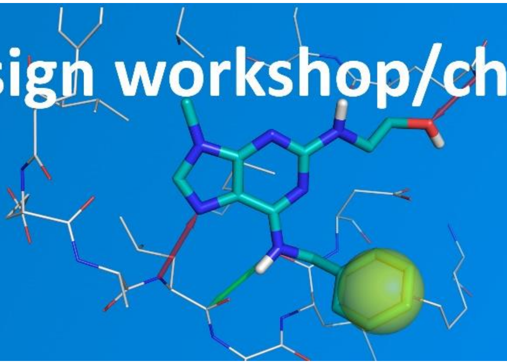
7th Advanced In silico Drug Design workshop/challenge

29 January - 2 February 2024
Olomouc, Czech Republic



Univerzita Palackého
v Olomouci

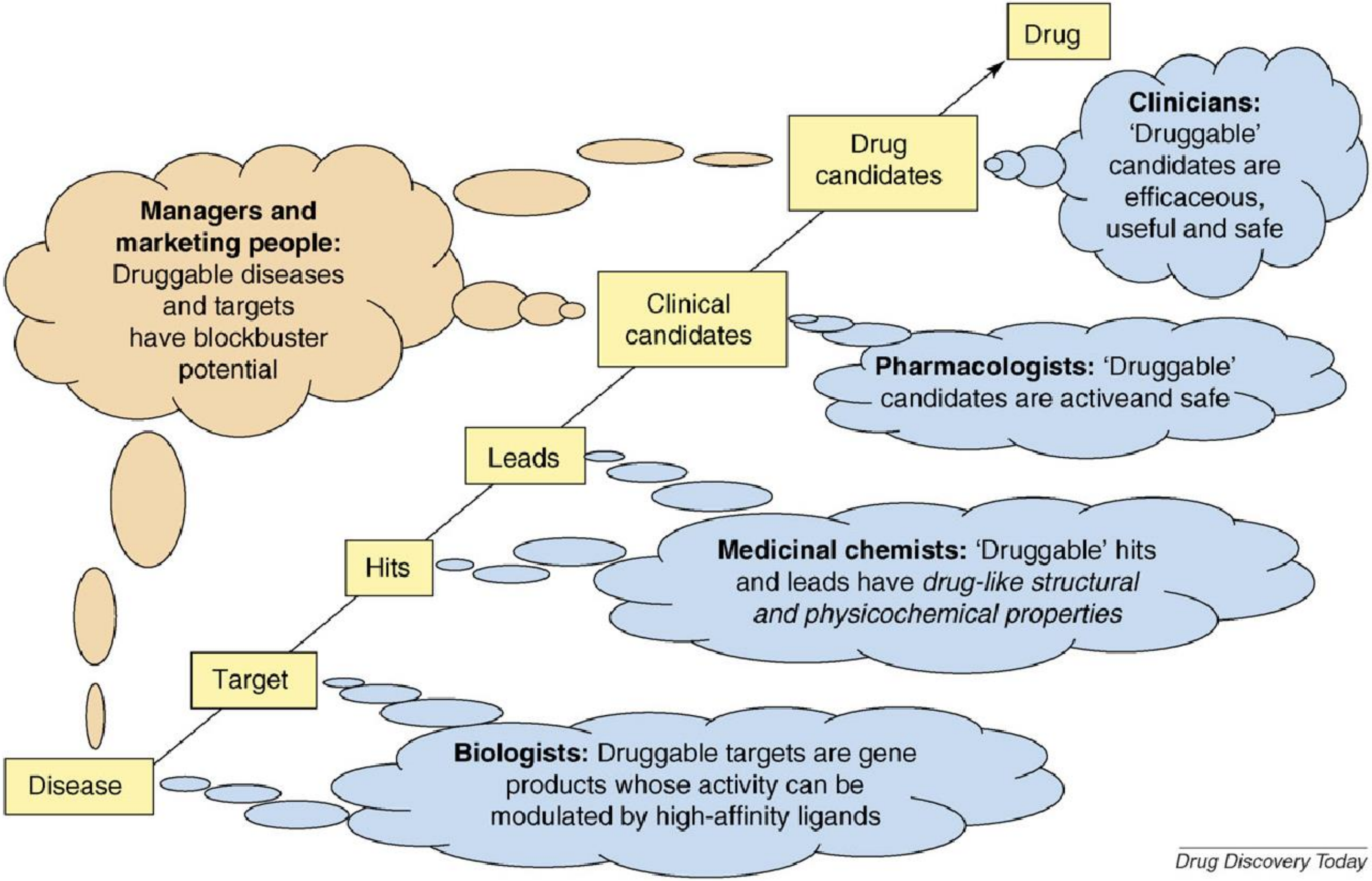
Virtual screening in drug discovery

Pavel Polishchuk

Institute of Molecular and Translational Medicine
Palacky University

pavlo.polishchuk@upol.cz

Drug development workflow



Disease

Target

Hits

Leads

Clinical candidates

Drug candidates

Drug

Managers and marketing people: Druggable diseases and targets have blockbuster potential

Biologists: Druggable targets are gene products whose activity can be modulated by high-affinity ligands

Medicinal chemists: ‘Druggable’ hits and leads have *drug-like structural and physicochemical properties*

