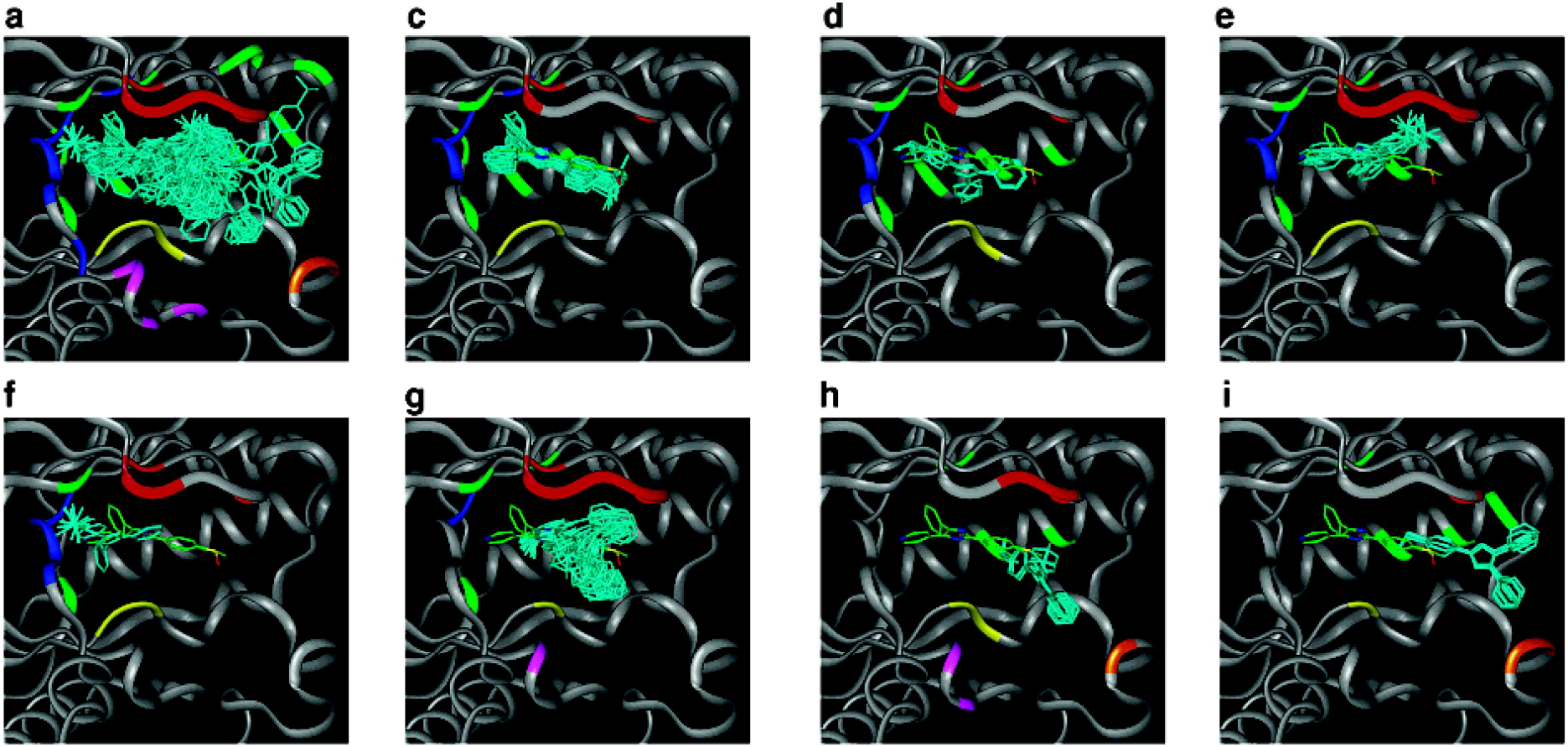
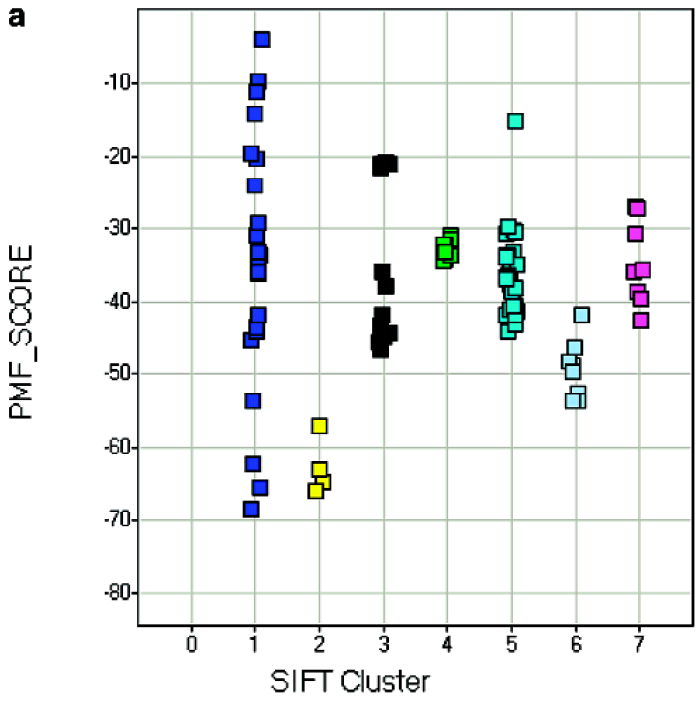


