Cheminformatics

- 2-week period
- Basic principles of cheminformatics and its terminology
- 2 home coding assignments
  - 2D structures representation (5 pts)
  - QSAR machine learning problem (5 pts)
- Prerequisites
  - Python + Jupyter notebooks
  - Familiarity with how ML algorithms run

Useful links

- https://training.galaxyproject.org/training-material/topics/computational-chemistry/tutorials/covid19-docking/tutorial.html
- https://chem.libretexts.org/Courses/Intercollegiate_Courses/Cheminformatics_OLCC_(2019)
- https://www.rdkit.org/docs/GettingStartedInPython.html
- Full course (CZ version): http://www.chemicke-listy.cz/ojs3/index.php/chemicke-listy/issue/view/250

The challenge

Chemical space

~10^{36} drug-like compounds [1]

Biological space

~10^{4} human proteins [2]

Drug discovery

10^{17} sec - the age of the Universe

(1) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A., *J Comput Aided Mol Des* **2013**, 27, 675-679.
(2) http://www.uniprot.otg

Vastness of chemical space

real datasets

SciFinder® A CAS SOLUTION — ~ 160 M compounds

REAXYS® — ~ 105 M compounds

Commercial

PubChem — ~ 102 M compounds — Free

ZINC — up to 1 B commercially available compounds

virtually enumerated dataset

GDB-17 — 166 B compounds = 1.66×10^{11}

estimated size of drug-like chemical space

10^{36} compounds

Screening

High-throughput screening

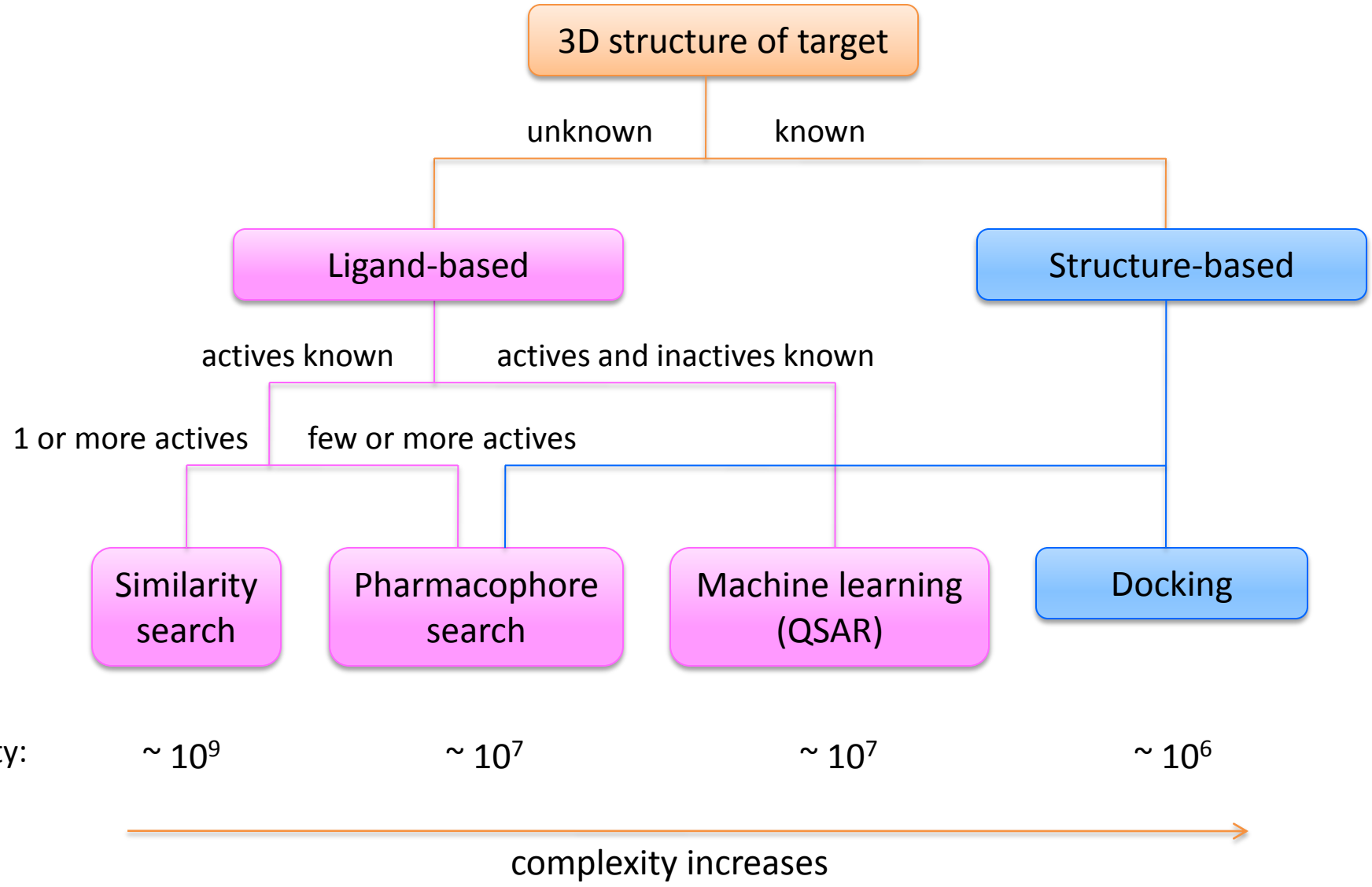
up to 10^6 of compounds can be tested

- expensive
- not all targets are suitable for HTS

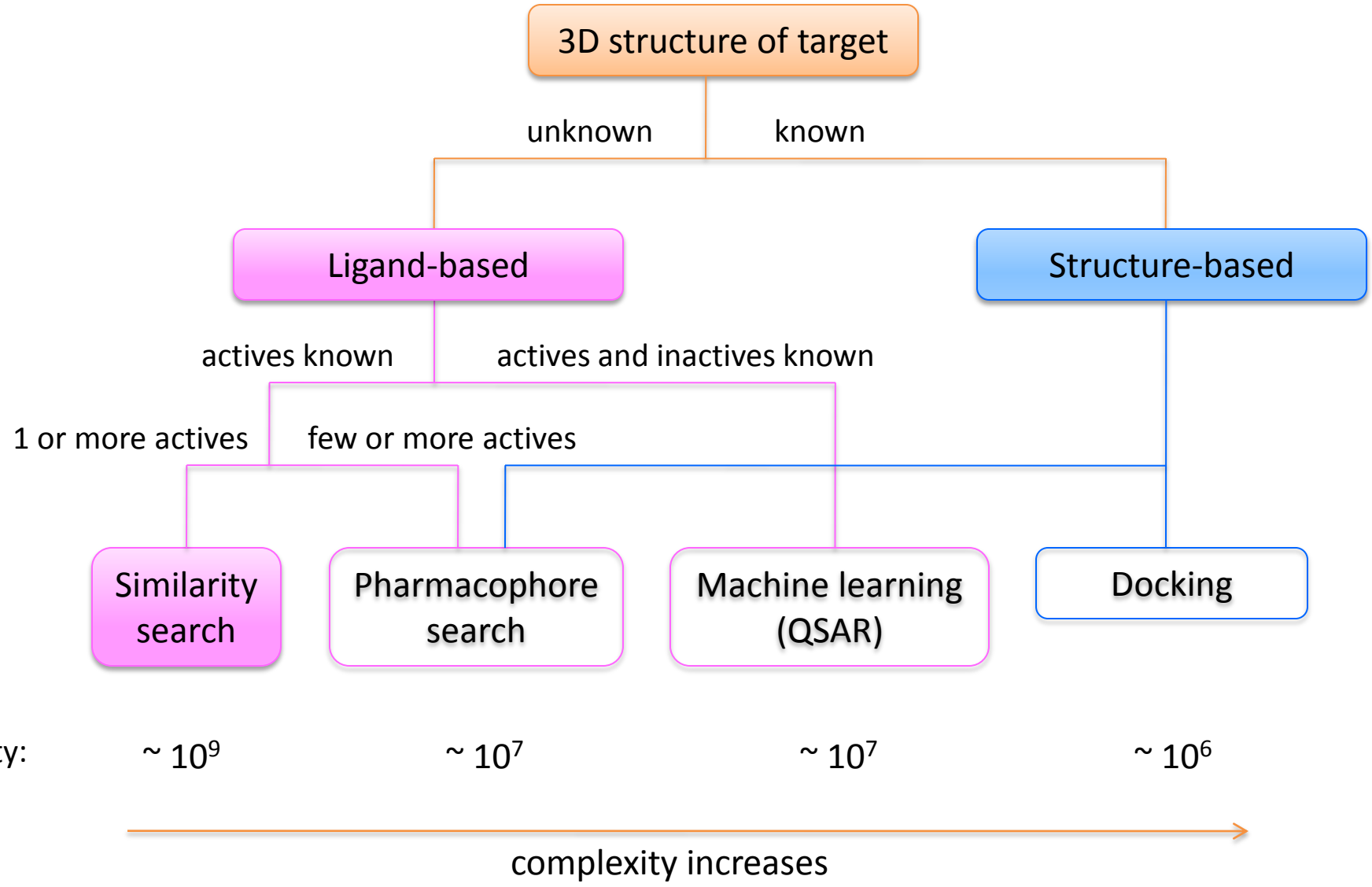
Virtual screening

up to 10^9 of compounds can be tested

Virtual screening methods

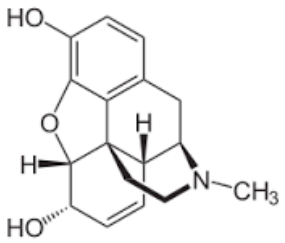


Similarity search

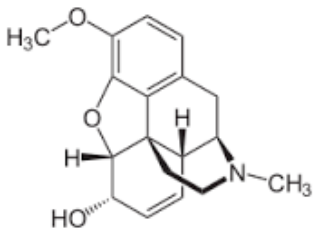


Similarity principle

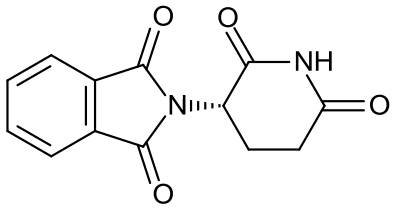
Similar compounds have similar properties