Pharmacologists: ‘Druggable’ candidates are activeand safe

Clinicians: ‘Druggable’ candidates are efficaceous, useful and safe

Drug Discovery Today

Vistoli G., et al., *Drug Discovery Today*, **2008**, 13, 285-294

IMTM

Vastness of chemical space

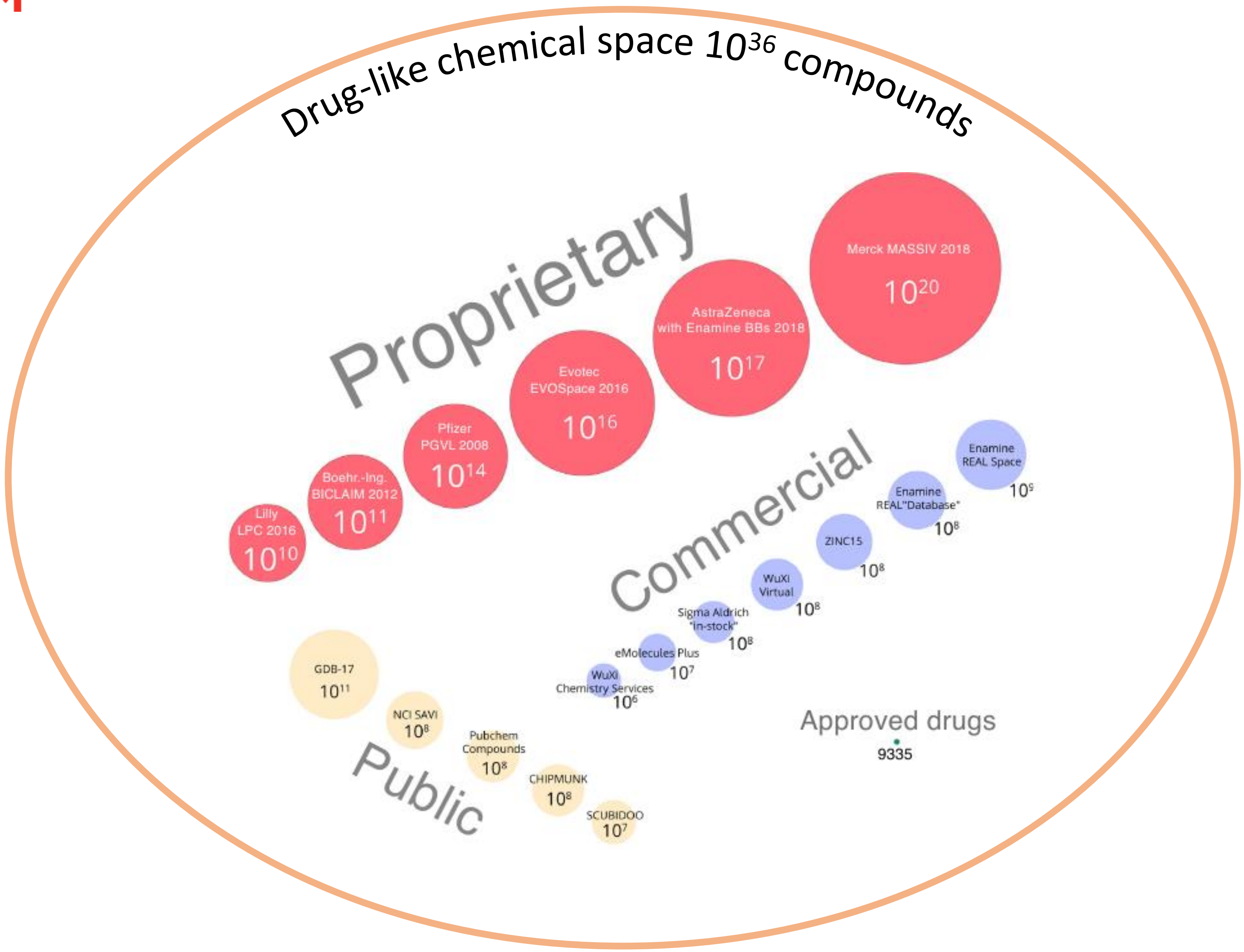
real datasets

SCIFINDER® A CAS SOLUTION	~ 160 M compounds	Commercial
REAXYS®	~ 105 M compounds	
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}

Vastness of chemical space



Hoffmann, T.; Gastreich, M., The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discovery Today* **2019**, 24, 1148-1156

Screening

High-throughput screening (HTS)

up to 10^6 of compounds can be tested

- expensive
- not all targets are suitable for HTS

DNA-encoded libraries (DEL)

up to 10^9 of compounds can be tested

- moderately expensive
- not all reactions can be adopted to DEL conditions

Virtual screening

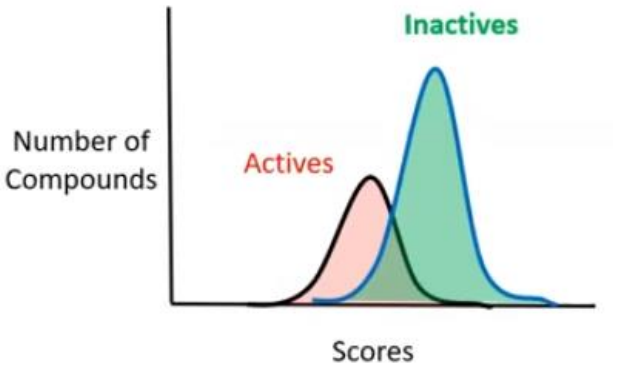
up to 10^{12} of compounds can be tested

