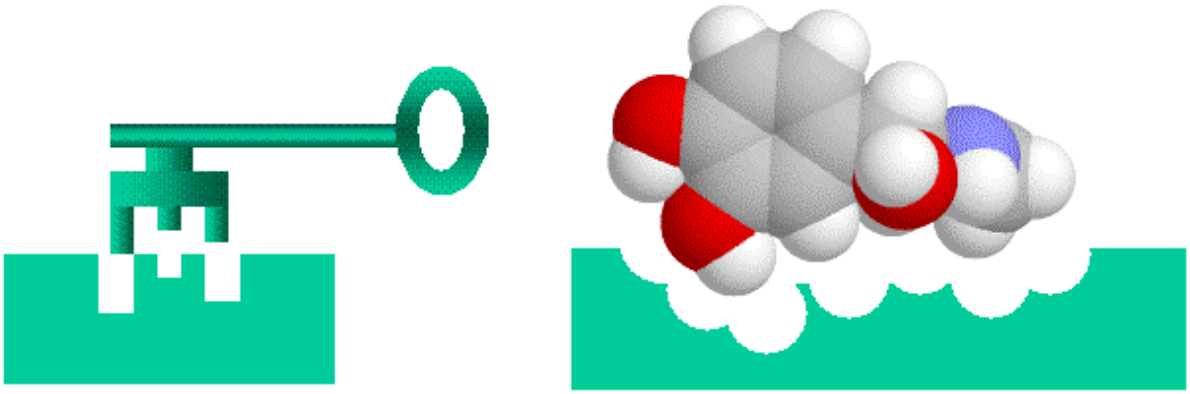
OPTIMIZATION OF SHAPE FINGERPRINTS FOR PROTEIN-LIGAND SYSTEMS

Joanna Zarnecka, Andrew G. Leach, Steven J. Enoch

LIVERPOOL JOHN MOORES UNIVERSITY

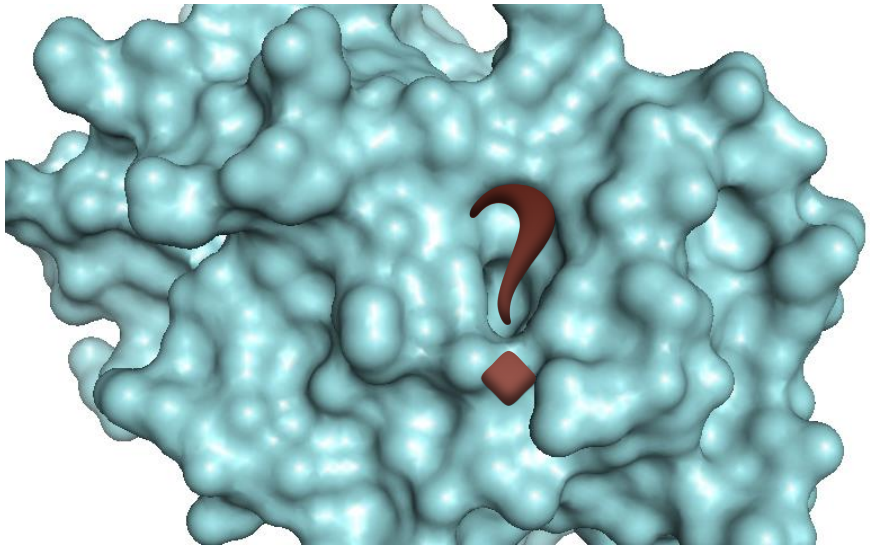
MedChemica
Creating a step change in Medicinal Chemistry

Introduction



- **Shape complementarity** – "lock and key" concept (or "hand in glove" concept)

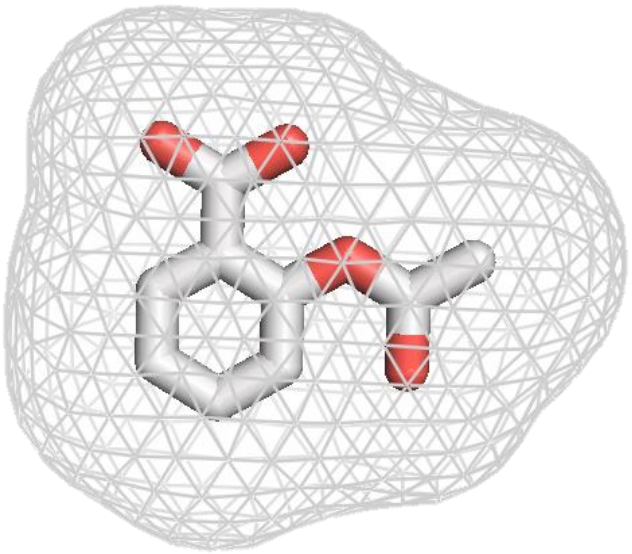
Introduction



Bioisosterism - Molecules similar in size and shape are more likely to show similar activity towards the same target macromolecule

Shape Fingerprints

- Binary bit strings that encode the shape of compounds
- Shape Database consists of diverse shapes of molecules



SHAPE DATABASE

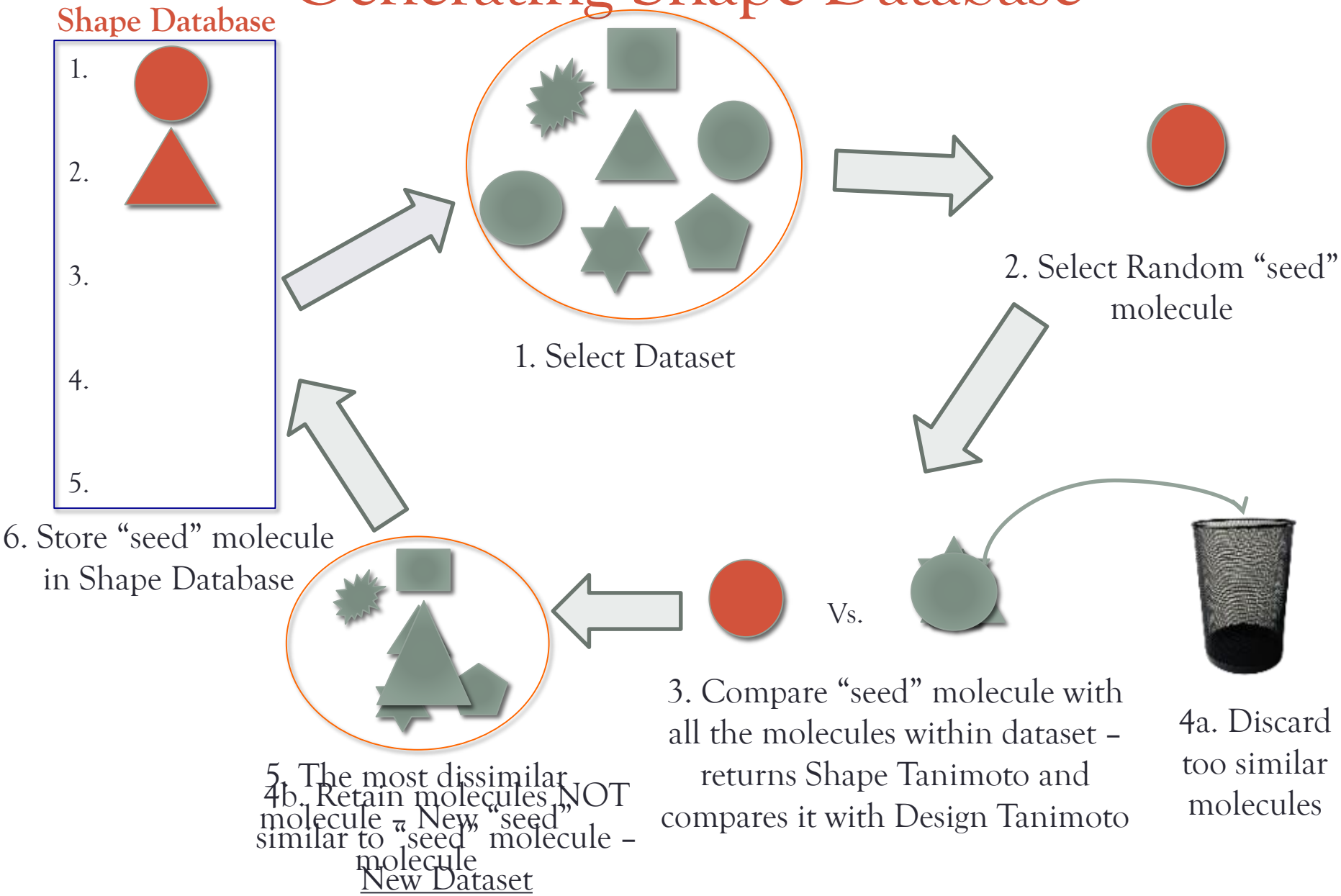
1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0

Haigh, J.A.; Pickup, B. T.; Grant, J.A.; Nicholls, A., Small Molecule Shape-Fingerprints, *J. Chem. Inf. Model.*, **2005**, 45, 673–84.