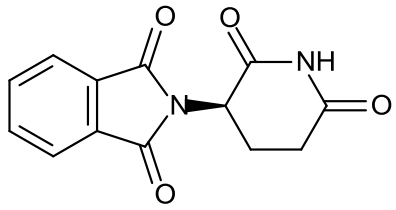
morphine



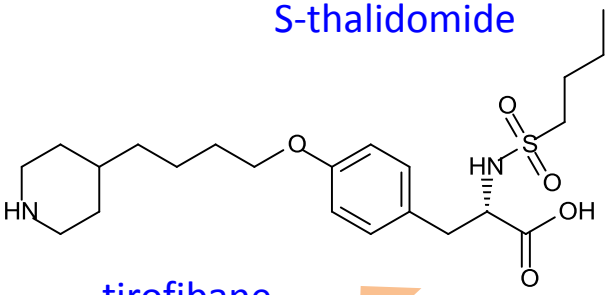
codeine



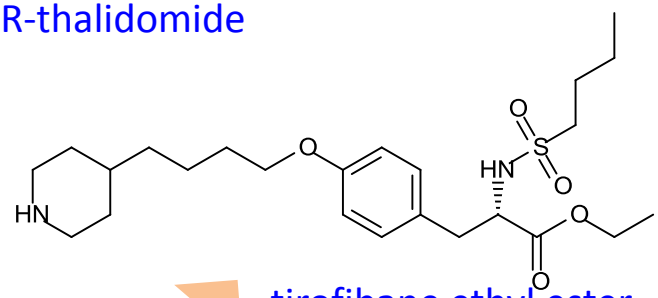
S-thalidomide



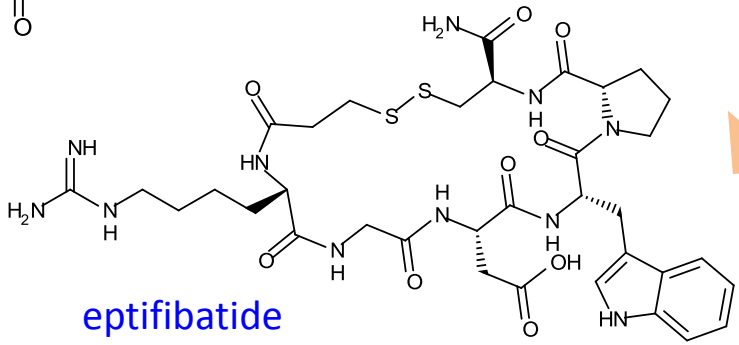
R-thalidomide



tirofibane



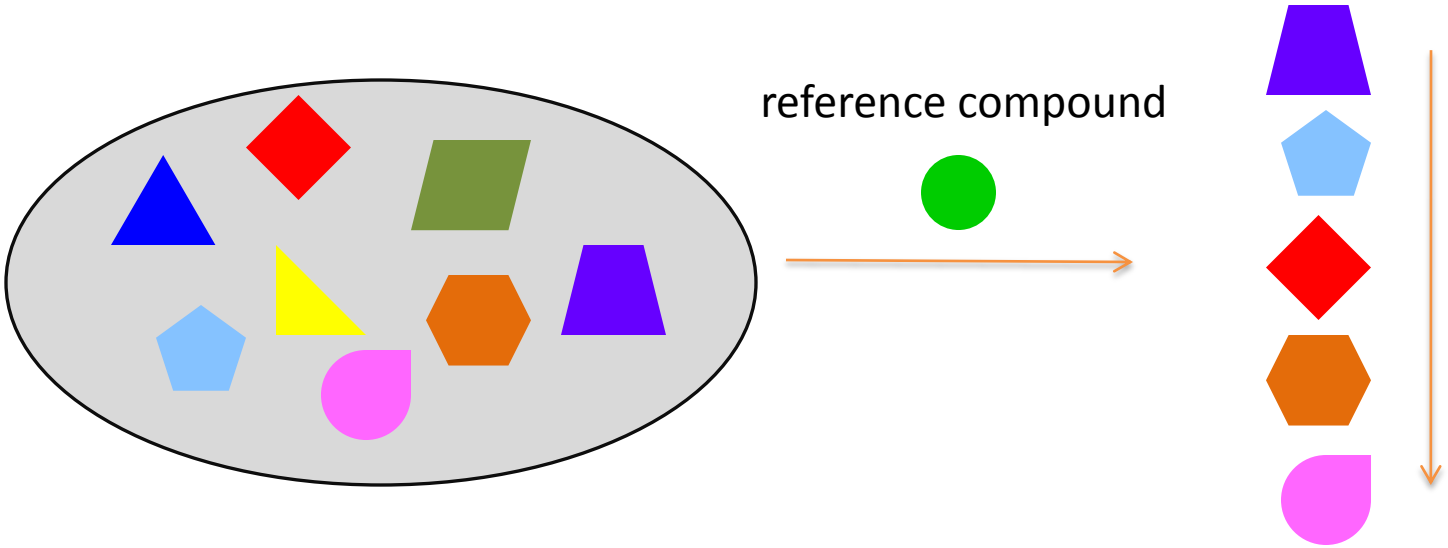
tirofibane ethyl ester



eptifibatide

Similarity search

Rank and select similar compounds



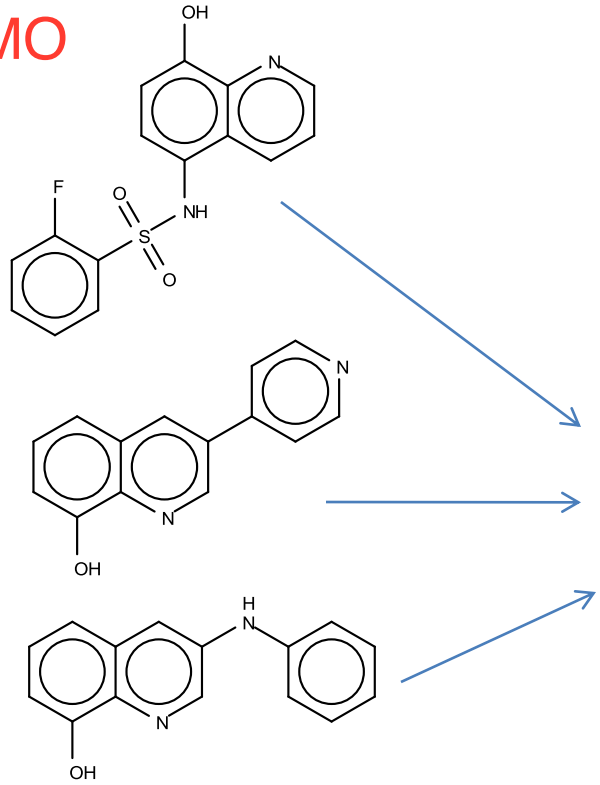
Ranking of compounds: example

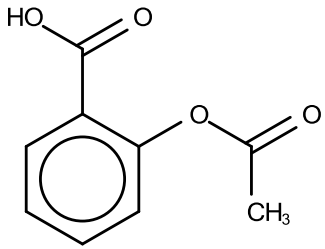
Structure representation spanning tree

- structural keys DEMO SMILES, InChI
- fingerprints

Similarity measure

- Tanimoto DEMO
- Dice
- Euclidian
- …





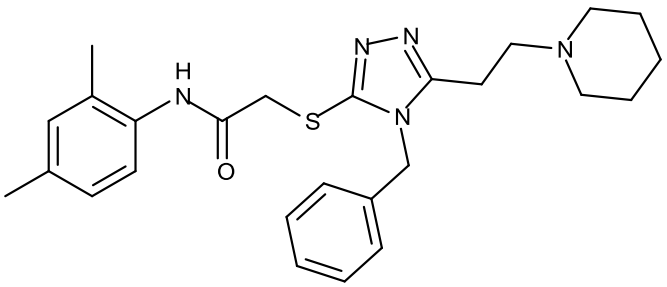
Dice		
Atom pairs	ECFP4	FCFP4
0.327 (3)	0.219 (2)	0.233 (1)
0.364 (1)	0.185 (3)	0.170 (2)
0.333 (2)	0.291 (1)	0.125 (3)

*binary fingerprints calculated with RDKit

Similarity search output depends on descriptors and similarity measure selected

Similarity search: example

agonists of CCR5



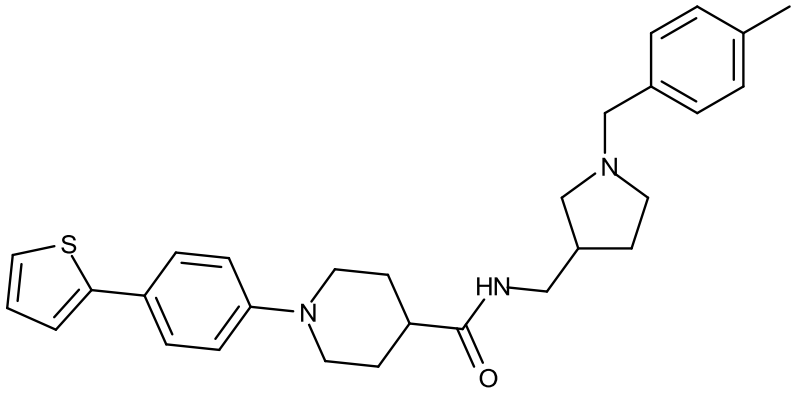
IC_{50} = 17 µM

60 000 compounds

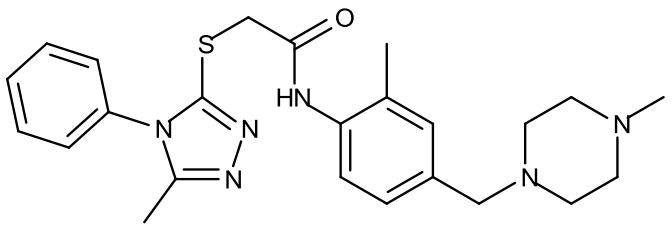
FCFP4

100 compounds

purchased & tested



IC_{50} = 5.8 µM



IC_{50} = 14.1 µM

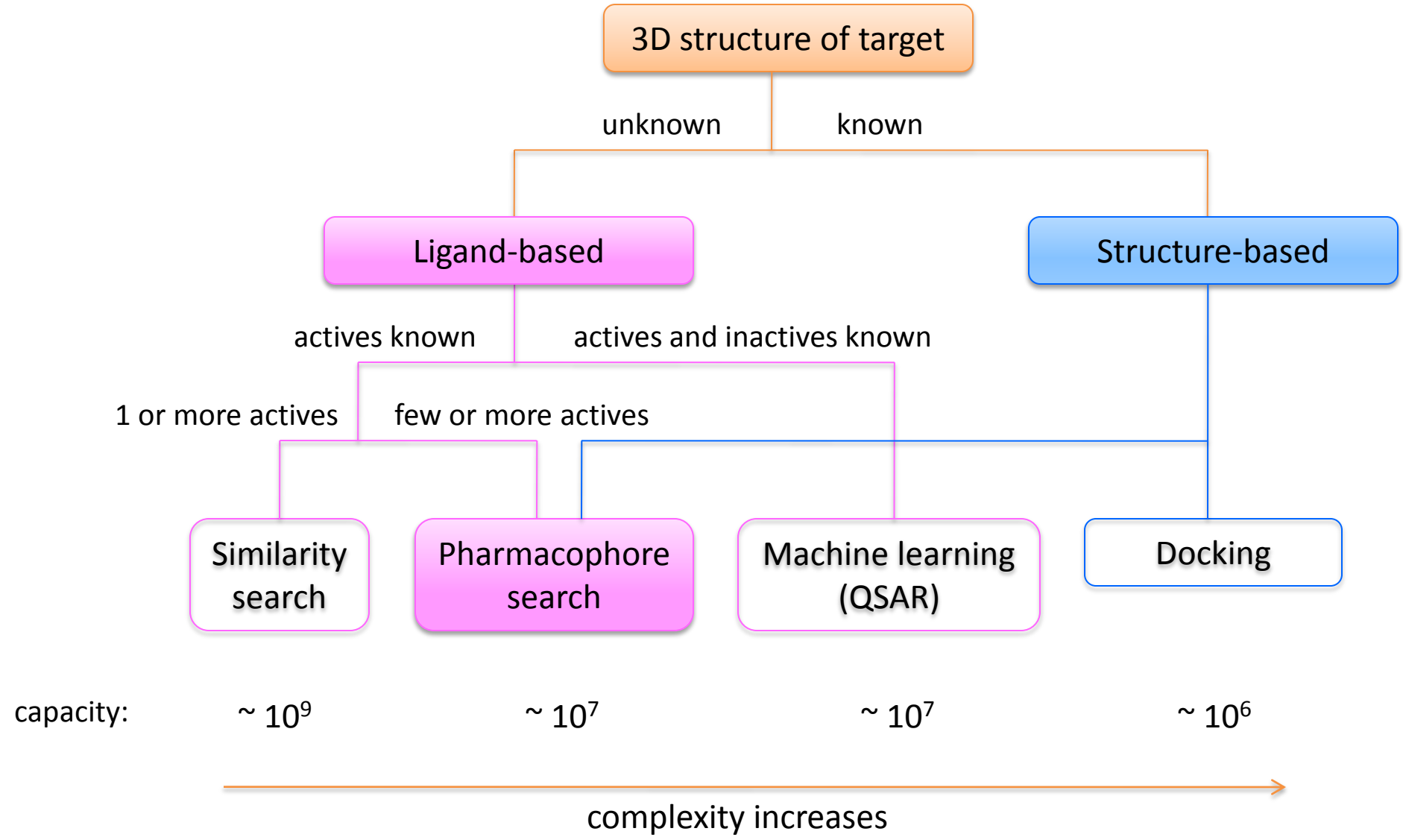
Kellenberger, E., et al., Identification of nonpeptide CCR5 receptor agonists by structure-based virtual screening. *Journal of Medicinal Chemistry* **2007**, 50, 1294–1303.

Similarity search: conclusion

+ Little information is required to start searching
+ Different chemotypes can be retrieved
+ Ultra fast screening
+ Scaffold hopping

- Hits will share common substructures with reference structures that may reduce their patentability
- Results depend on chosen descriptors and similarity measure
- Chemical similarity is not always followed by biological one