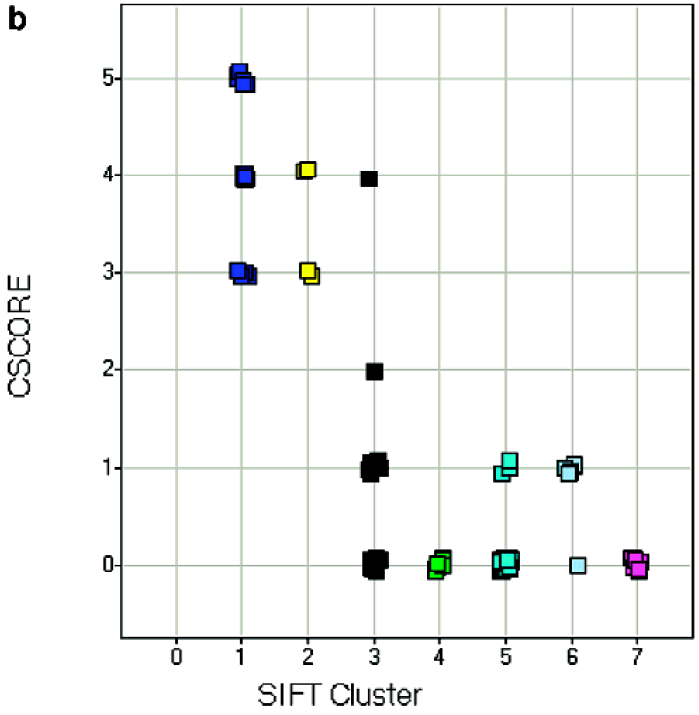
Figure shows the 100 poses generated in Docking study #1, SB203580 docked to p38

J. Med. Chem. **2004,** *47,* 337-344

SB203580 Clusters in P38





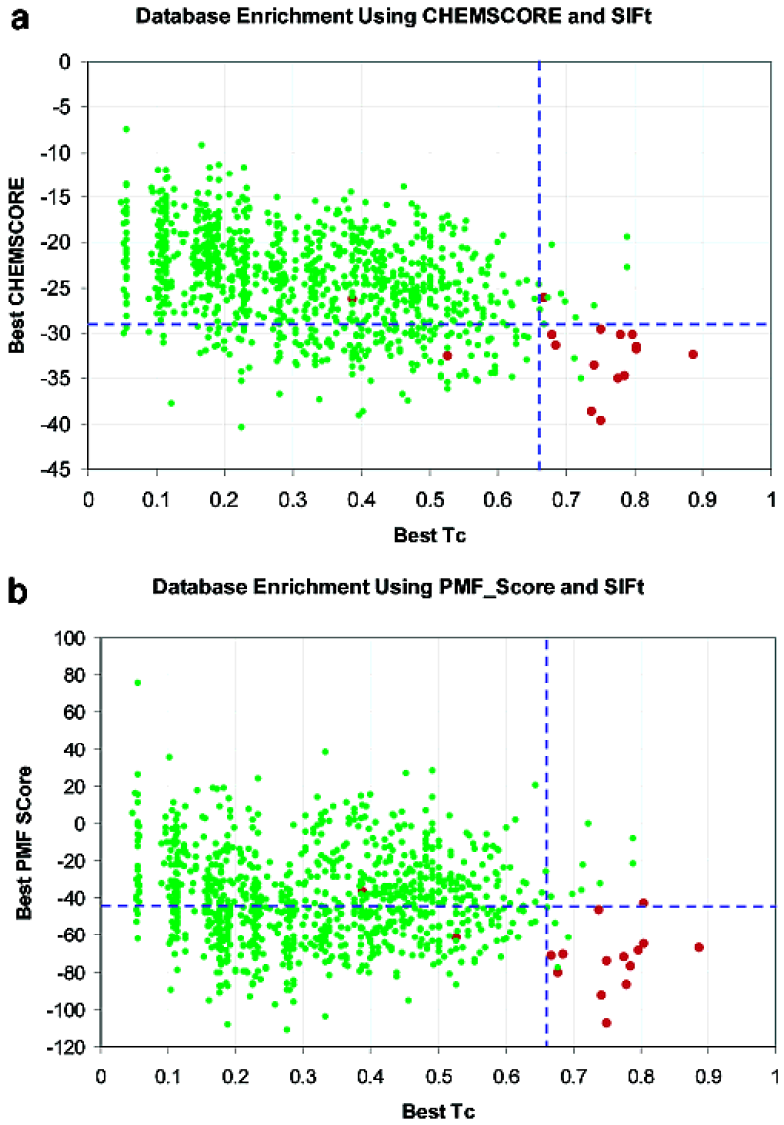


SIFt Clusters

- Scores are not able to identify the binding mode (SIFt)