Our Aim

Optimization of description of **SHAPE** via fingerprints for the explanation of **BIOLOGICAL ACTIVITY**

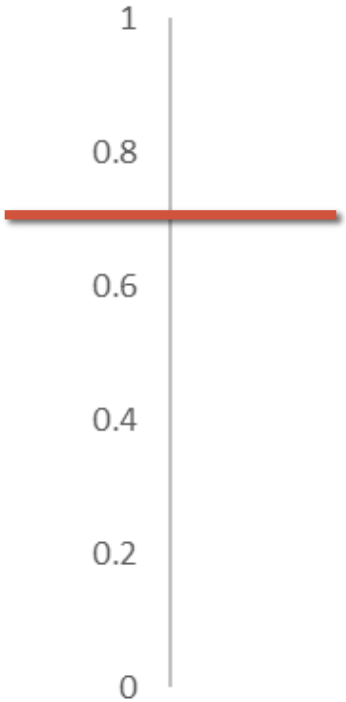
Generating Shape Database



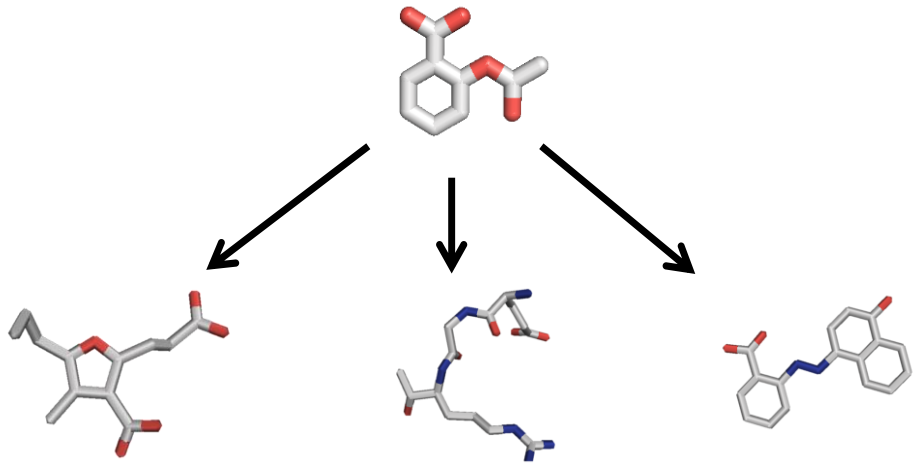
Shape Tanimoto and Design Tanimoto

- **Shape Tanimoto** – measure of similarity of two molecules based on their volume overlay; calculated by Openeye's Shape TK

- **Design Tanimoto** – user-defined cut-off value for Shape Database generation



Generating Shape Fingerprints : Bit On Value



Reference molecule (aspirin)

Hypothetical Shape Database

BOV = 0.5

0.77	0.46	0.57

Calculated Shape Tanimoto

1	0	1

Generated bit string

- The bit corresponding to each reference shape is set On (1), if the Tanimoto is above a user-defined cut-off (the Bit On Value), or Off (0) if it is not

Comparing Shape Fingerprints: Fingerprint Tanimoto

A	1	0	1	0
B	1	1	0	0

$$FT_{AB} = \frac{bothAB}{onlyA + onlyB + bothAB} = \frac{1}{1+1+1} = 0.33$$

onlyA, onlyB - the numbers of unique bits On in the bit strings for A and B respectively

bothAB - the number of bits On in common to A and B

Fingerprint Tanimoto (FT) similarity values vary from 0 (for dissimilar compounds) to 1 (for the most similar molecules).

Our Aim

Optimization of description of **SHAPE** via fingerprints for the explanation of **BIOLOGICAL ACTIVITY**

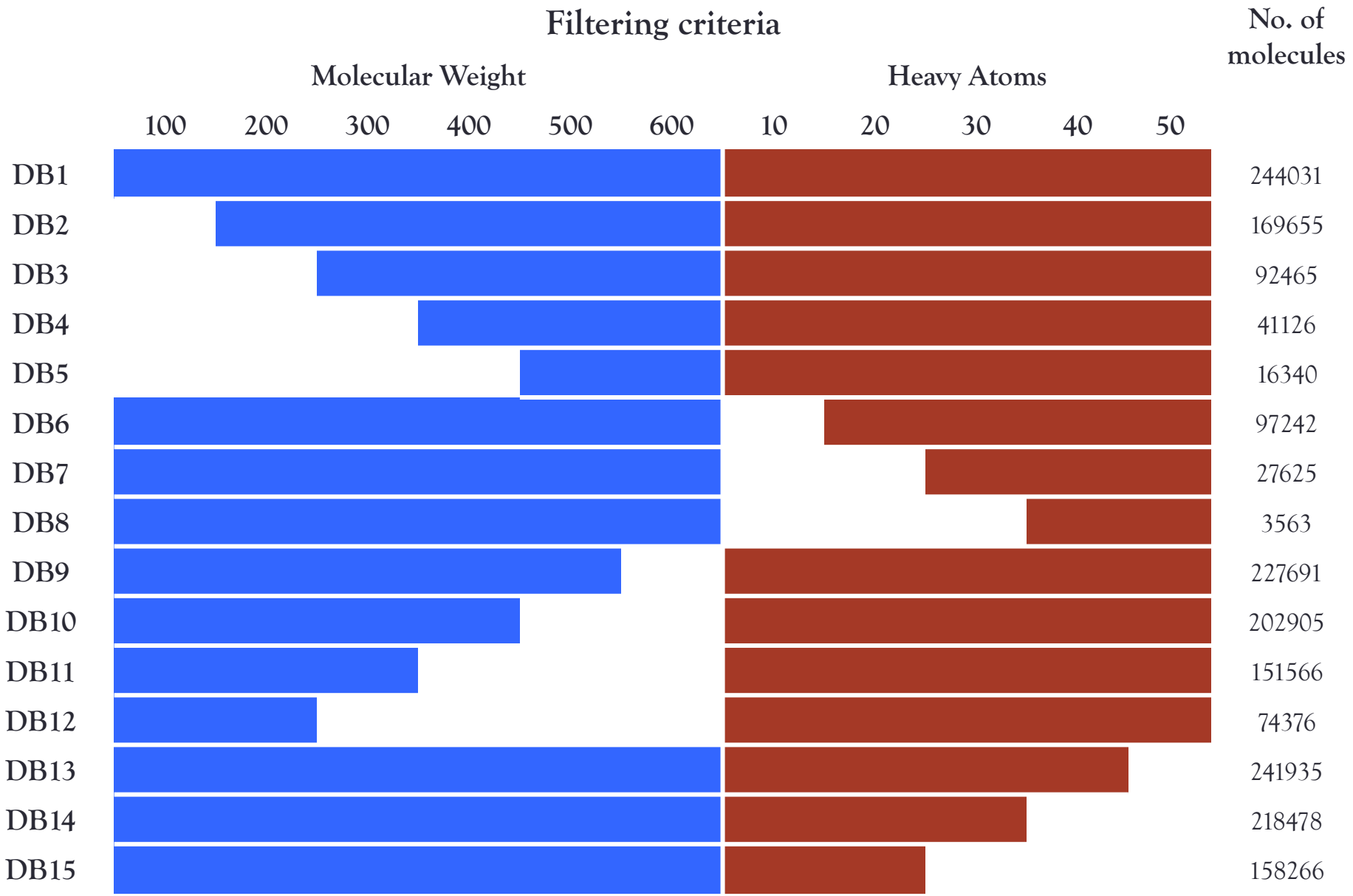
Selecting Dataset

- The Ligand Expo (**http://ligand-expo.rcsb.org**) dataset was used to derive our Shape Database

- consists of 1 158 763 experimental coordinates for non-polymer molecules and non-standard amino acids and nucleotides