- cheap
- fast
- not very accurate

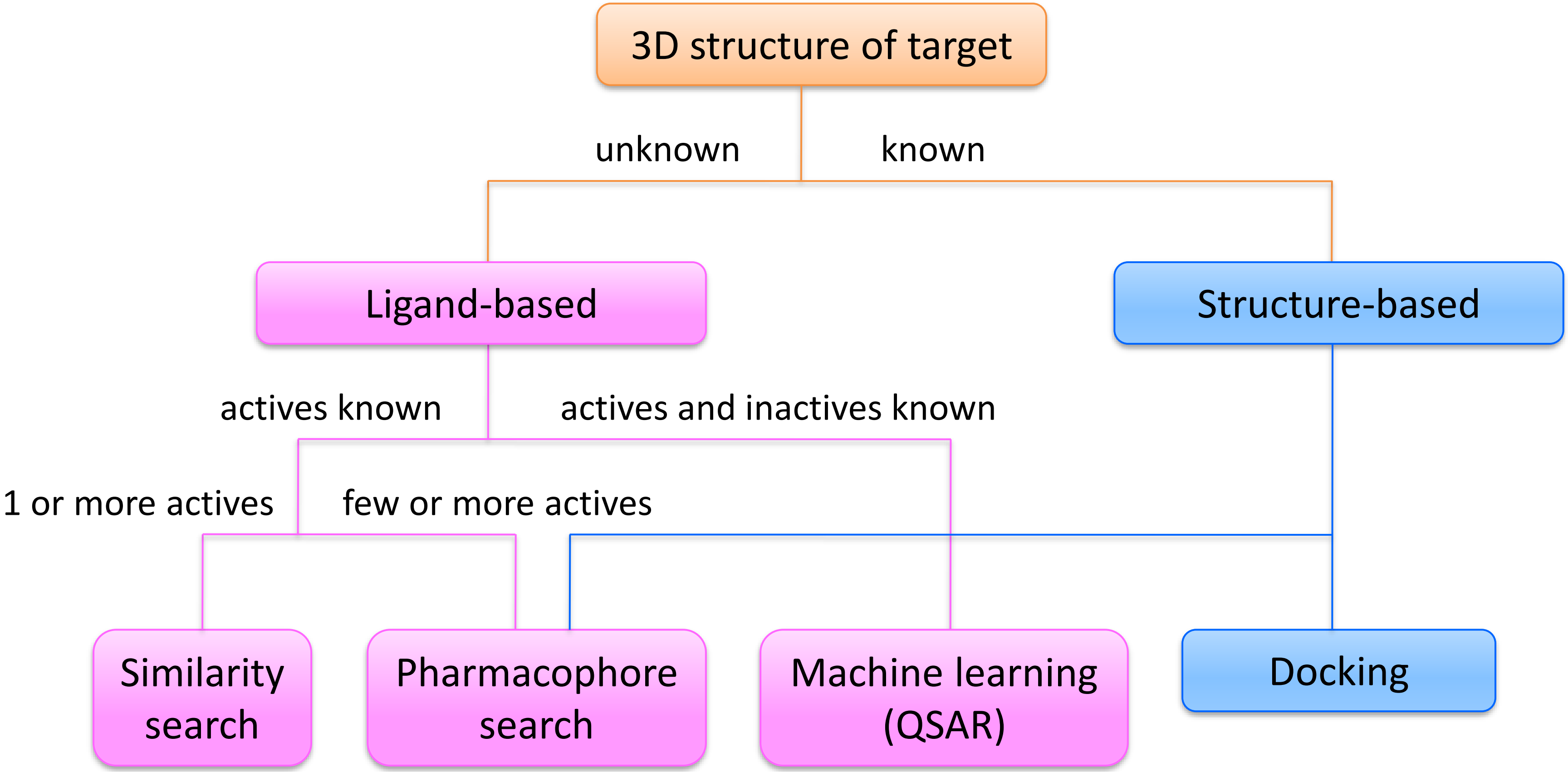
Virtual screening concept

Molecule ID	Score
CHEMBL1367590	0.127
CHEMBL2403348	0.715
CHEMBL4209434	0.585
CHEMBL204341	0.599
CHEMBL494704	0.072
CHEMBL1581690	0.554
CHEMBL4869612	0.686
CHEMBL447111	0.660
CHEMBL152972	0.108
CHEMBL4851230	0.438
CHEMBL494705	0.118
CHEMBL398456	0.347
CHEMBL4760508	0.828
CHEMBL196509	0.214
CHEMBL522471	0.471
CHEMBL3657154	0.538
CHEMBL361258	0.465
CHEMBL1370	0.122
CHEMBL296411	0.189
CHEMBL511492	0.143
CHEMBL4850019	0.171
CHEMBL441537	0.591
CHEMBL399142	0.661
CHEMBL235386	0.639
CHEMBL1342736	0.030
CHEMBL106773	0.965
CHEMBL3427390	0.776
CHEMBL3827784	0.206
CHEMBL192325	0.486
CHEMBL1301796	0.162
CHEMBL4243739	0.755
CHEMBL1347829	0.004
CHEMBL1676	0.027

Molecule ID	Score
CHEMBL106773	0.965
CHEMBL4760508	0.828
CHEMBL3427390	0.776
CHEMBL4243739	0.755
CHEMBL2403348	0.715
CHEMBL4869612	0.686
CHEMBL399142	0.661
CHEMBL447111	0.660
CHEMBL235386	0.639
CHEMBL204341	0.599
CHEMBL441537	0.591
CHEMBL4209434	0.585
CHEMBL1581690	0.554
CHEMBL3657154	0.538
CHEMBL192325	0.486
CHEMBL522471	0.471
CHEMBL361258	0.465
CHEMBL4851230	0.438
CHEMBL398456	0.347
CHEMBL196509	0.214
CHEMBL3827784	0.206
CHEMBL296411	0.189
CHEMBL4850019	0.171
CHEMBL1301796	0.162
CHEMBL511492	0.143
CHEMBL1367590	0.127
CHEMBL1370	0.122
CHEMBL494705	0.118
CHEMBL152972	0.108
CHEMBL494704	0.072
CHEMBL1342736	0.030
CHEMBL1676	0.027
CHEMBL1347829	0.004