Pharmacophore search

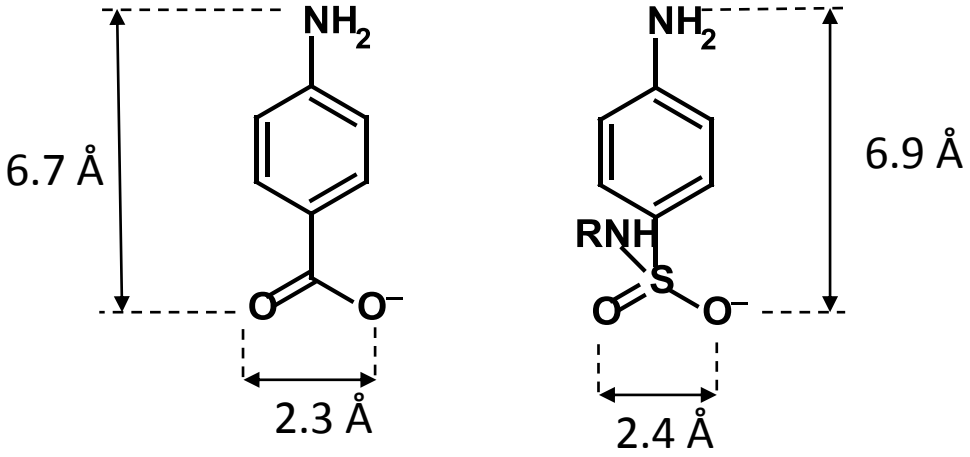


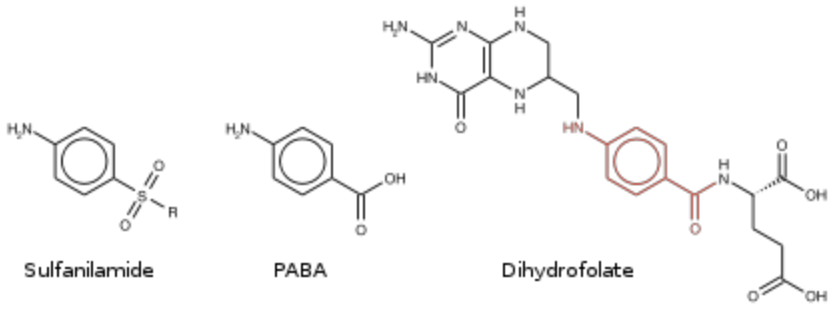
Pharmacophore definition

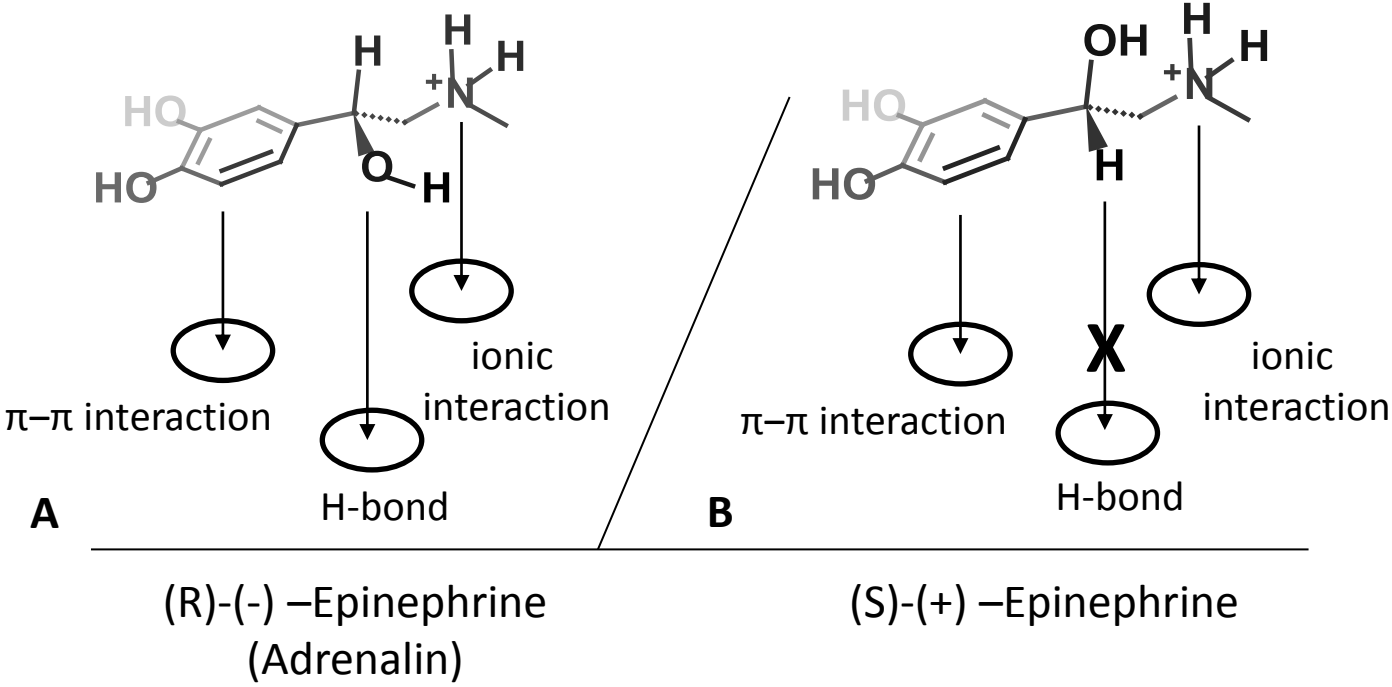
A **pharmacophore** is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interaction with a specific biological target structure and to trigger (or block) its biological response.

Annu. Rep. Med. Chem. 1998, 33, 385–395

Early pharmacophore hypotheses

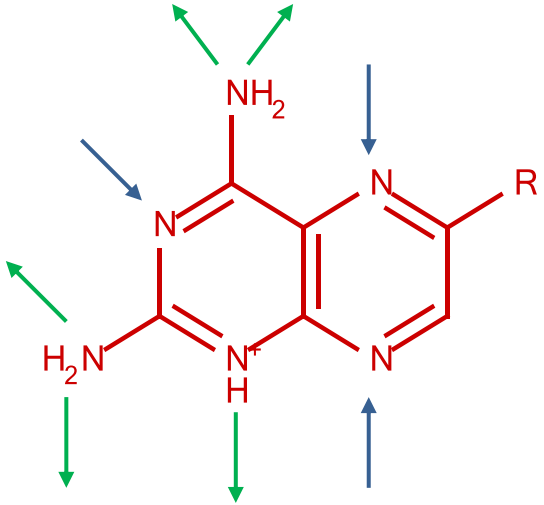




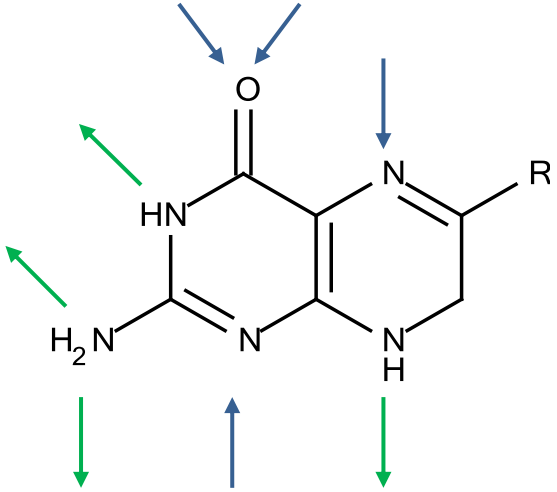


Atom- and pharmacophore-based alignment

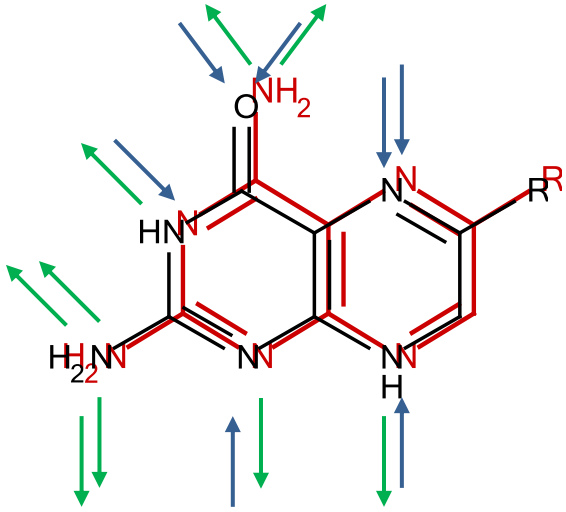
Methotrexate



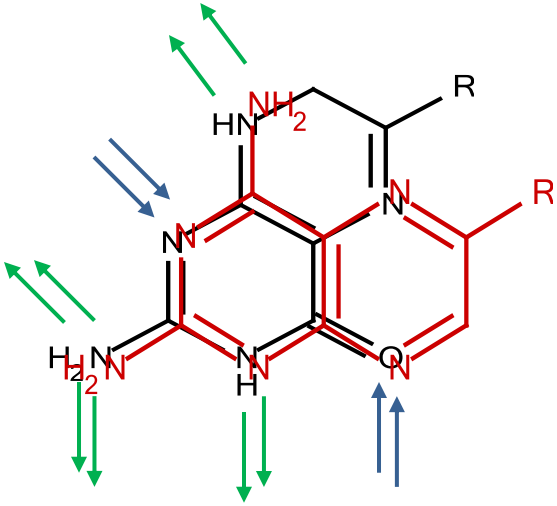
Dihydrofolate



Hydrogen bonding patterns



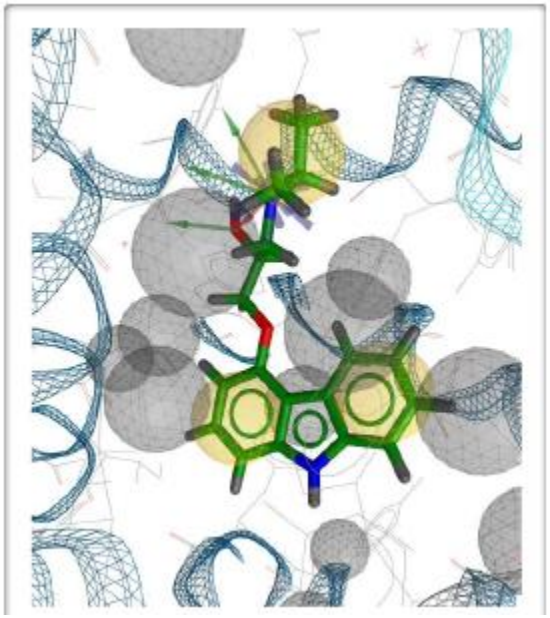
Atom-based alignment

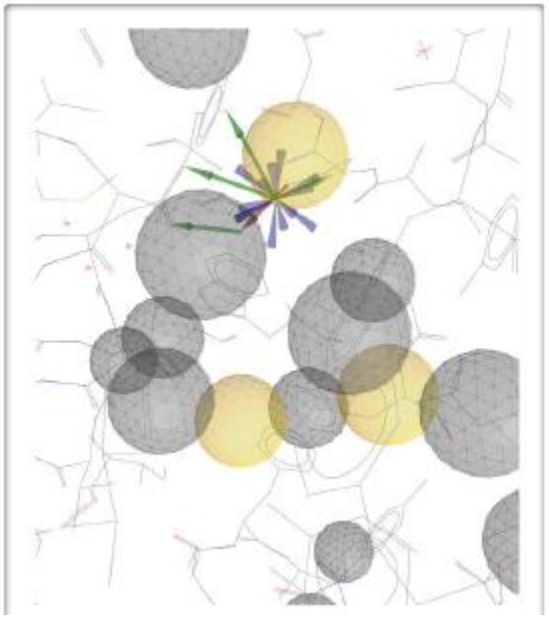


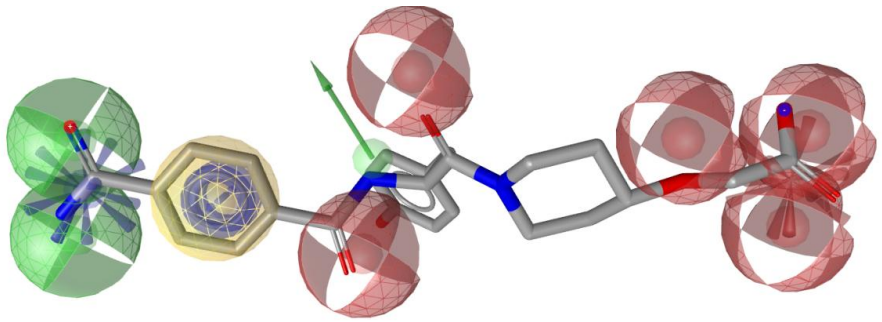
Pharmacophore alignment

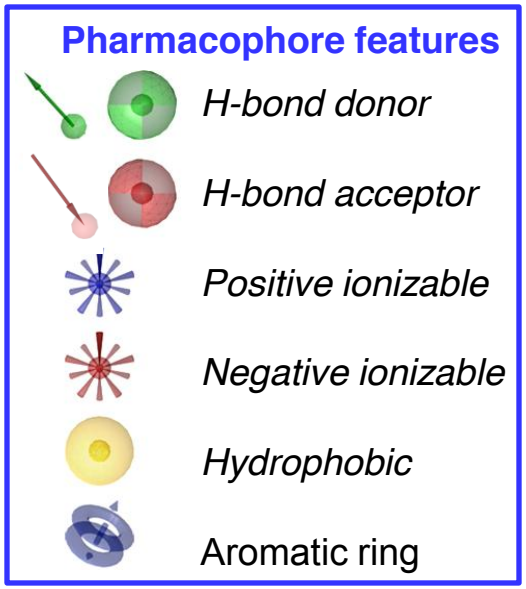
Feature-based pharmacophore models

Features: Electrostatic interactions, H-bonding, aromatic interactions, hydrophobic regions, coordination to metal ions ...