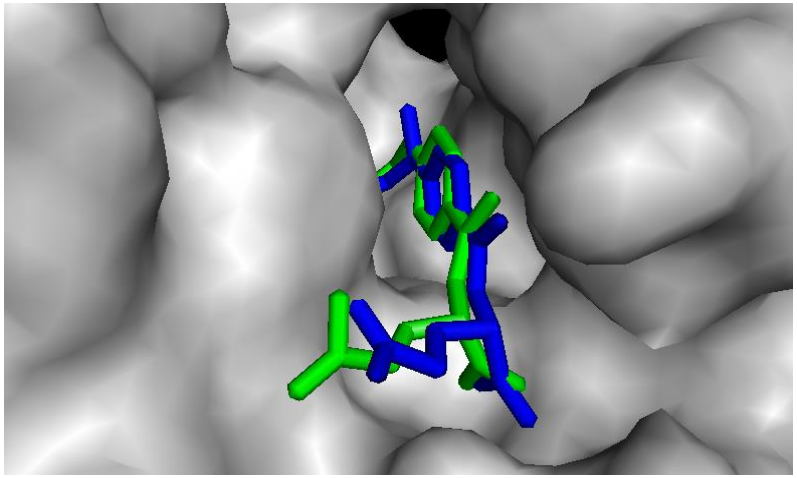
RCSB PDB
PROTEIN DATA BANK

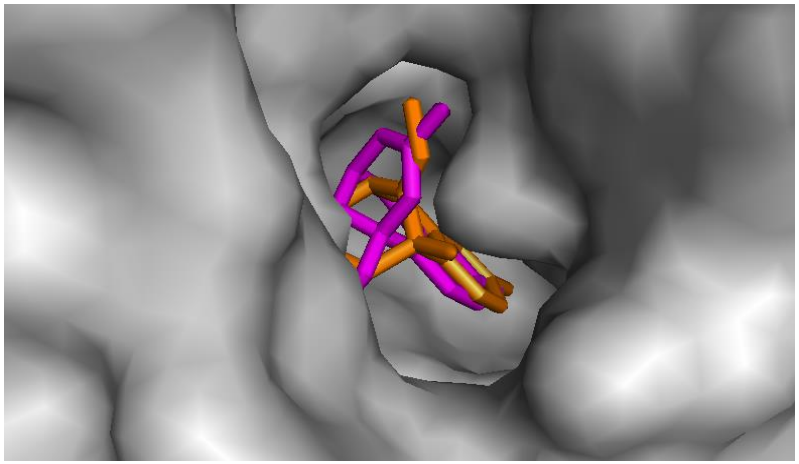
Filtering Dataset



Selecting Test Set – Taylor's Set

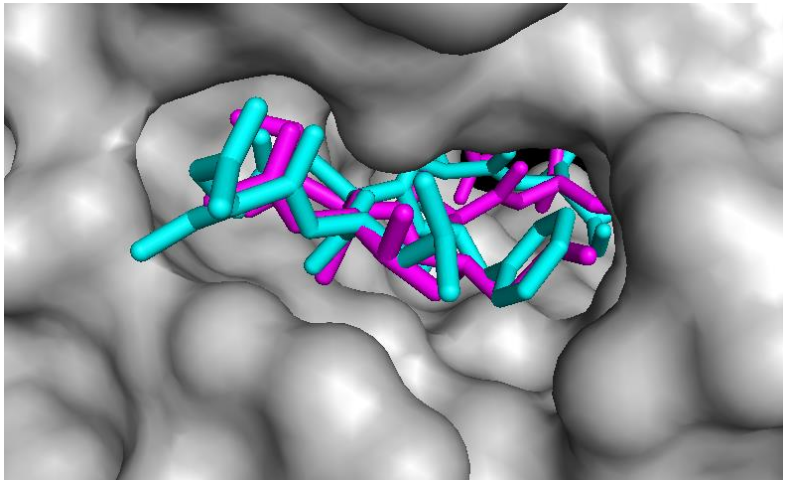
Protein Name	Abbreviation	No. of complexes (no of bound ligands)
Protein kinase 5	PK5	2
Fatty acid binding protein	FABP	3
Neprilysin	NEP	4
Dihydrofolate reductase	DHFR	6
Checkpoint kinase	CHk1	16
Neuraminidase	NEU	11
Carbonic anhydrase	CA	13
Adenosine deaminase	ADA	11
Heat shock protein 90	HSP	10
Acetylcholinesterase	AChE	11

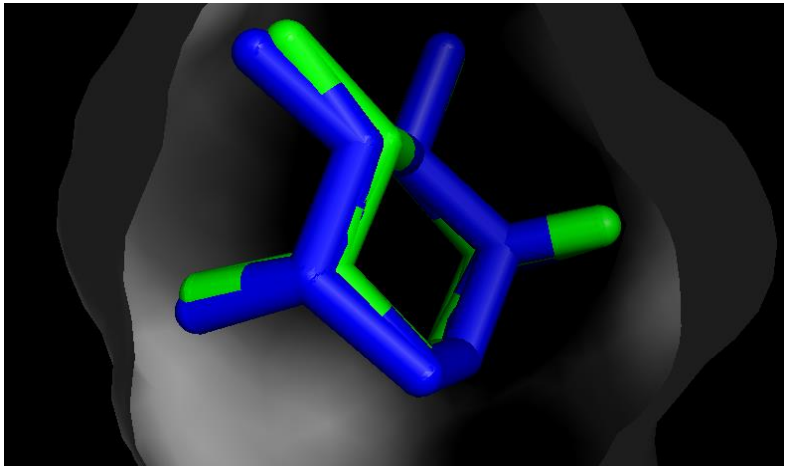




Taylor, R.; Cole, J.C.; Cosgrove, D.A.; Gardiner E.J.; Gillet V.J.; Korb, O. Development and Validation of an Improved Algorithm for Overlaying Flexible Molecules, *J. Comput-Aided Mol Des*, **2012**, 26, 451–72.

Selecting Test Set - Head's Set

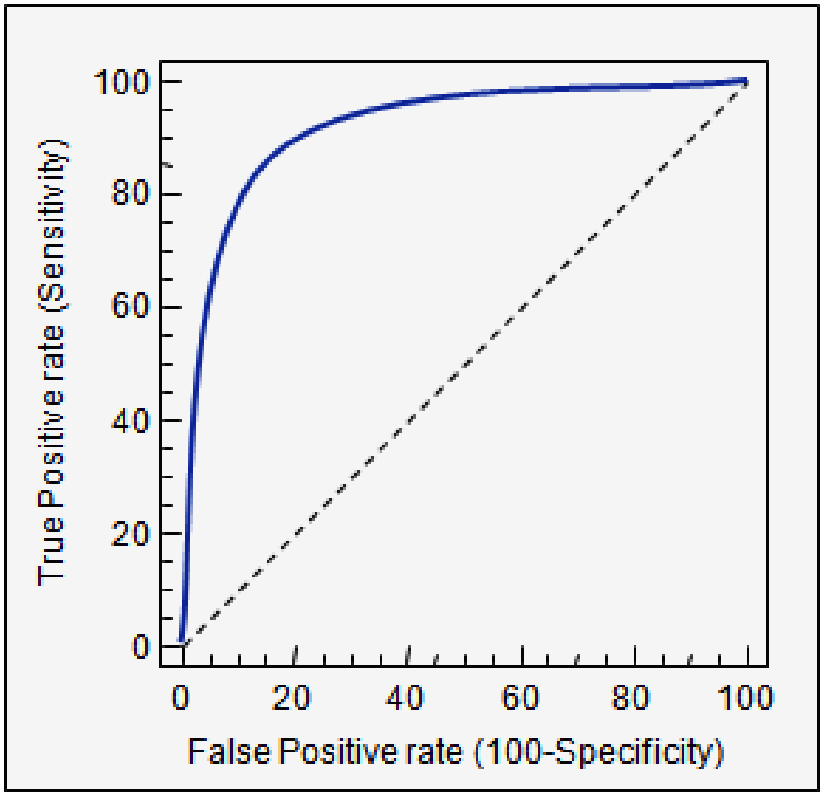




Protein Name	No. of complexes (no. of bound ligands)
HIV-1 PROTEASE	6
THERMOLYSIN	9
ENDOTHIAPEPSIN	9
L-ARABINOSE BIND PROTEIN	12

Head, R.D.; Smythe, M.L.; Oprea, T.I.; Waller, C.L.; Green, S.M.; Marshall, G.R. VALIDATE: A New Method for the Receptor-Based Prediction of Binding Affinities of Novel Ligands, *J. Am. Chem. Soc.* **1996**, 118, 3959-3969