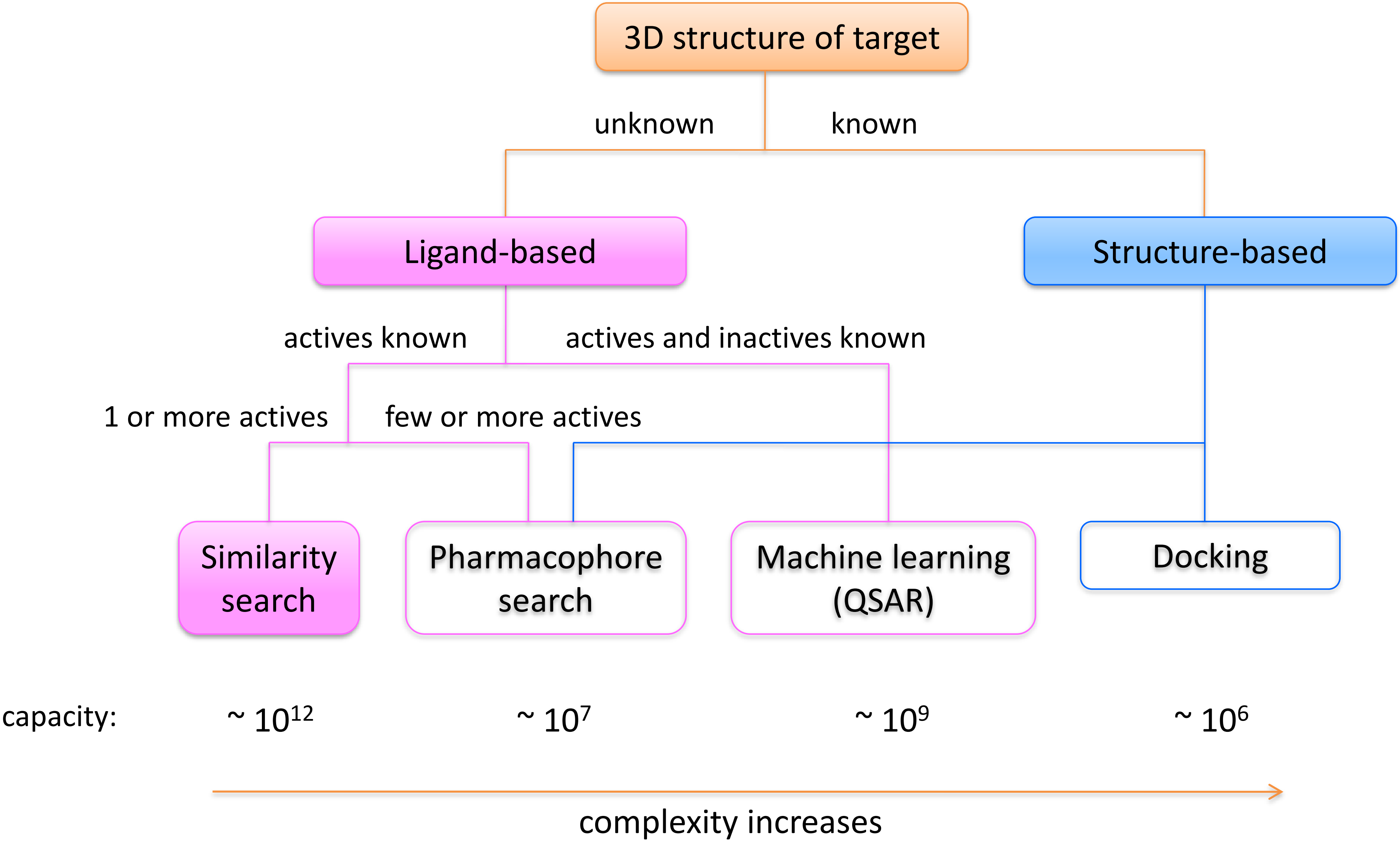
Virtual screening methods



capacity: ~ 10^{12} ~ 10^{7} ~ 10^{9} ~ 10^{6}

complexity increases

Similarity search



3D structure of target

unknown

known

Ligand-based

Structure-based

actives known

actives and inactives known

1 or more actives

few or more actives

Similarity search

Pharmacophore search

Machine learning (QSAR)