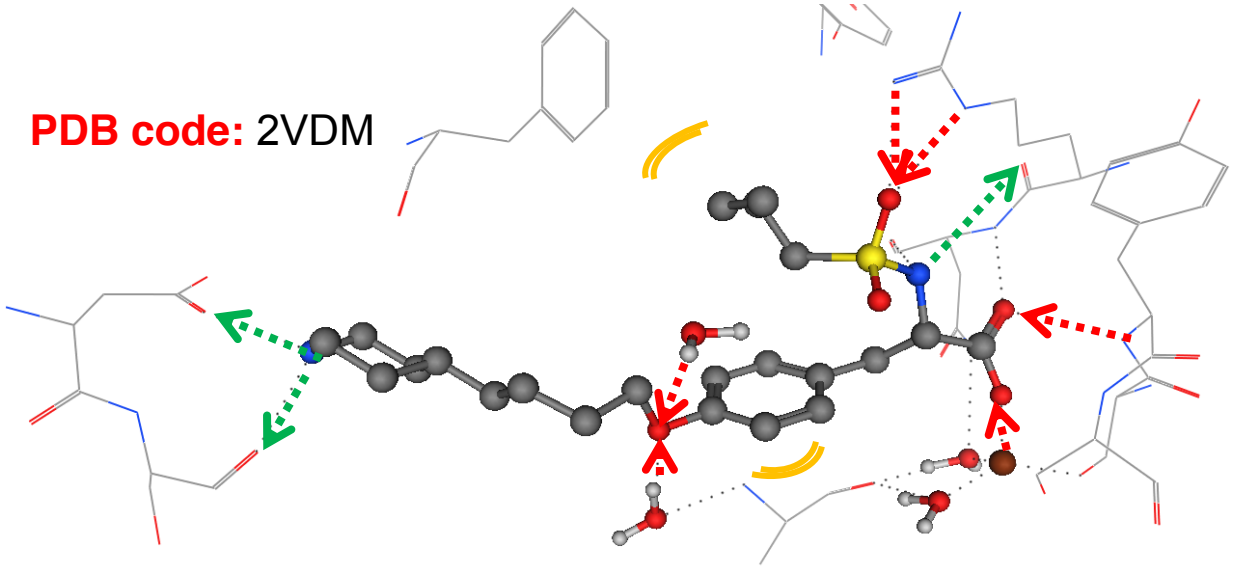






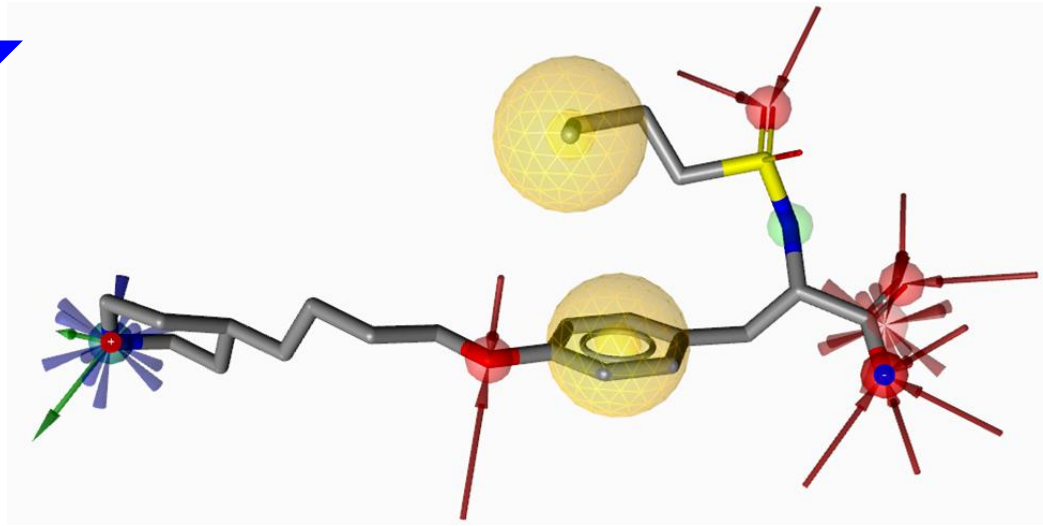
Feature-based pharmacophores (LigandScout)

PDB code: 2VDM



H-bonds formed by the ligand

Hydrophobic interaction



Pharmacophore features

H-bond donor

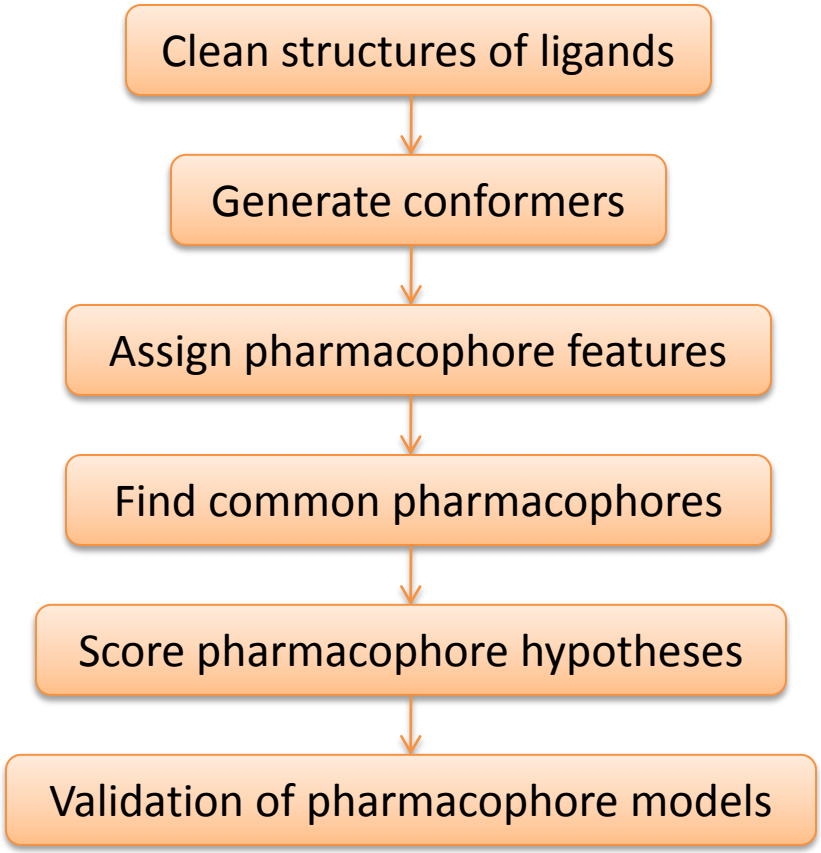
H-bond acceptor

Positive ionizable

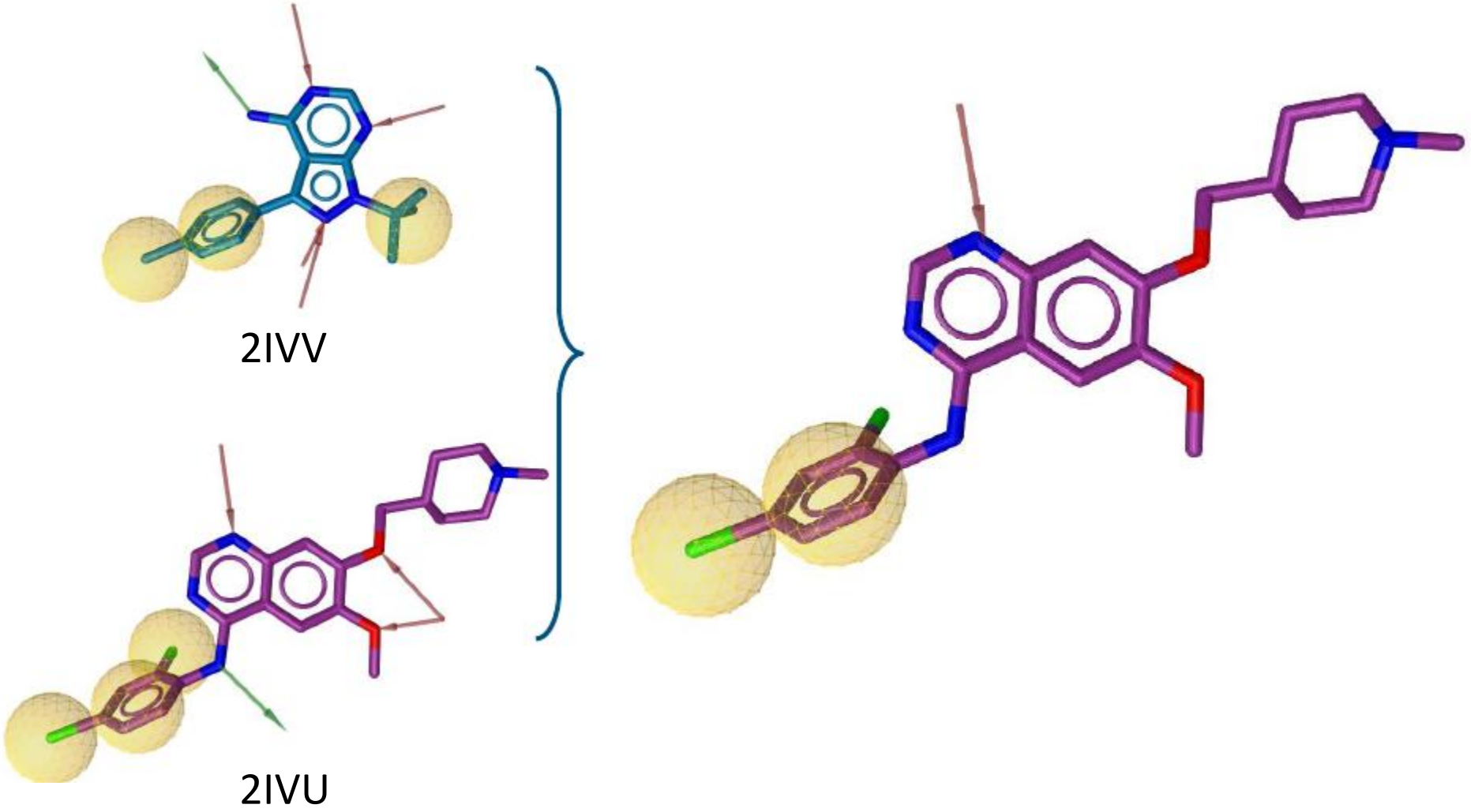
Negative ionizable

Hydrophobic

Typical ligand-based pharmacophore modeling workflow



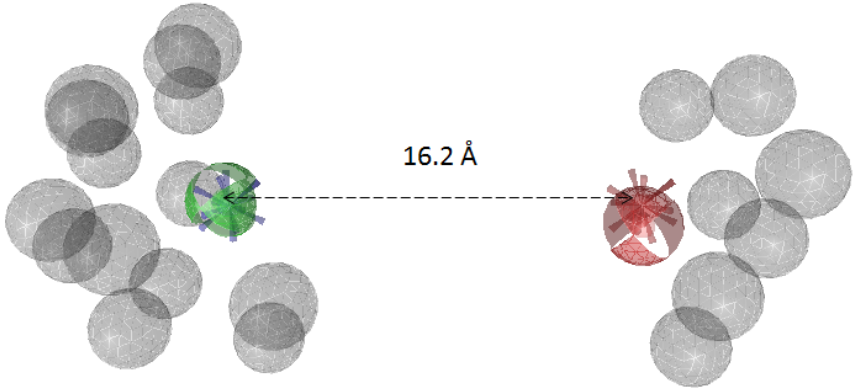
Shared consensus pharmacophore (LigandScout)



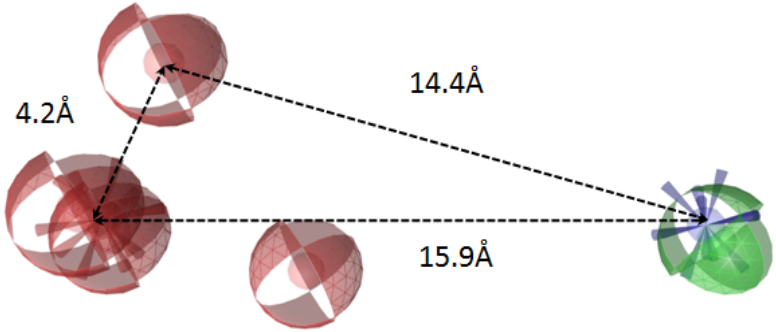
RET Kinase Inhibitors

Pharmacophore examples

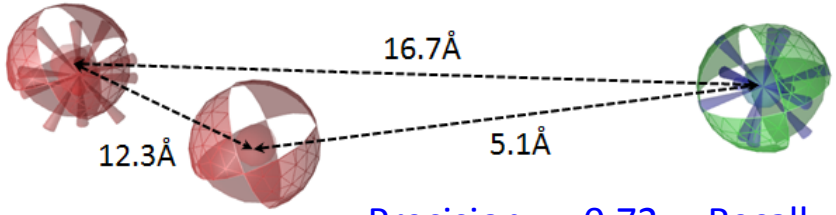
Shared model on 83 antagonists of fibrinogen receptor



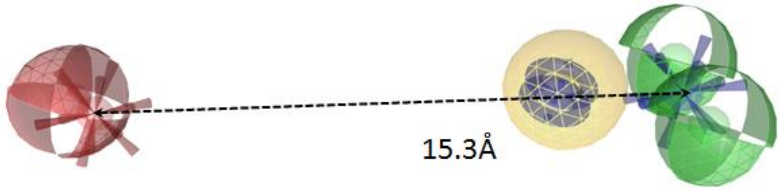
Pharmacophore models obtained for clusters of compounds



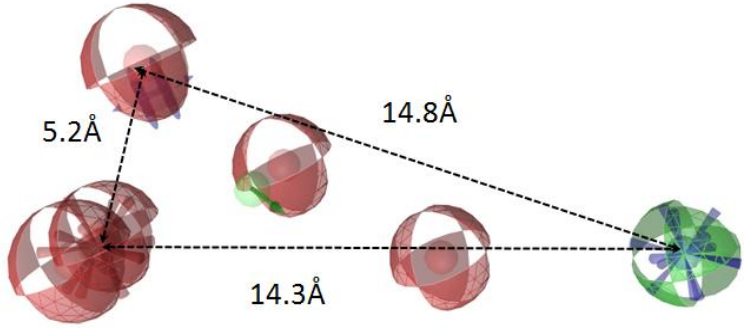
Precision: 0.77 Recall: 0.27



Precision: 0.72 Recall: 0.77



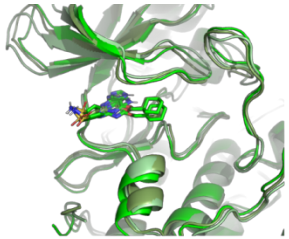
Precision: 0.67 Recall: 0.29



Precision: 0.90 Recall: 0.04

Polishchuk, P. G. et al., *Journal of Medicinal Chemistry* **2015**, 58, 7681-7694.