Analysis – ROC Curves

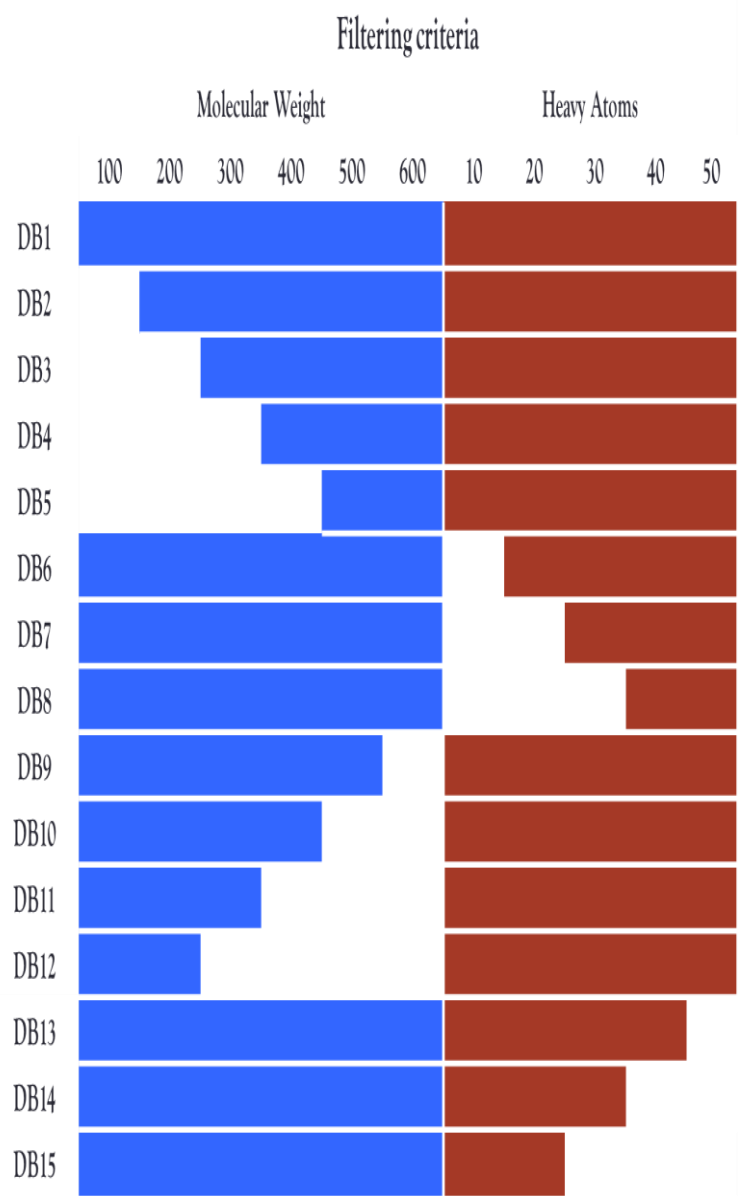


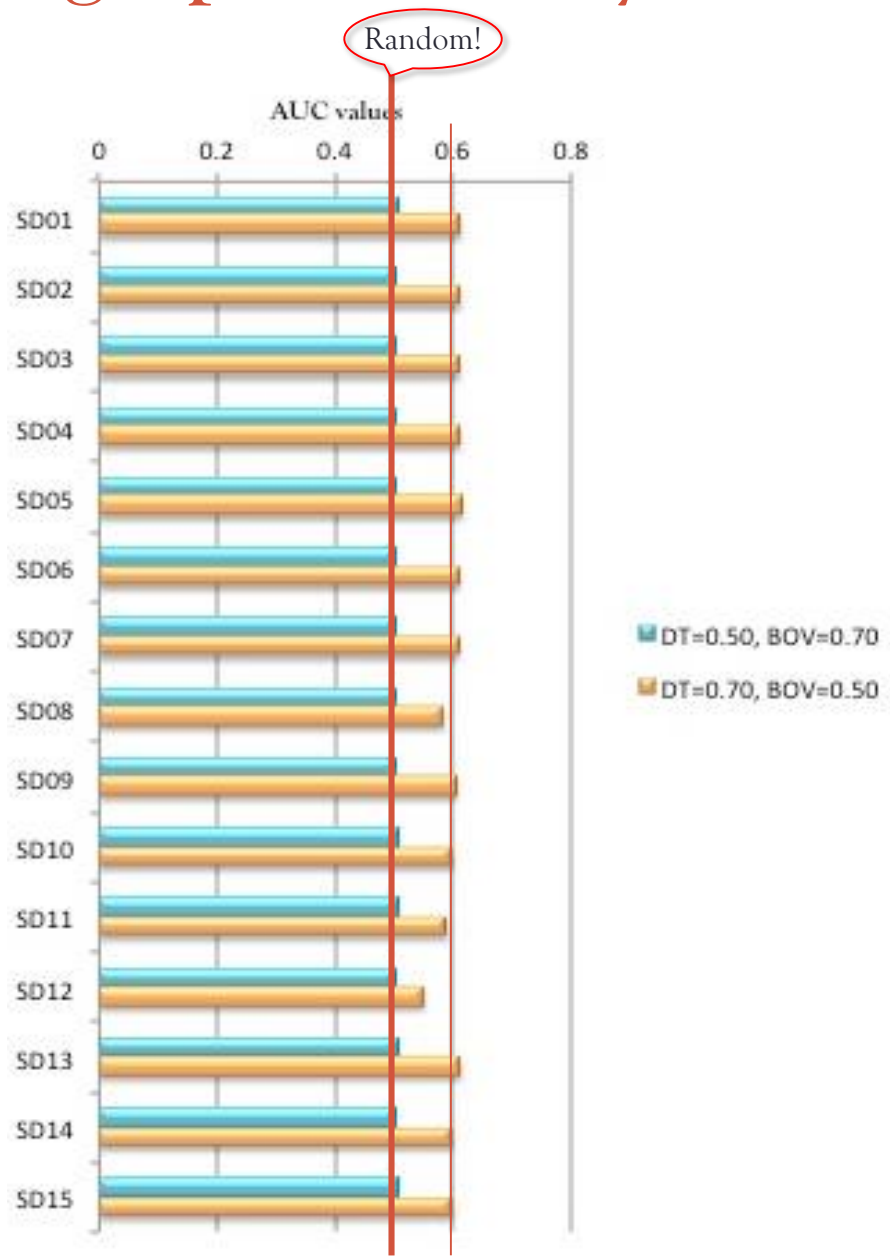
ROC curve – a tool for diagnostic test evaluation

AUC (Area Under Curve) - a measure of how well a parameter can distinguish between two diagnostic groups

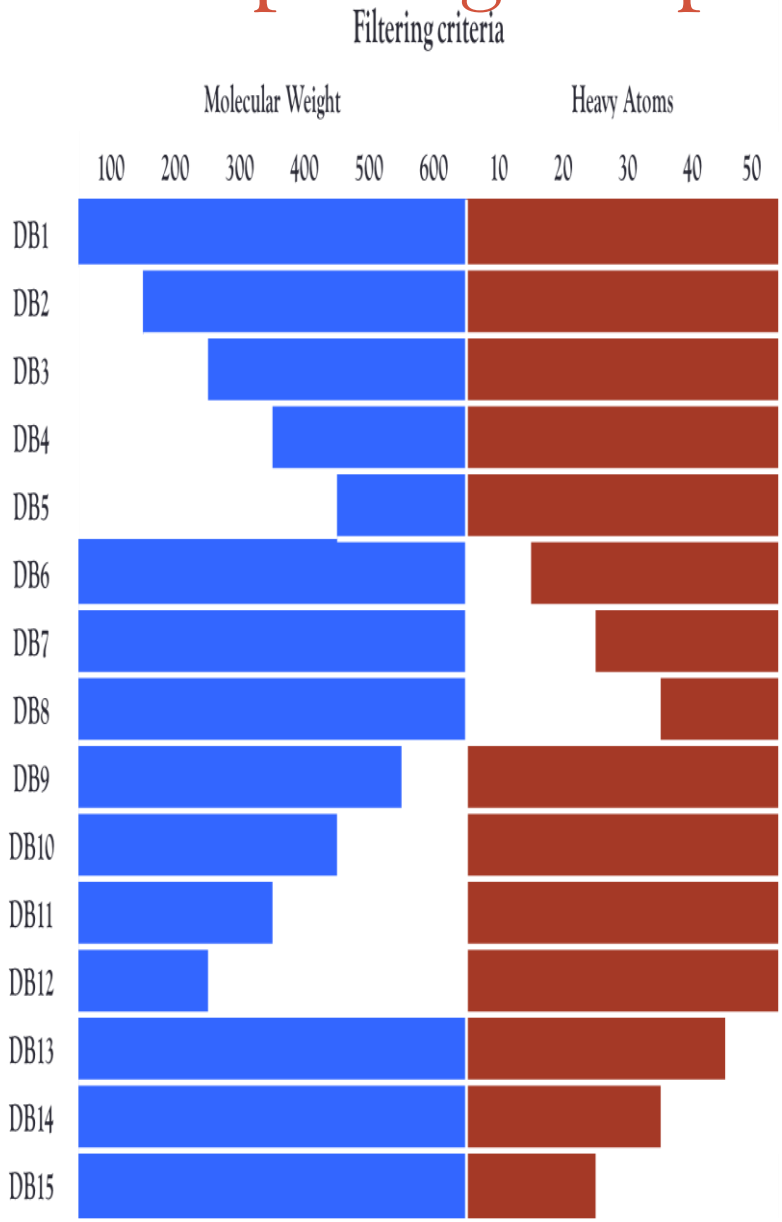
0.90-1 = excellent (A)
0.80-0.90 = good (B)
0.70-0.80 = fair (C)
0.60-0.70 = poor (D)
0.50-0.60 = fail (F)

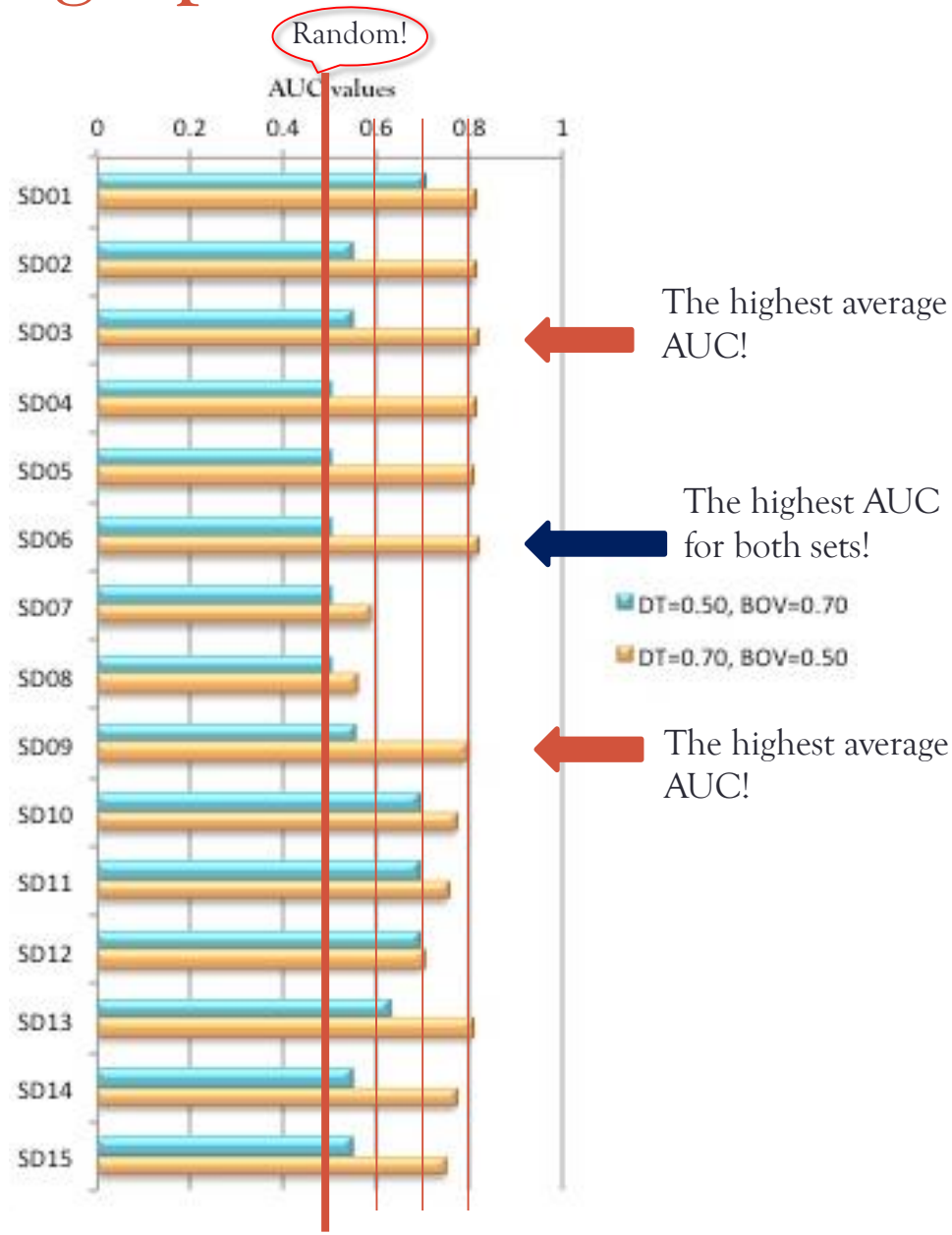
Comparing Shape Fingerprints – Taylor's Set