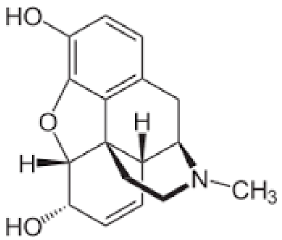
Docking

capacity: ~ 10^{12} ~ 10^{7} ~ 10^{9} ~ 10^{6}

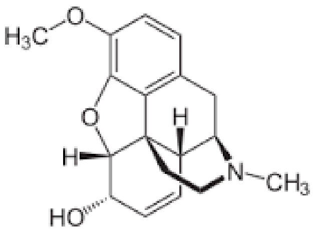
complexity increases

Similarity principle

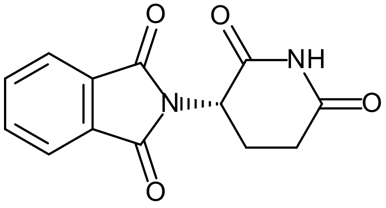
Similar compounds have similar properties



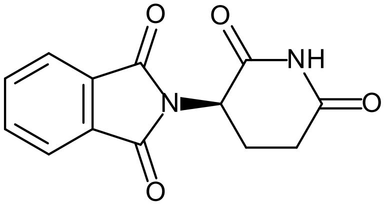
morphine



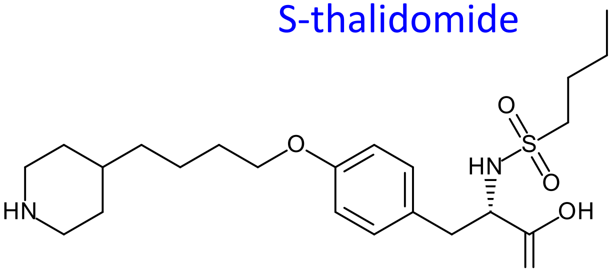
codeine



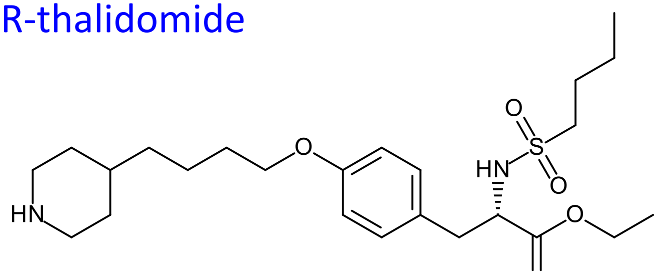
S-thalidomide



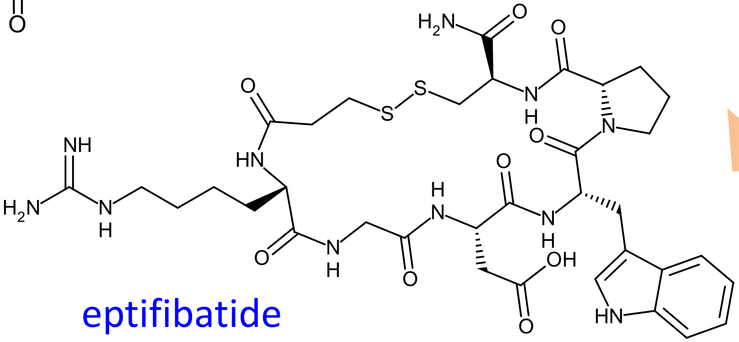
R-thalidomide



tirofibane



tirofibane ethyl ester



eptifibatide

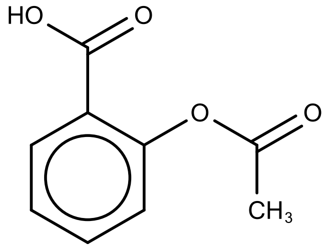
Ranking of compounds: example

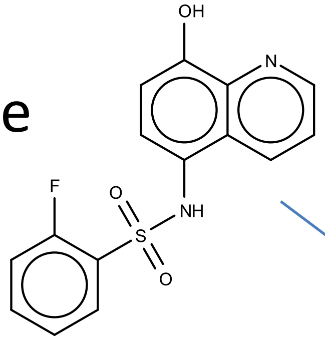
Structure representation

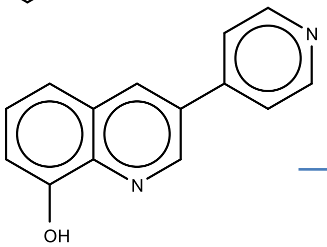
- structural keys
- fingerprints
- molecular shape
- …