MD pharmacophores



MD snapshot timeline

A	B	C	D	E	F	G	H	I	J	K
h6rhf..3o	dkrti..21	dkrti..21	341kq..pb	dkrti..21	341kq..pb	91m9b..1a	e2e4k..a8	91m9b..1a	91m9b..1a	kiqp1..dp

MD pharmacophores

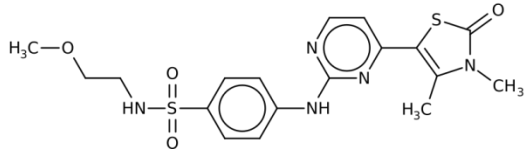
3D pharmacophore hashes

representative pharmacophore models

A	B	D	G	H	K

compound conformers

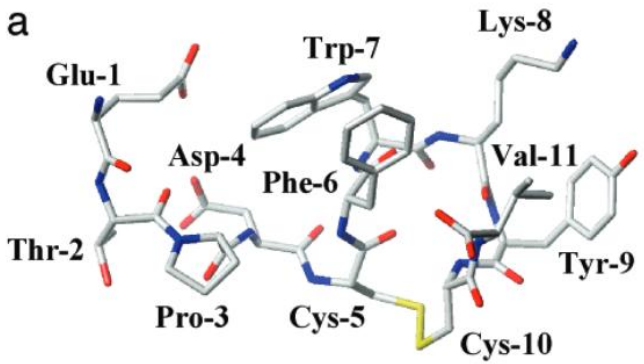
1	2	3	4	5



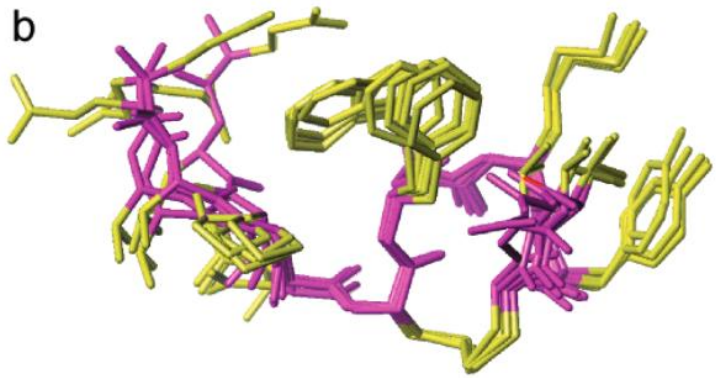
score	scoring approach
$\frac{3}{6} = \mathbf{0.5}$	Common HIts Approach (CHA)
$\frac{4}{5} = \mathbf{0.8}$	Conformers coverage Approach (CCA)

Polishchuk, P. et al. Virtual Screening Using Pharmacophore Models Retrieved from Molecular Dynamic Simulations. *International Journal of Molecular Sciences* **2019**, 20, (23), 5834.

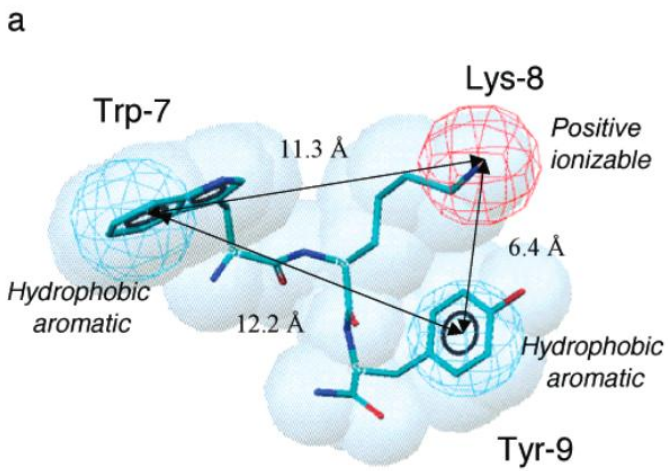
Pharmacophore example



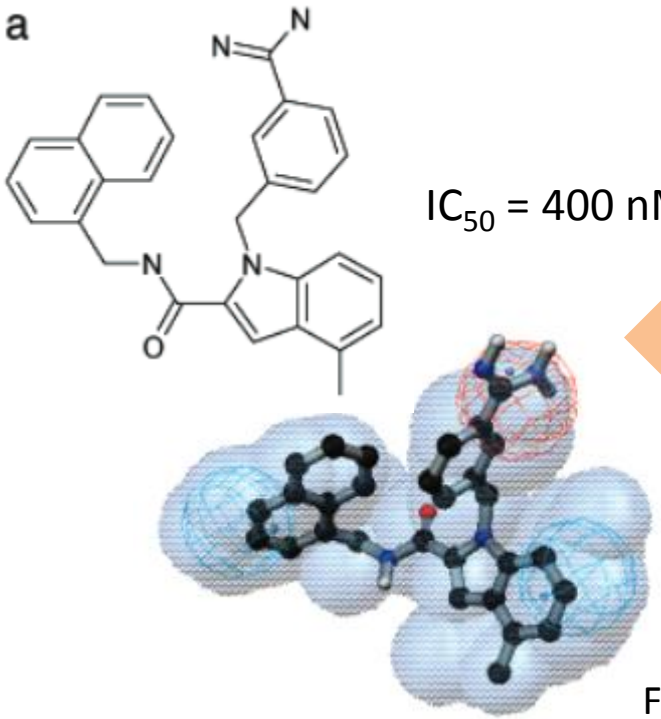
Ala scan
NMR
MD



Urotensin II - ETPDc[CFWKYCV]
potent vasoconstrictor



500 hits



IC_{50} = 400 nM

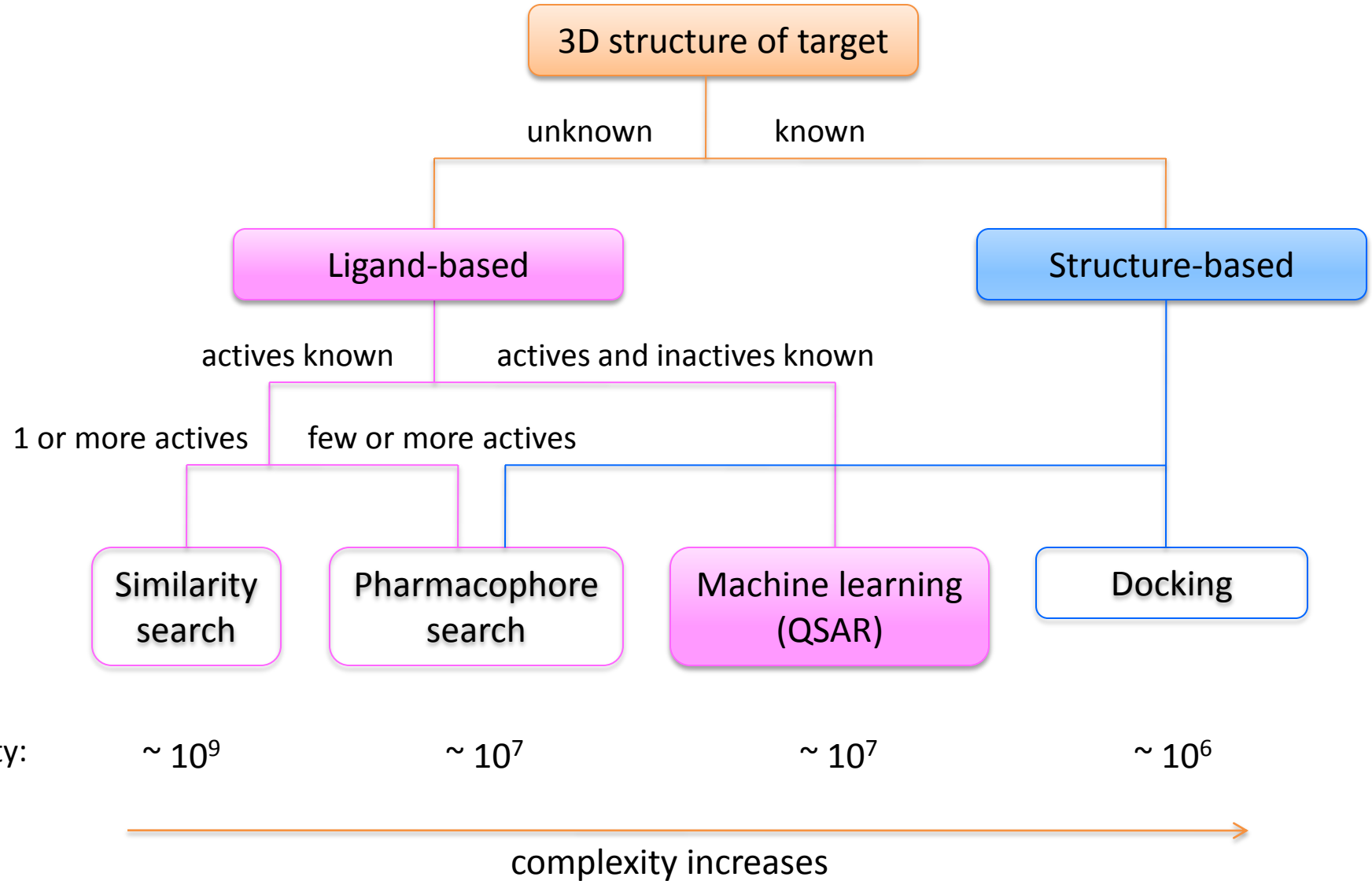
Flohr S. et al., J. Med. Chem., **2002**, 45 (9), pp 1799–1805

Pharmacophores: conclusion

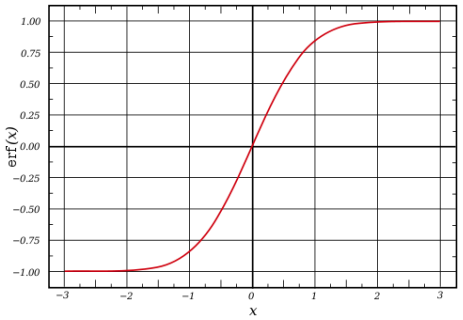
+ Universal representation of binding pattern
+ Qualitative output
+ Very fast screening
+ Scaffold hopping

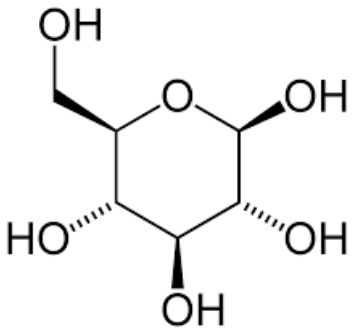
- Structure-based models can be very specific
- Ligand-based models depend on conformational sampling

Machine learning (QSAR)

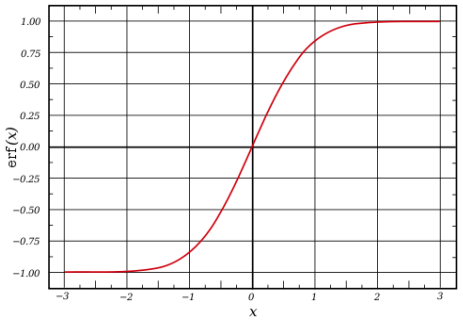


Modeling of compounds properties





Activity = F(structure)



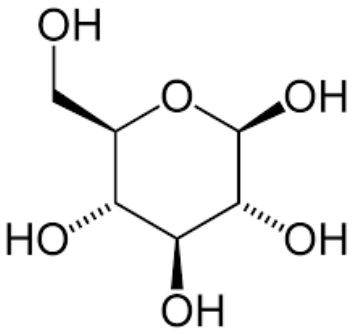
X_1	X_2	X_3	X_4	X_5	X_6	...	X_N
1	0	9	0	11	1	...	1
4	0	1	0	0	0	...	1
0	0	0	0	0	4	...	6
0	2	3	6	0	0	...	3
...	...	...	...	...	...	...	...
4	0	0	0	1	2	...	1

Activity = M(E(structure))

M – mapping function
E – encoding function

QSAR modeling workflow

Structure



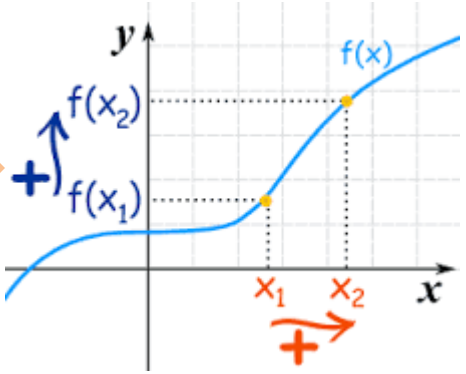
Descriptors (features)

X_1	X_2	X_3	X_4	X_5	X_6	...	X_N
1	0	9	0	11	1	...	1
4	0	1	0	0	0	...	1
0	0	0	0	0	4	...	6
0	2	3	6	0	0	...	3
...	...	...	...	...	...	...	...
4	0	0	0	1	2	...	1