Comparing Shape Fingerprints – Head's Set



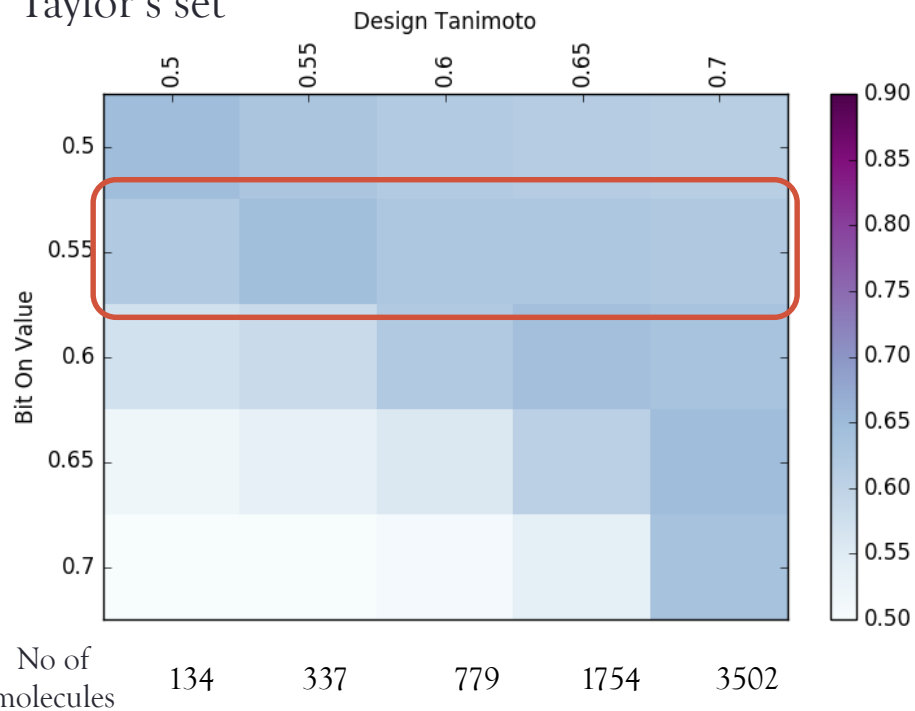


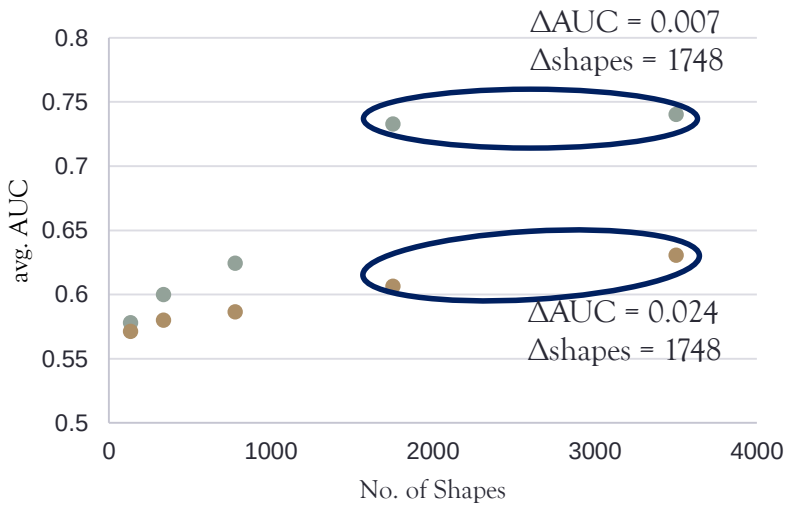
- SD06: MW 100 – 600, HAC 20 – 50

- SD16: MW 300 – 500, HAC 10 – 50 new!
  (SD03 + SD09)