J. Med. Chem. **2004,** *47,* 337-344

Enrichment



- comparison of SIFt with 2 alternative scoring functions
- SIFt gives good enrichment

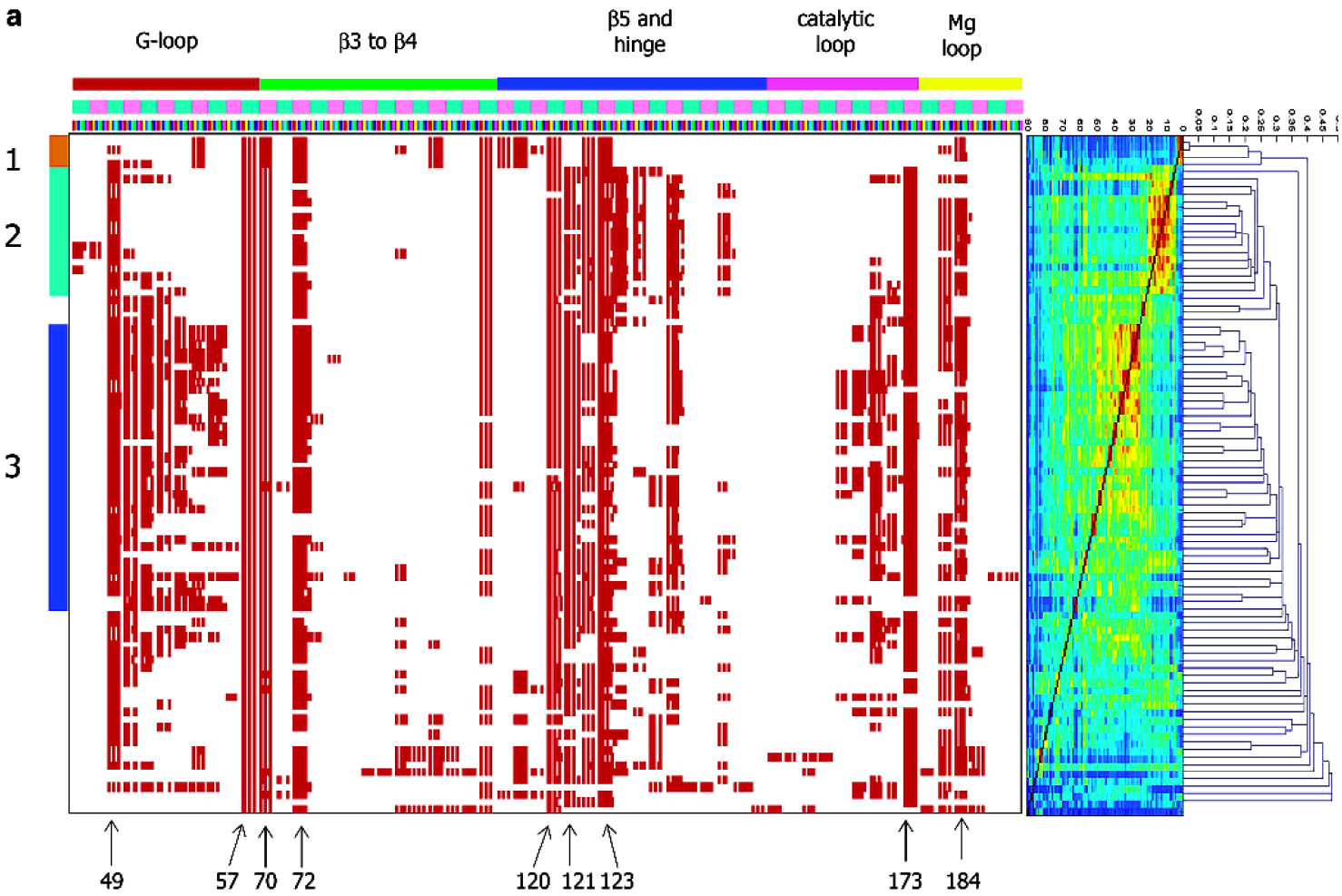
Table 1. Comparison of the Database Enrichment Performances of SIFt with ChemScore and PMF Score

filtering method	EF[a]
PMF Score	2.0
ChemScore	5.4
SIFt	37.0
SIFt + ChemScore	42.3

Crystal Structure study

- Study #3 (Kinase family analysis)
  - 89 kinase-ligand complexes
    - inhibitor or substrate in ATP binding cleft
    - all active site residues are present in structure
  - 25 different kinases
  - 14 different protein kinase subfamilies
  - 54 unique compounds

J. Med. Chem. **2004,** *47,* 337-344



J. Med. Chem. **2004,** *47*, 337-344

Conclusions

- SIFt is a powerful tool
  - pose clustering
  - family clustering
  - filtering screening results
- possible improvements
  - incorporate more types of interactions in the fingerprint
  - uses only subset of residues
  - uses scaled numeric data representing interactions

J. Med. Chem. **2004,** *47,* 337-344