End-point values

Y
1.1
1.4
6.8
3.0
...
1.5

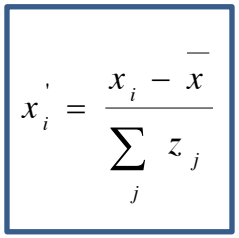
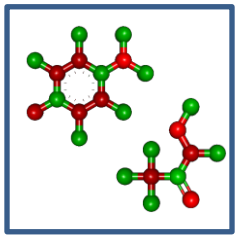
Model



Encoding
(represent structure with numerical features)

Mapping
(machine learning)

Overall QSAR workflow

Input data	Preprocessing	Feature engineering	Model training	Model validation	Interpretation
					
Bioassays	*Data normalization & curation*	*Feature selection*	*Classification*	*Cross-validation*	
Databases	*Feature extraction*	*Feature combination*	*Regression*	*Bootstrap*	
			Clustering	*Test set*	
				Applicability Domain	

OECD principles for the validation, for regulatory purposes, of (Q)SAR models

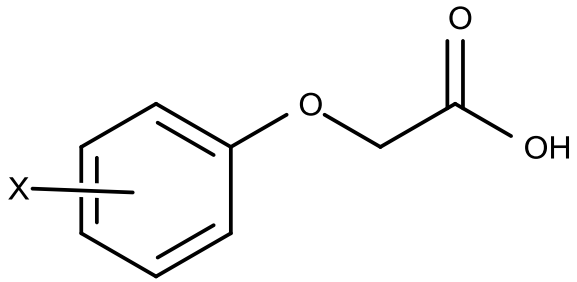
1) a defined endpoint
2) an unambiguous algorithm
3) a defined domain of applicability
4) appropriate measures of goodness-of–fit, robustness and predictivity
5) a mechanistic interpretation, if possible

QSAR model building

plant growth inhibition activity of phenoxyacetic acids

Hansch equation

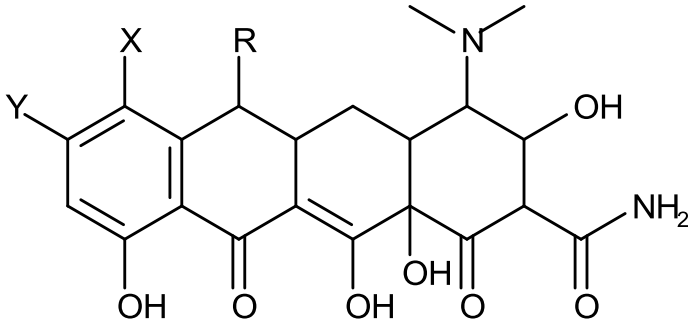
$$1/C = 4.08\pi - 2.14\pi^2 + 2.78\sigma + 3.38$$



$\pi = \log P_X - \log P_H$

σ - Hammet constant

Free-Wilson models



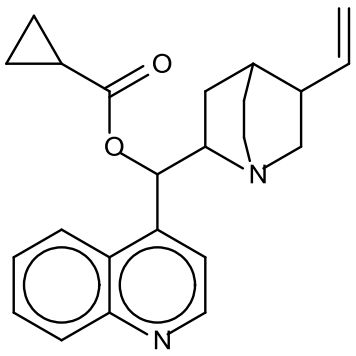
Inhibition activity of compounds against *Staphylococcus aureus*

R is H or CH_3;
X is Br, Cl, NO_2 and
Y is NO_2, NH_2, $NHC(=O)CH_3$

$$Act = 75R_H - 112R_{CH3} + 84X_{Cl} - 16X_{Br} - 26X_{NO2} + 123Y_{NH2} + 18Y_{NHC(=O)CH3} - 218Y_{NO2}$$

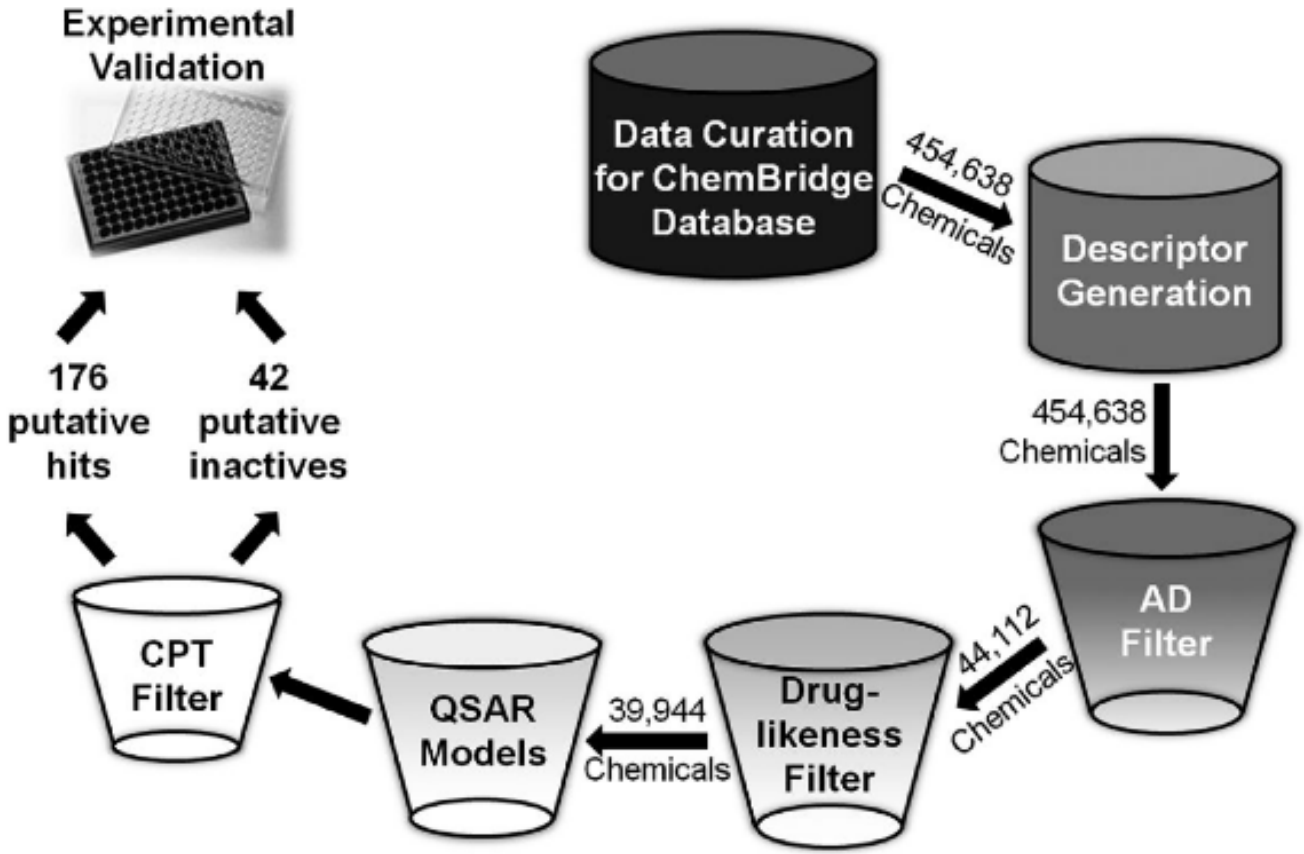
QSAR: example

Antimalarial activity



EC_{50} = 95 nM

7 hits, EC_{50} < 2μM



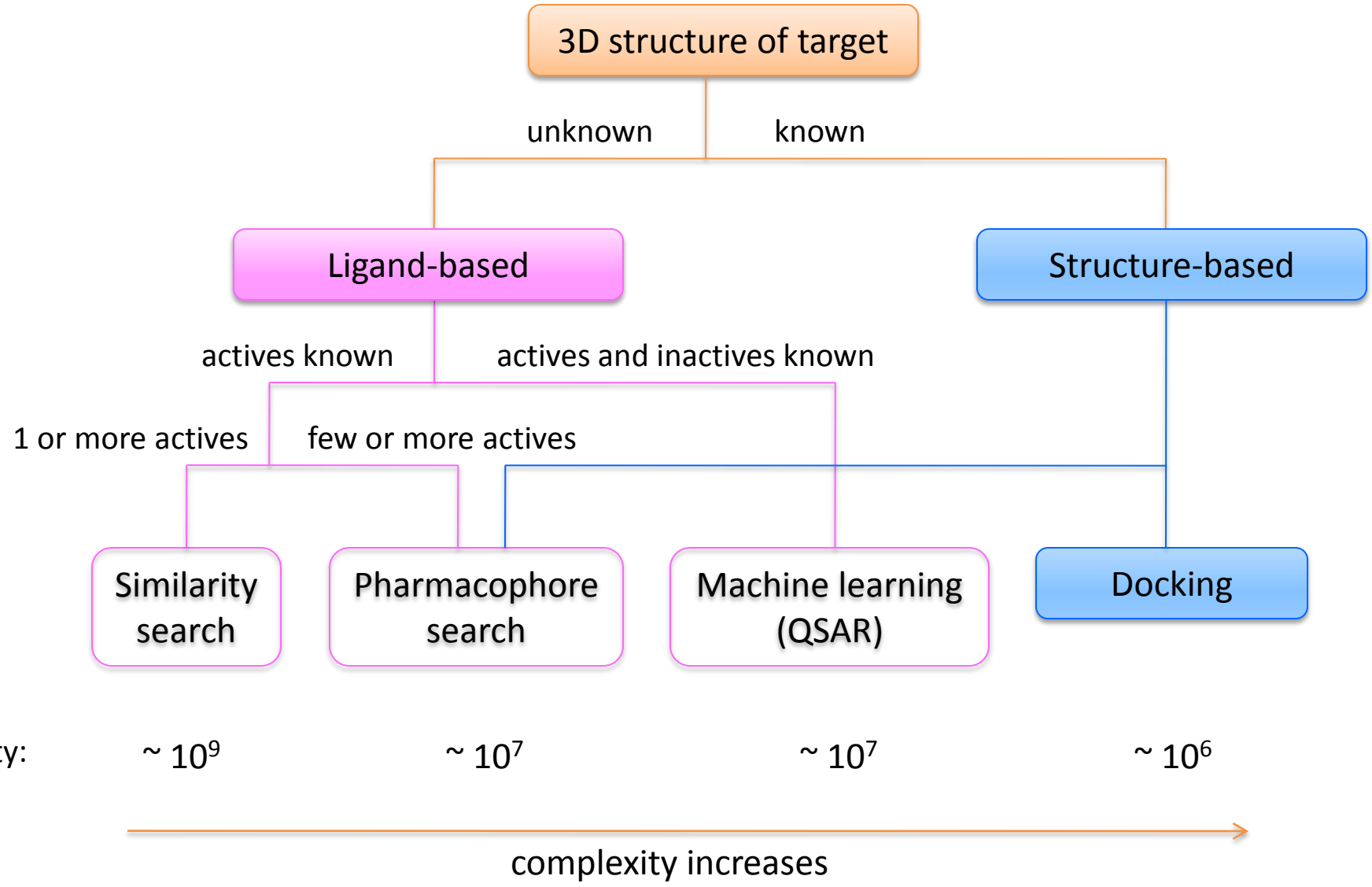
Zhang L. et al., J. Chem. Inf. Model., **2013**, 53 (2), pp 475–492

QSAR: conclusion

+ Qualitative and quantitative output
+ May work for compounds having different mechanisms of action
+ Fast screening

- Very demanding to the quality of input data
- Applicability limited by the training set structures
- Hard to encode stereochemistry

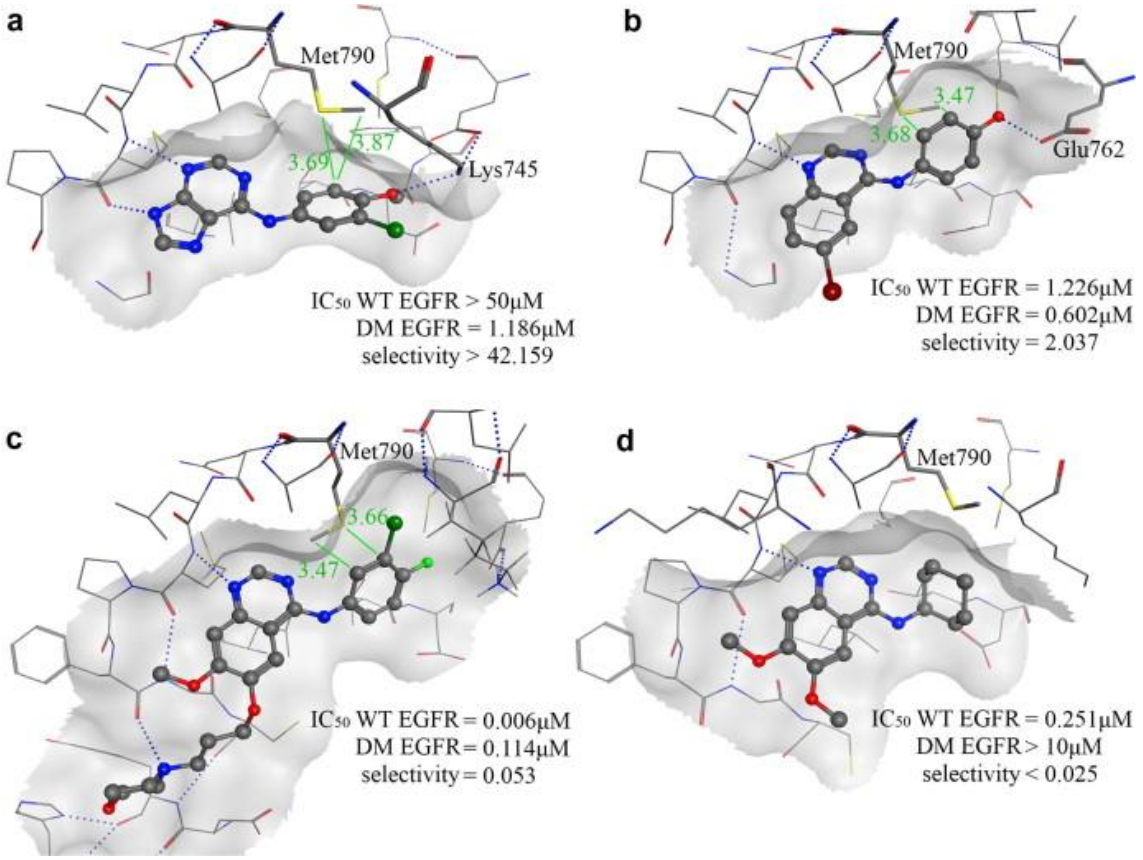
Molecular docking



Docking is an *in silico* tool which predicts

Pose – a possible relative orientation of a ligand and a receptor as well as conformation of a ligand and a receptor when they are form complex

Score – the strength of binding of the ligand and the receptor.