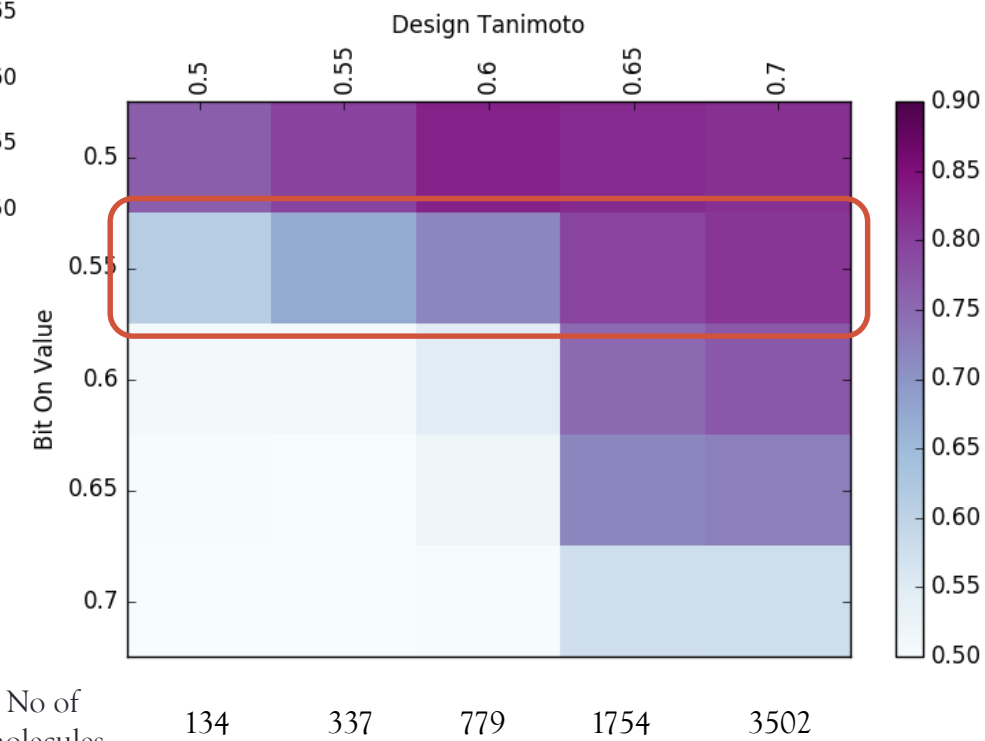
Shape Databases – SD06

Taylor's set



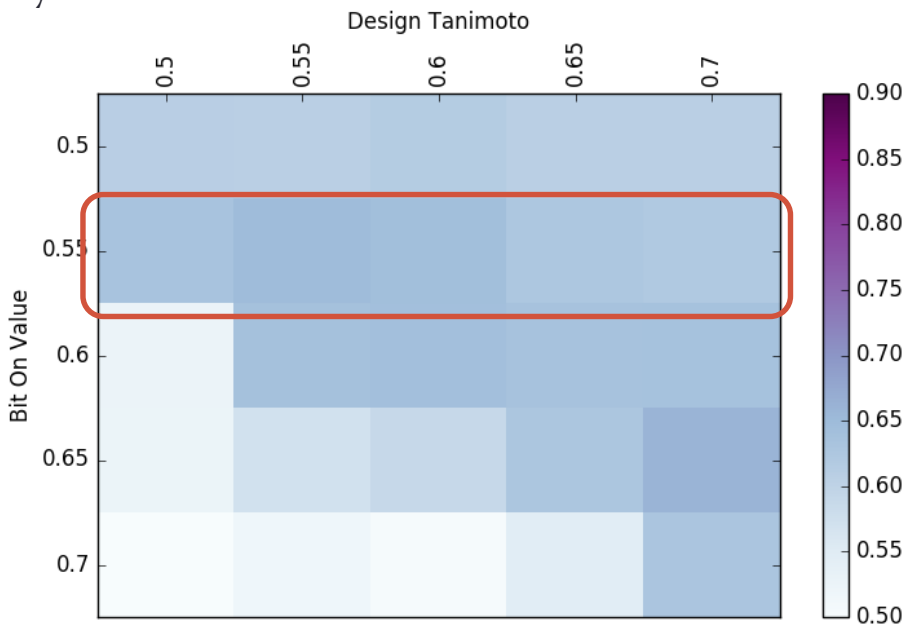


Head's set

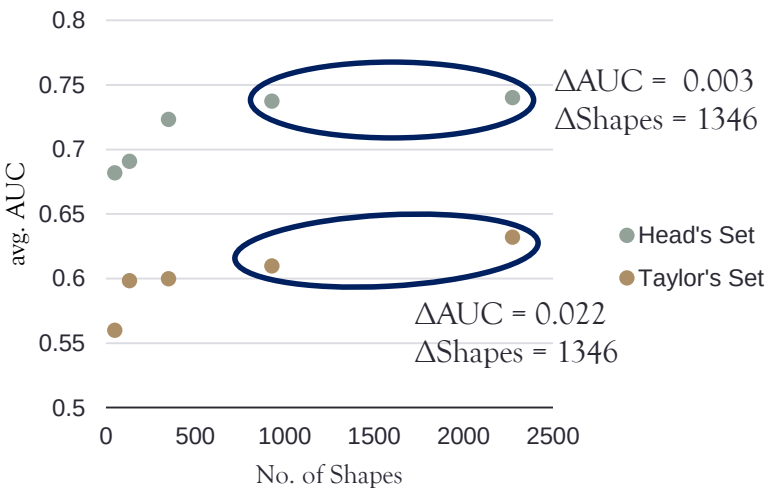


Shape Databases – SD16

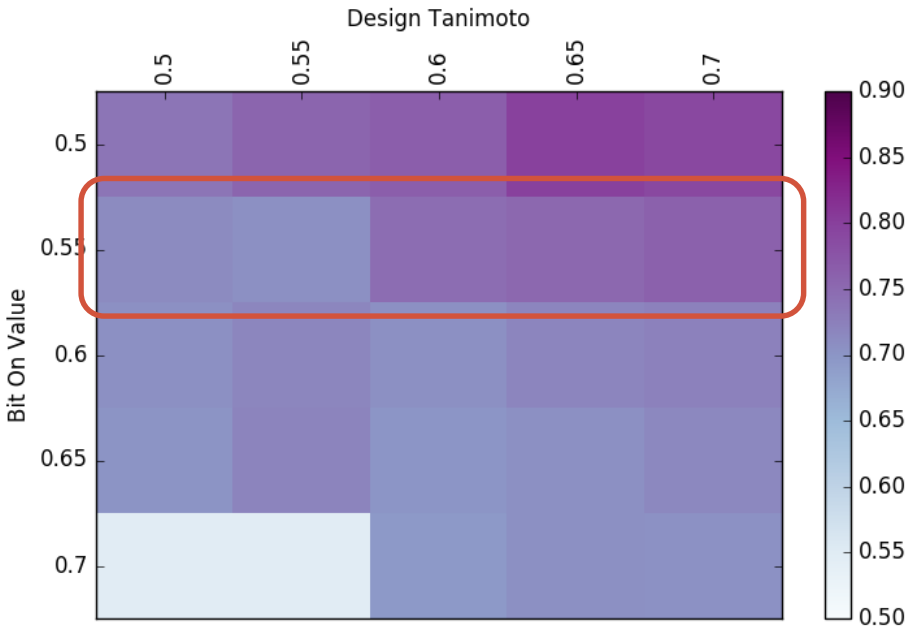
Taylor's set



No of molecules 49 132 351 929 2275



Head's set



No of molecules 49 132 351 929 2275

Conclusions

- We grouped compounds with shared BIOACTIVITY based only on their SHAPES
  - Our best Shape Databases: SD16 (MW 300 – 500, HA 10 – 50)
    - The best settings: DT=0.65 and BOV=0.55
- We have extended our work by using conformations generated from 2D of the test sets molecules
  - We made the Shape Database freely available at https://github.com/LeachResearchGroup/ShapeFingerprints

In future:

- Investigate the applicability of the best Shape Database

Acknowledgments

- Phil Rowe
- Iva Lukac
- Openeye for license
- MedChemica for funding

Thank you for your attention!

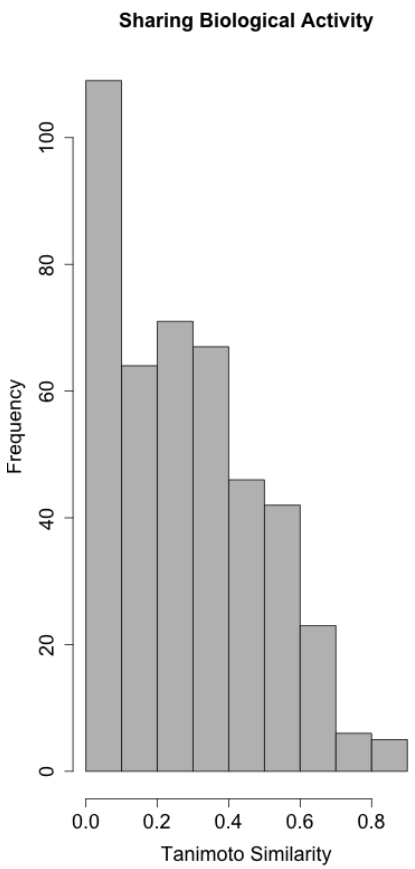
Shape Databases – SD16

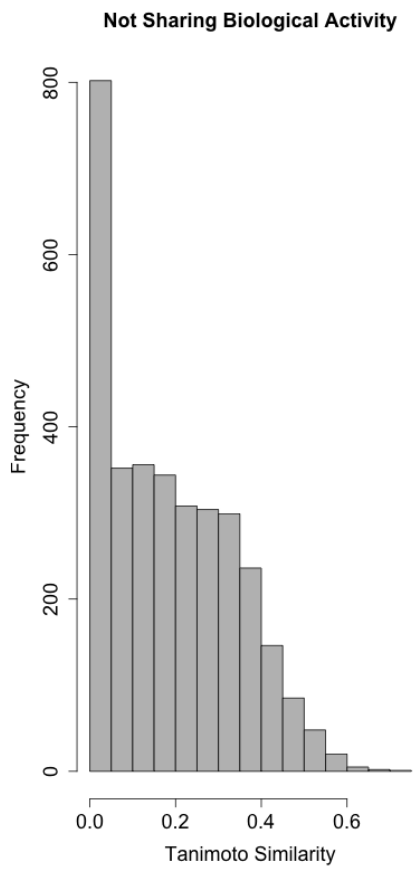
Taylor's set		DT 0.50	0.55	0.60	0.65	0.70
BOV	0.50	0.6104	0.6092	0.6163	0.6091	0.6082
	0.55	0.6354	0.6491	0.6438	0.6272	0.6227
	0.60	0.5268	0.6403	0.6427	0.636	0.6389
	0.65	0.5264	0.5734	0.5904	0.6286	0.6599
	0.70	0.5002	0.5197	0.5052	0.548	0.6298
No of molecules		49	132	351	929	2275

Head's set		DT 0.50	0.55	0.60	0.65	0.70
BOV	0.50	0.738	0.7572	0.765	0.7999	0.792
	0.55	0.7127	0.7078	0.7475	0.7533	0.7614
	0.60	0.7064	0.7181	0.7063	0.72	0.7249
	0.65	0.7031	0.7218	0.7004	0.7063	0.7163
	0.70	0.549	0.549	0.6967	0.7071	0.7054
No of molecules		49	132	351	929	2275