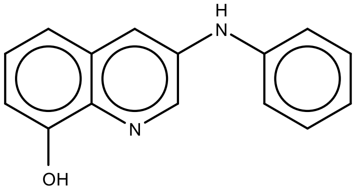
Similarity measure

- Tanimoto
- Dice
- Euclidian
- …







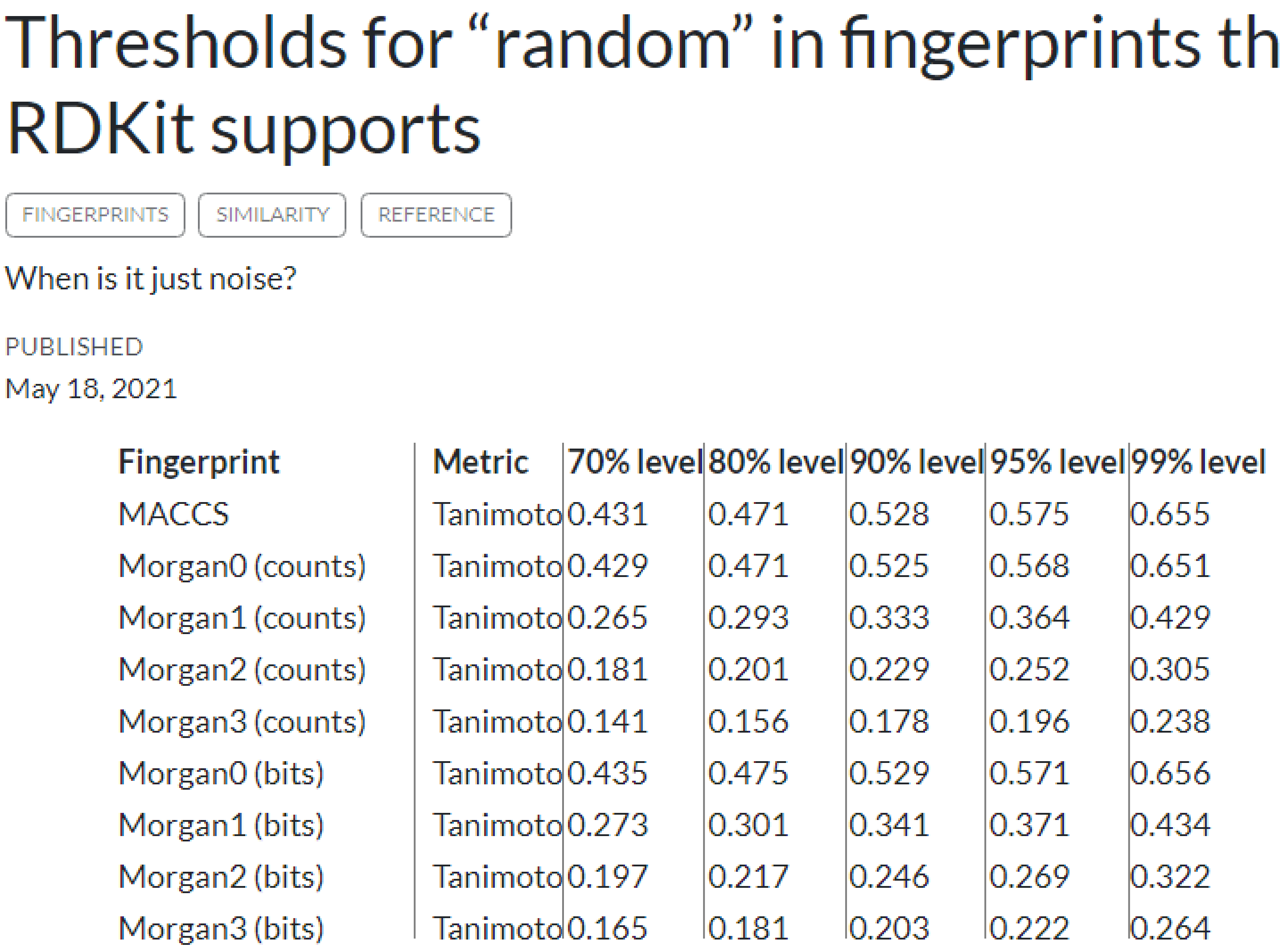


Dice		
Atom pairs	ECFP4	FCFP4
0.327 (3)	0.219 (2)	0.233 (1)
0.364 (1)	0.185 (3)	0.170 (2)
0.333 (2)	0.291 (1)	0.125 (3)

*binary fingerprints calculated with RDKit

Similarity search output depends on descriptors and similarity measure selected

What is similarity between random compounds



Thresholds for “random” in fingerprints the RDKit supports

FINGERPRINTS SIMILARITY REFERENCE

When is it just noise?