Bioorganic & Medicinal Chemistry, 20 (12), 2012, 3756–3767.

Why docking is a hard task

Complex 3D jigsaw puzzle

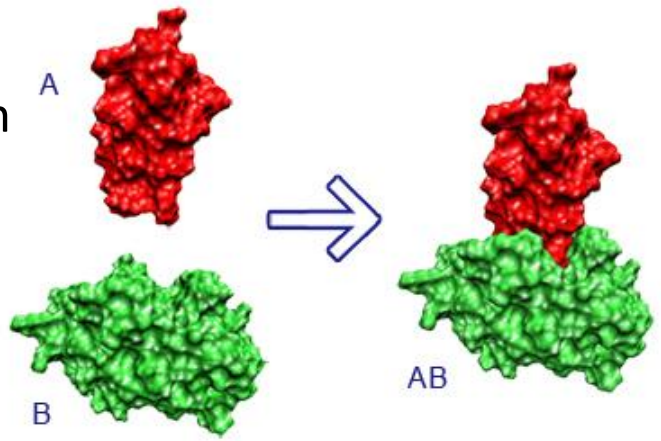
Conformational flexibility – many degrees of freedom

Mutual adaptation (“induced fit”)

Solvation in aqueous media

Complexity of thermodynamic contribution

No easy route to evaluation of ΔG



Simplification and heuristic approaches are necessary

“At its simplest level, this is a problem of subtraction of large numbers, inaccurately calculated, to arrive at a small number.”

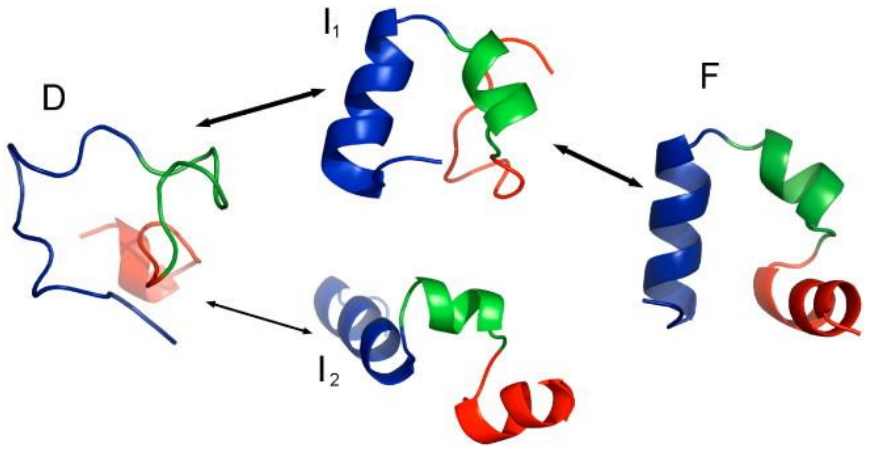
(Leach A.R., Shoichet B.K., Peishoff C.E.. *J. Med. Chem.* 2006, *49*, 5851-5855)

Sampling and scoring

Protein-ligand docking software consists of two main components which work together:

1. **Search algorithm (sampling)** - generates a large number of poses of a molecule in the binding site.

2. **Scoring function** - calculates a score or binding affinity for a particular pose



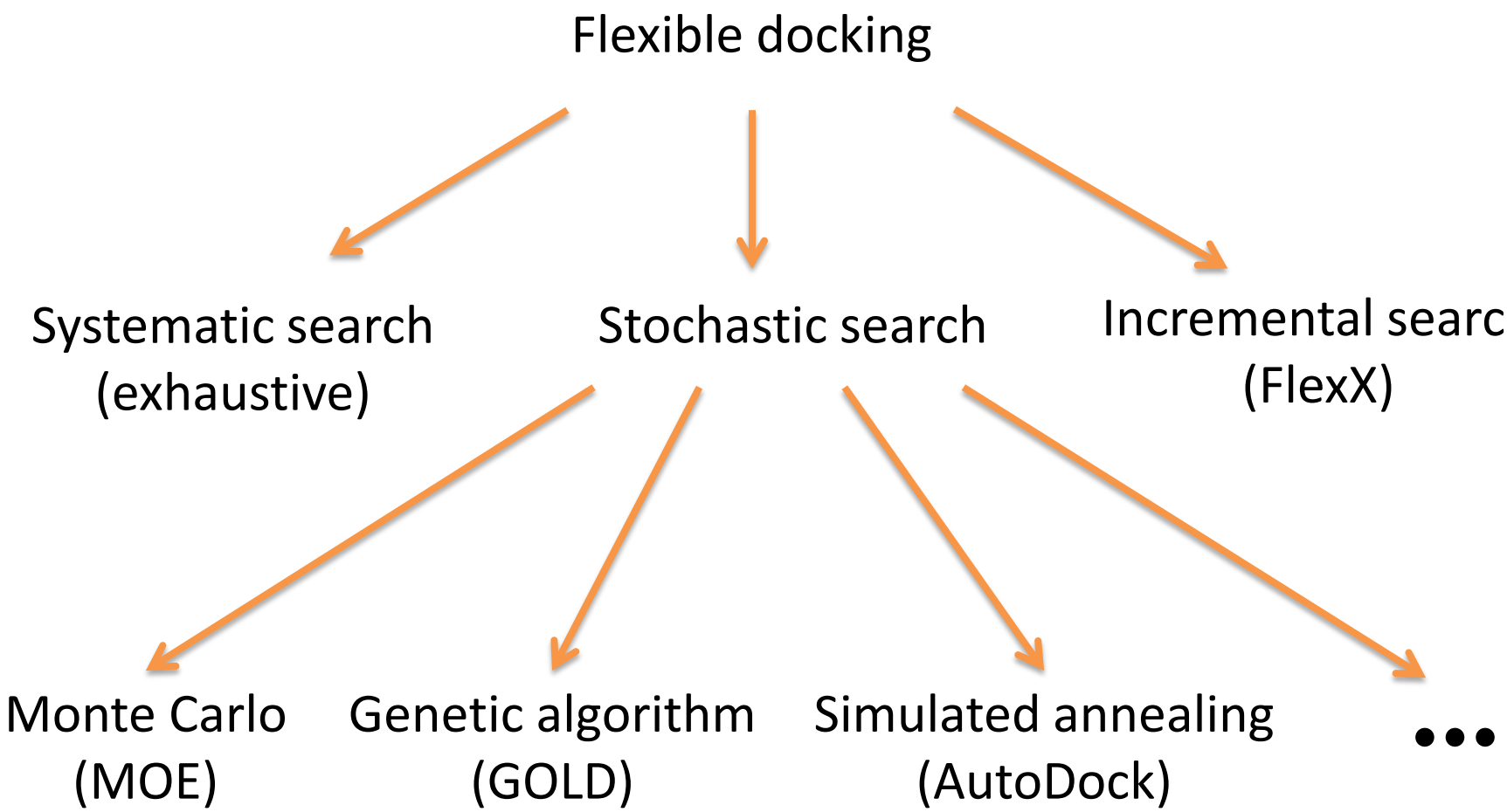
Search algorithms (sampling)

Ligand	Receptor
Rigid	Rigid
Flexible	Rigid
Flexible	Flexible

Fast & Simple

Slow & Complex

Search algorithms (sampling)



Classes of scoring function

Forcefield-based

Based on terms from molecular mechanics forcefields

GoldScore, DOCK, AutoDock

Empirical

Parameterised against experimental binding affinities

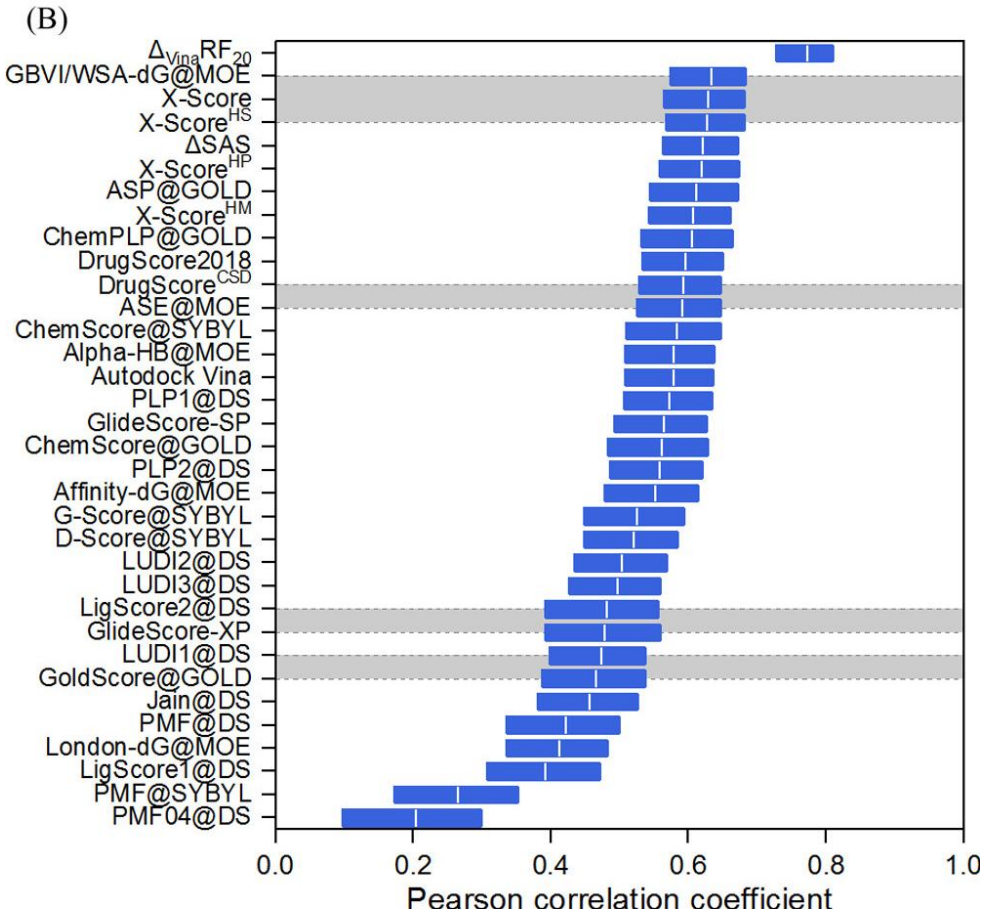
ChemScore, PLP, Glide SP/XP

Knowledge-based potentials

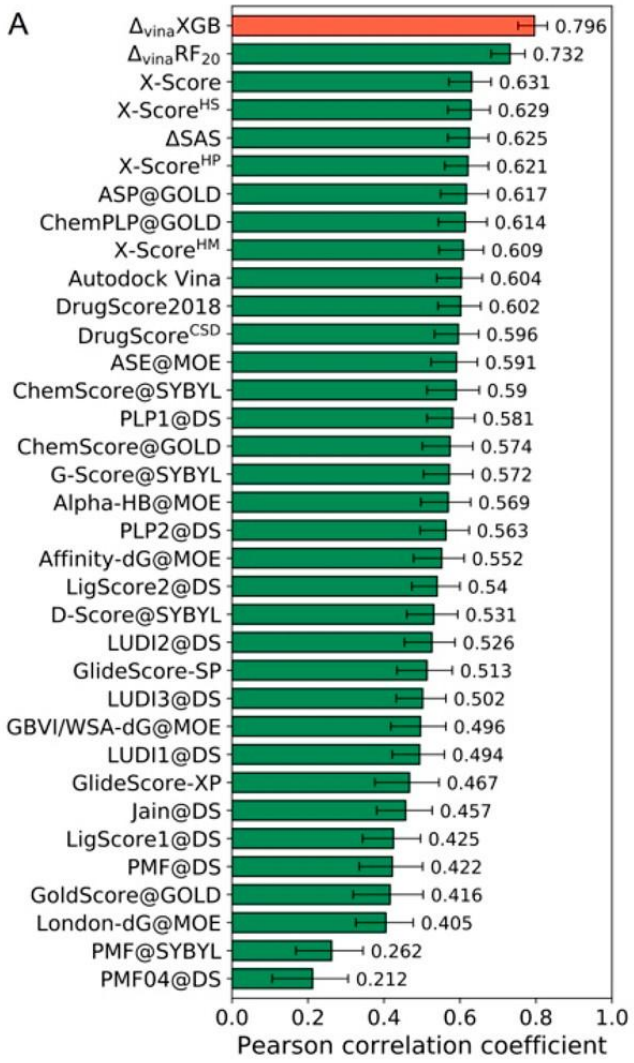
Based on statistical analysis of observed pairwise distributions

PMF, DrugScore, ASP

Docking quality assessment



Su, M. et al Comparative Assessment of Scoring Functions: The CASF-2016 Update. *Journal of Chemical Information and Modeling* **2019**, 59, (2), 895-913.