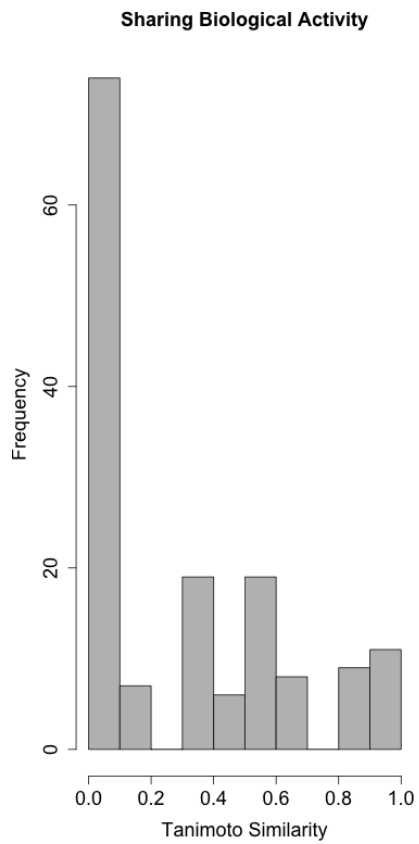
SD16; DT = 0.65; BOV = 0.55

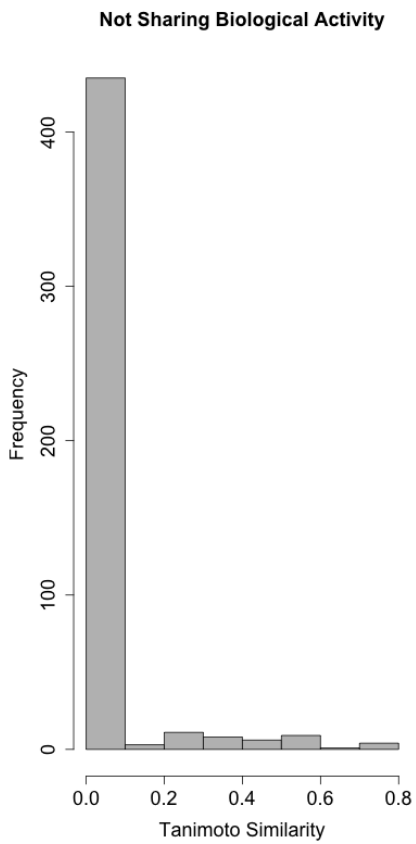
Taylor's set





Head's set





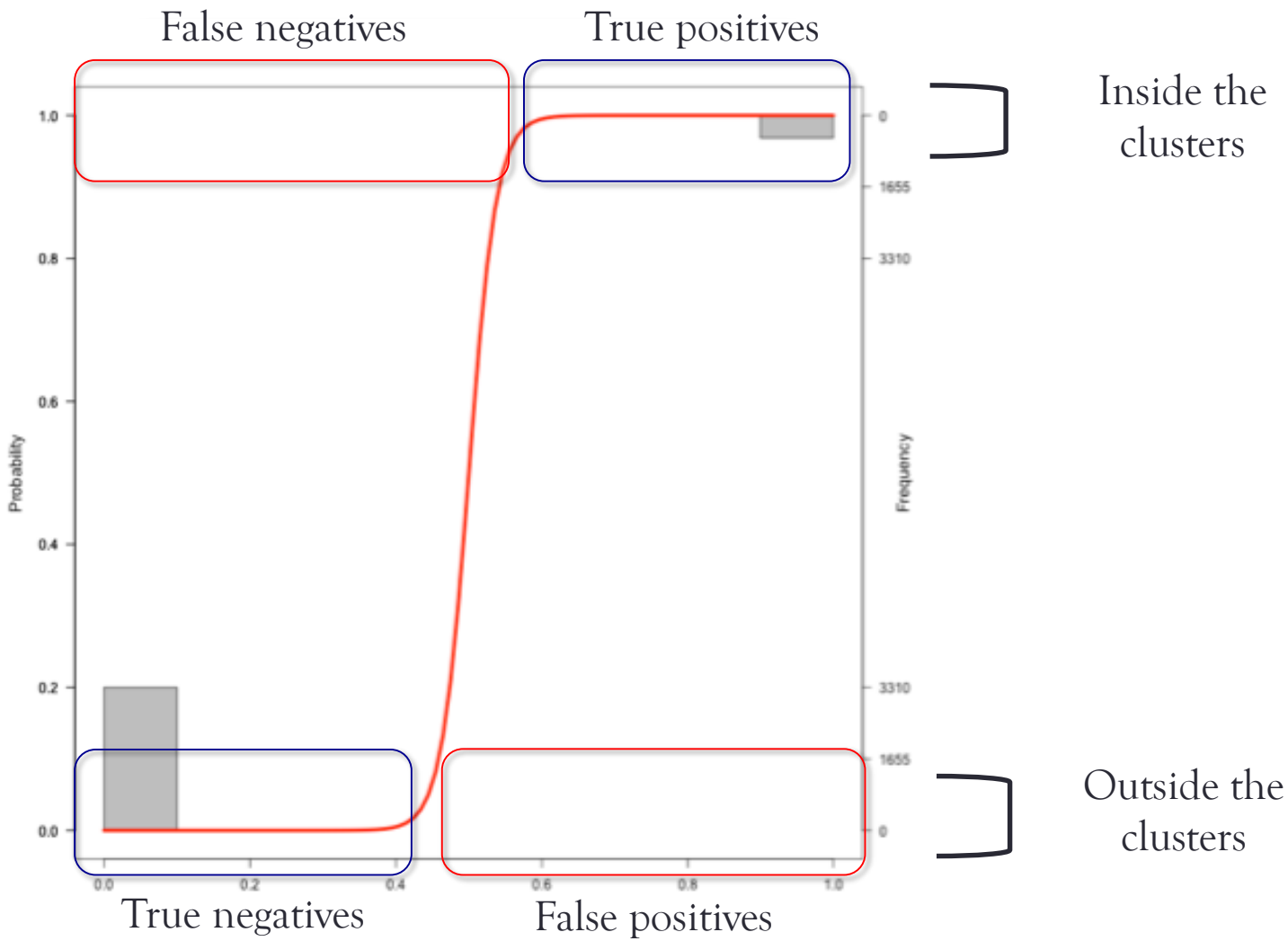
Step by Step

- 1. Selecting dataset as input for Shape Database generation
- 2. Filtering dataset - molecular weight and heavy atoms
- 3. Generating Shape Databases with different user-defined cut-offs (Design Tanimoto)
- 4. Selecting test set
- 5. Generating shape fingerprints for the test set
- 6. Comparing molecules
- 7. Analysis of results

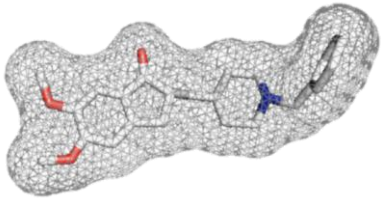
SI - Shape Databases

Dataset	Dataset size	DT = 0.5	DT = 0.7
DB1	244031	139	3560
DB2	169655	134	3567
DB3	92465	126	3482
DB4	41126	123	2814
DB5	16340	109	1628
DB6	97242	134	3502
DB7	27625	115	2466
DB8	3563	68	487
DB9	227691	56	2378
DB10	202905	30	1098
DB11	151566	13	230
DB12	74376	4	34
DB13	241935	116	3378
DB14	218478	43	1565
DB15	158266	16	231

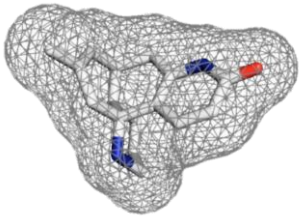
Analysis –Logistic Regression



False negatives

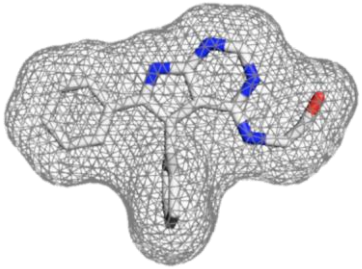


AChE ligand (1eve)

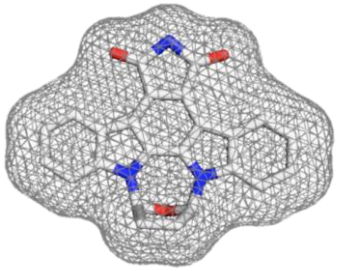


AChE ligand (1gpn)

True positives

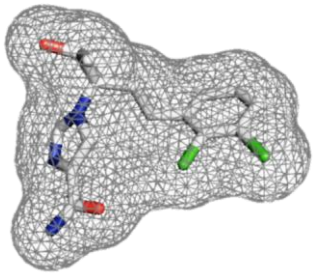


Chk1 ligand (2brm)

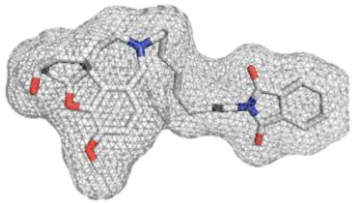


Chk1 ligand (1nvs)

True negatives

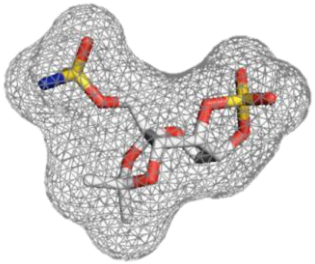


ADA ligand (1v79)

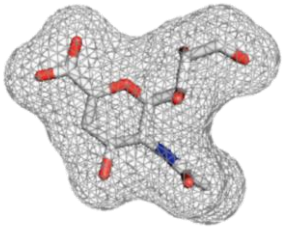


AChE ligand (1w4l)