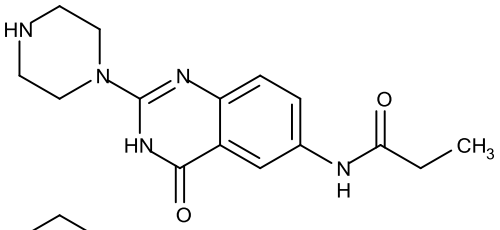


Lu, J et al, *Journal of Chemical Information and Modeling* **2019**, 59, (11), 4540-4549.

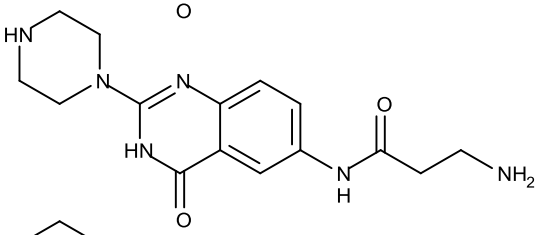
Docking quality assessment

Antagonists of $\alpha_{IIb}\beta_3$

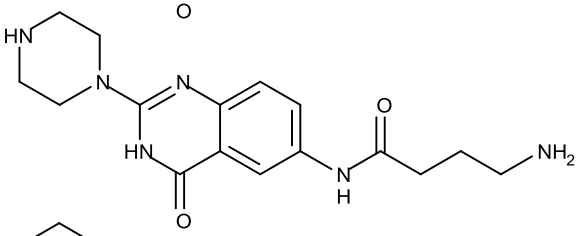
1



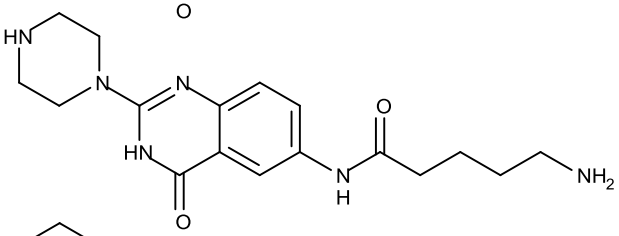
2



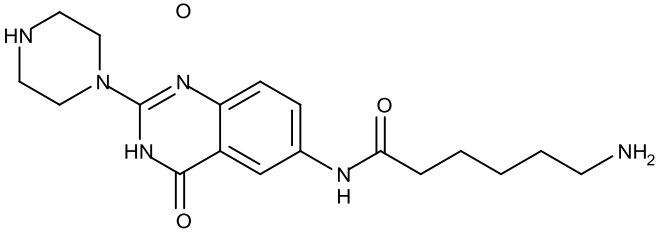
3

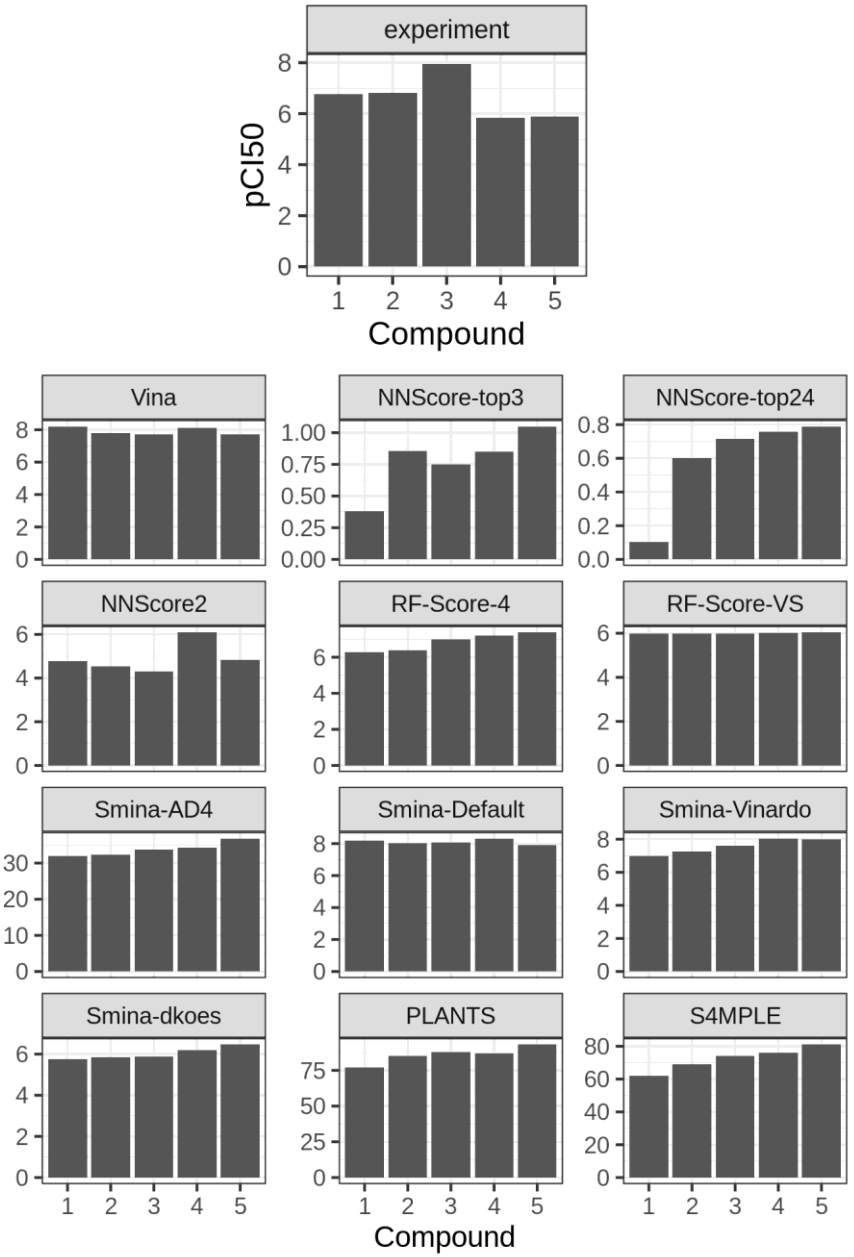


4



5



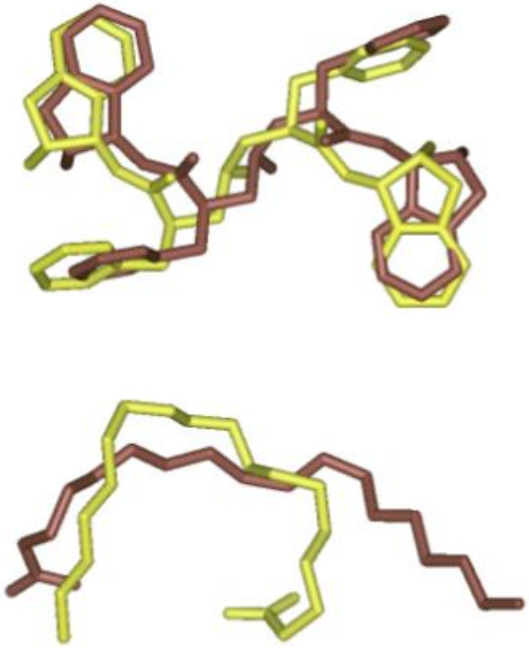


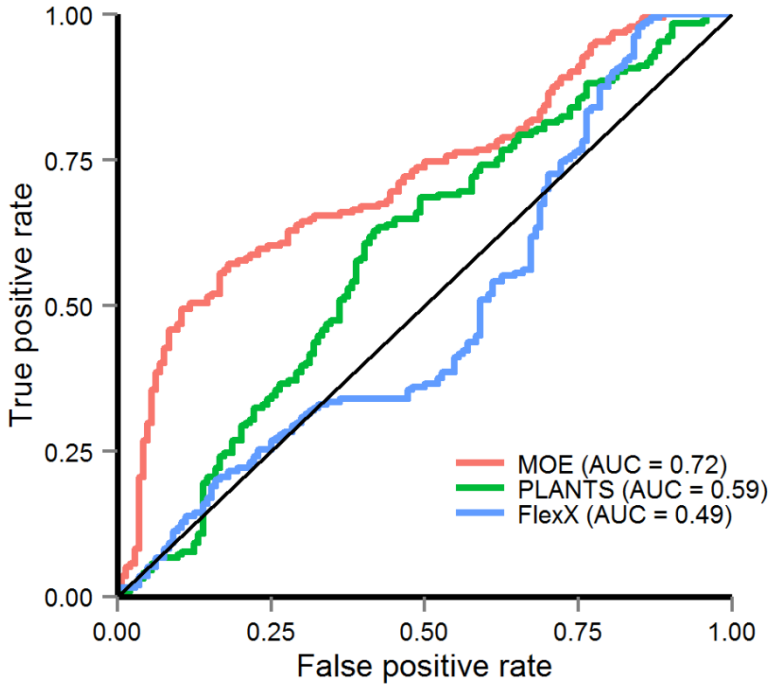
Krysko, A.A., et al., *Bioorganic & Medicinal Chemistry Letters,* **2016**, 26, 1839-1843

Validation

Self-docking – reproducibility of a pose

Docking of a set of ligands with known affinity – reproducibility of affinity





Molecular docking: example

ligands of D4 receptor

enumerated library

138 M compounds

DOCK

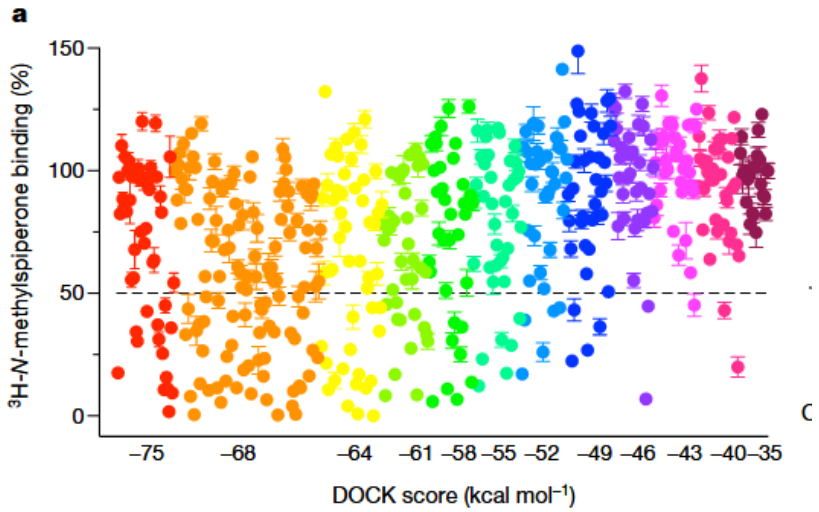
remove similar to known (ChEMBL) and in 3.5 M in-stock library

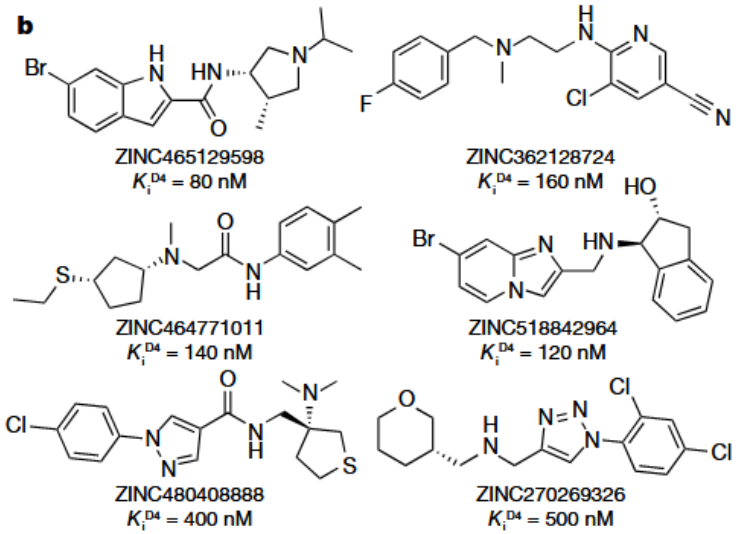
1000 clusters

124 + 444 selected

K_i < 8.3 µM

81 compounds





Lyu, J. et al Ultra-large library docking for discovering new chemotypes. *Nature* **2019**, 566, 224-229.

Molecular docking: conclusion

+ Relatively fast
+ Determine binding poses
+ Good in ranking ligands for virtual screening

- Low accuracy of binding energy estimation
- Require knowledge about binding site