False positives

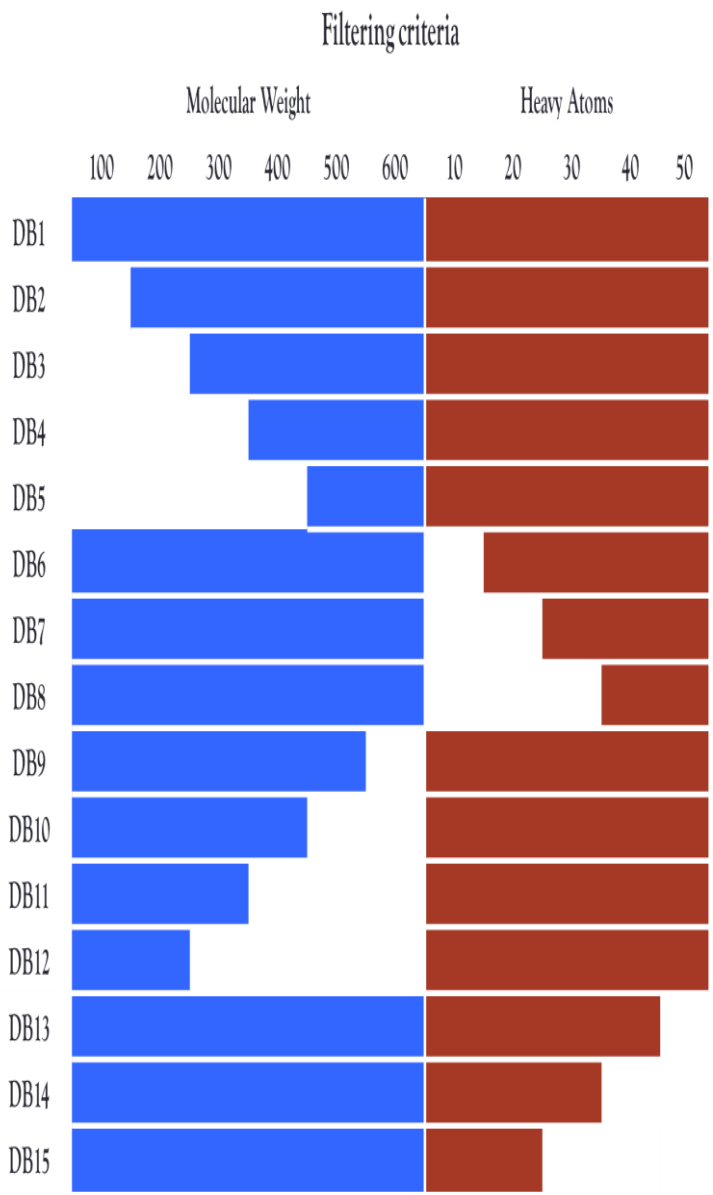


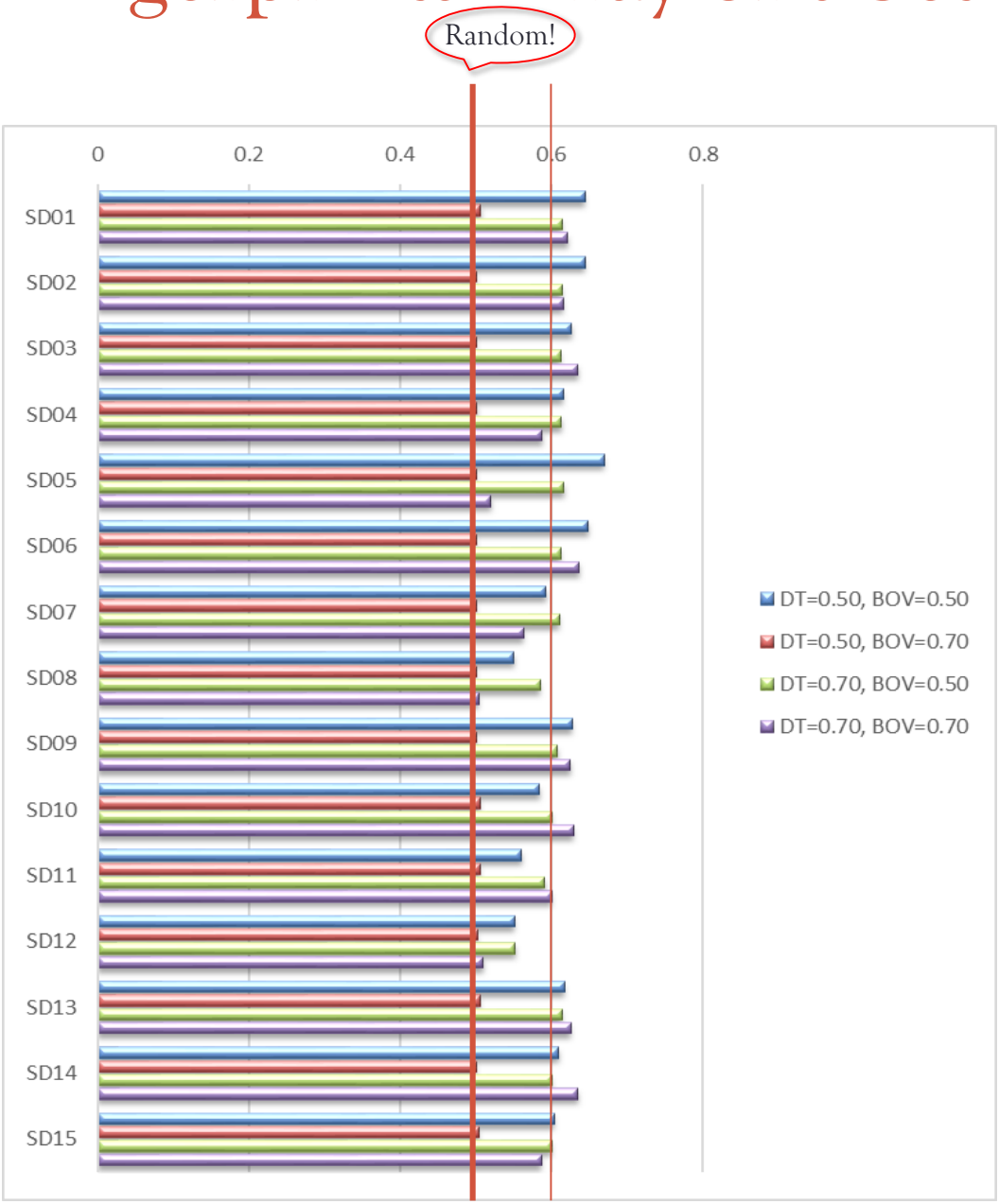
CA ligand (1eou)



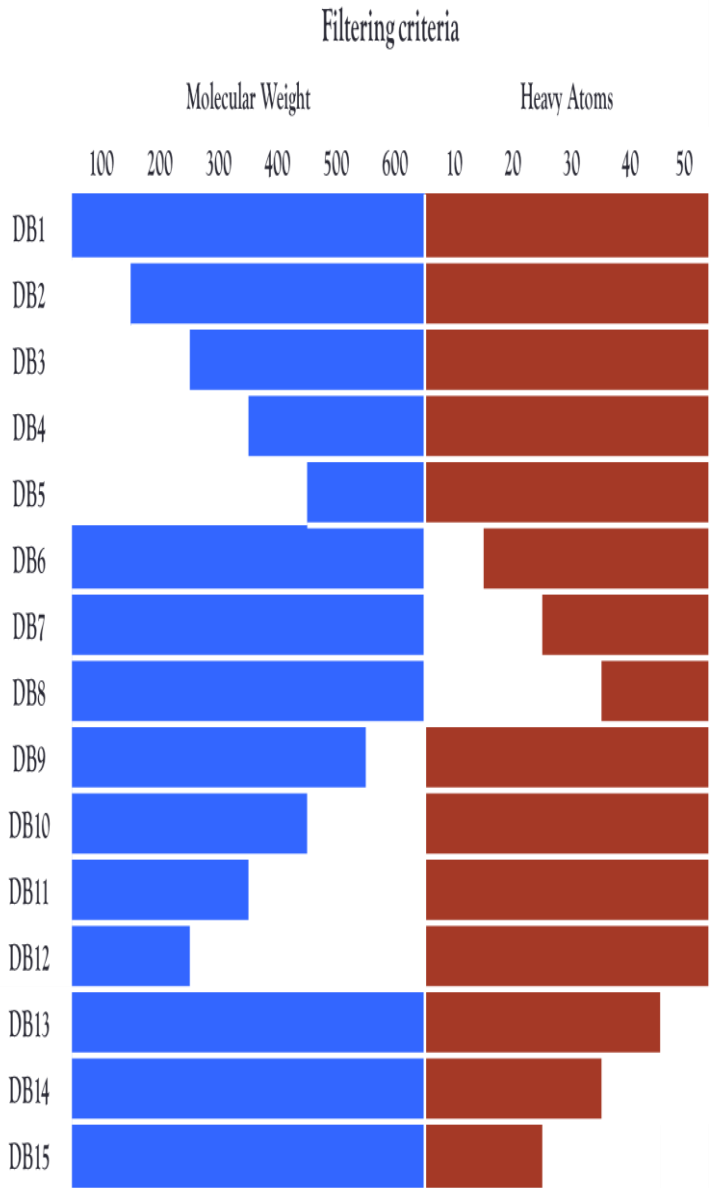
NEU ligand (1nsd)

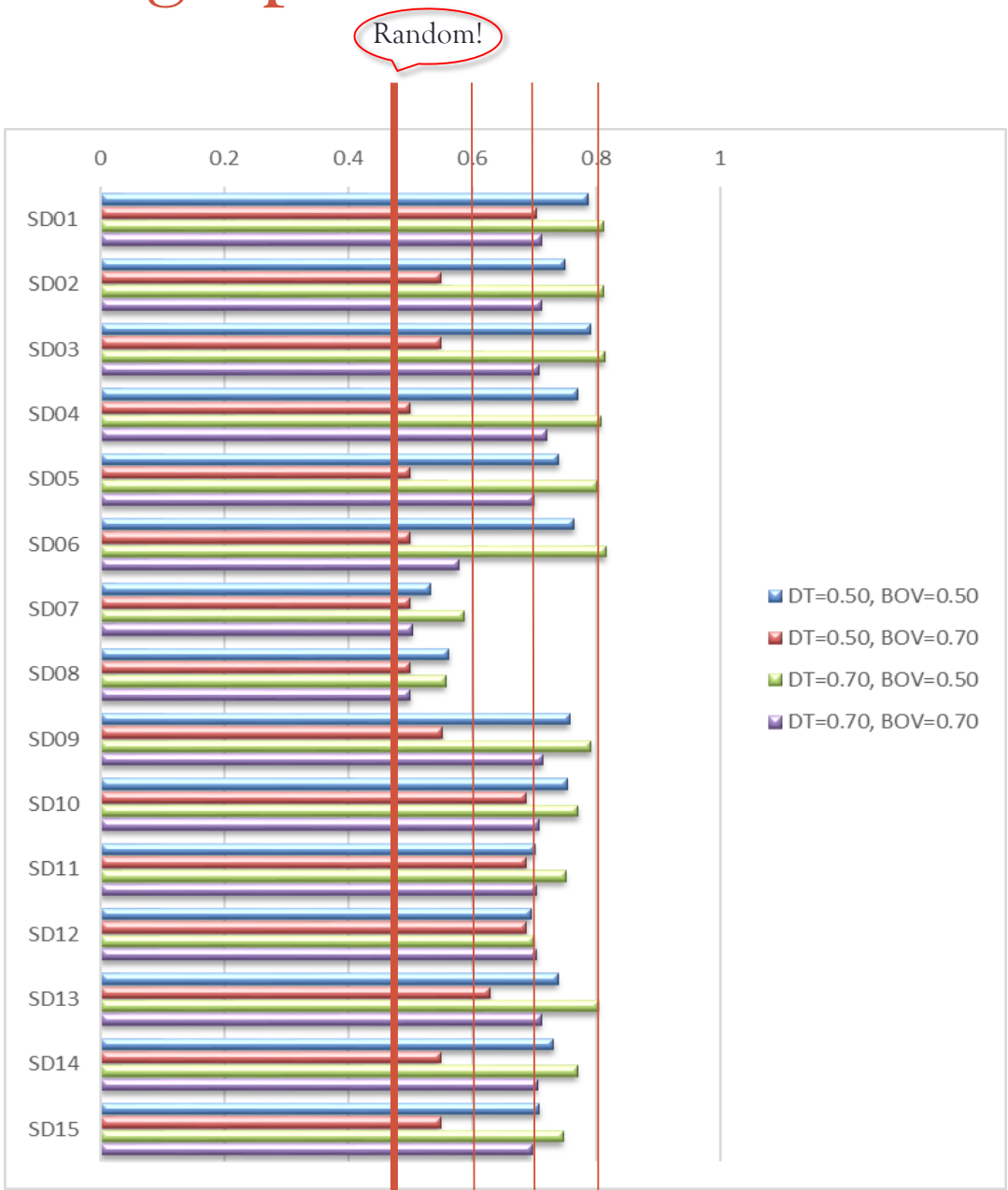
Comparing Shape Fingerprints – Taylor's Set





Comparing Shape Fingerprints – Head's Set





Shape Databases – SD06

Taylor's set		DT 0.50	0.55	0.60	0.65	0.70
BOV	0.50	0.6479	0.6304	0.6184	0.6155	0.6116
	0.55	0.6188	0.6454	0.6272	0.6266	0.6219
	0.60	0.5726	0.5855	0.6196	0.6431	0.6356
	0.65	0.518	0.5379	0.5591	0.6071	0.6482
	0.70	0.5	0.501	0.5084	0.5398	0.6367
No of molecules		134	337	779	1754	3502

Head's set		DT 0.50	0.55	0.60	0.65	0.70
BOV	0.50	0.7648	0.7969	0.8303	0.8229	0.8158
	0.55	0.6126	0.6743	0.7195	0.7955	0.8101
	0.60	0.5098	0.5278	0.5498	0.7523	0.7722
	0.65	0.5033	0.5012	0.5196	0.7185	0.7251
	0.70	0.5	0.5	0.5033	0.5752	0.5784
No of molecules		134	337	779	1754	3502