PUBLISHED

May 18, 2021

Fingerprint	Metric	70% level	80% level	90% level	95% level	99% level
MACCS	Tanimoto	0.431	0.471	0.528	0.575	0.655
Morgan0 (counts)	Tanimoto	0.429	0.471	0.525	0.568	0.651
Morgan1 (counts)	Tanimoto	0.265	0.293	0.333	0.364	0.429
Morgan2 (counts)	Tanimoto	0.181	0.201	0.229	0.252	0.305
Morgan3 (counts)	Tanimoto	0.141	0.156	0.178	0.196	0.238
Morgan0 (bits)	Tanimoto	0.435	0.475	0.529	0.571	0.656
Morgan1 (bits)	Tanimoto	0.273	0.301	0.341	0.371	0.434
Morgan2 (bits)	Tanimoto	0.197	0.217	0.246	0.269	0.322
Morgan3 (bits)	Tanimoto	0.165	0.181	0.203	0.222	0.264

https://greglandrum.github.io/rdkit-blog/posts/2021-05-18-fingerprint-thresholds1.html

Similarity search: chemfp project

Dalke *J Cheminform* (2019) 11:76
https://doi.org/10.1186/s13321-019-0398-8

Journal of Cheminformatics

METHODOLOGY Open Access

The chemfp project

Andrew Dalke*

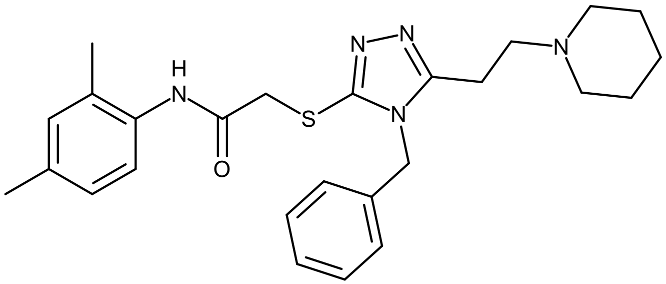
Check for updates

Fingerprints supported:

- RDKit
- CDK
- OpenEye
- OpenBabel
- PubChem
- ChemFP

Similarity search: example

agonists of CCR5



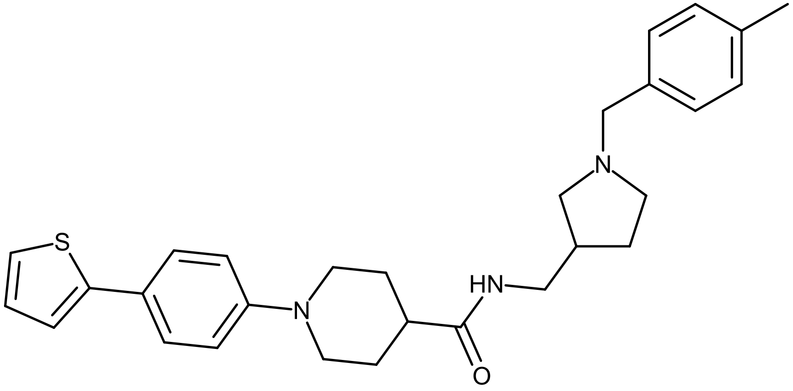
IC$_{50}$ = 17 µM

60 000 compounds

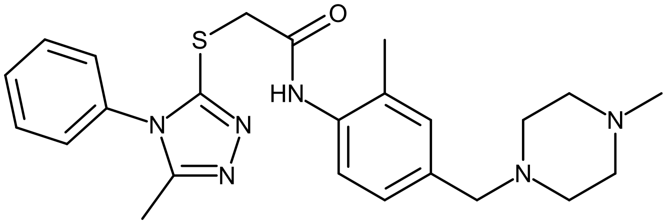
FCFP4

100 compounds

purchased & tested



IC$_{50}$ = 5.8 µM



IC$_{50}$ = 14.1 µM

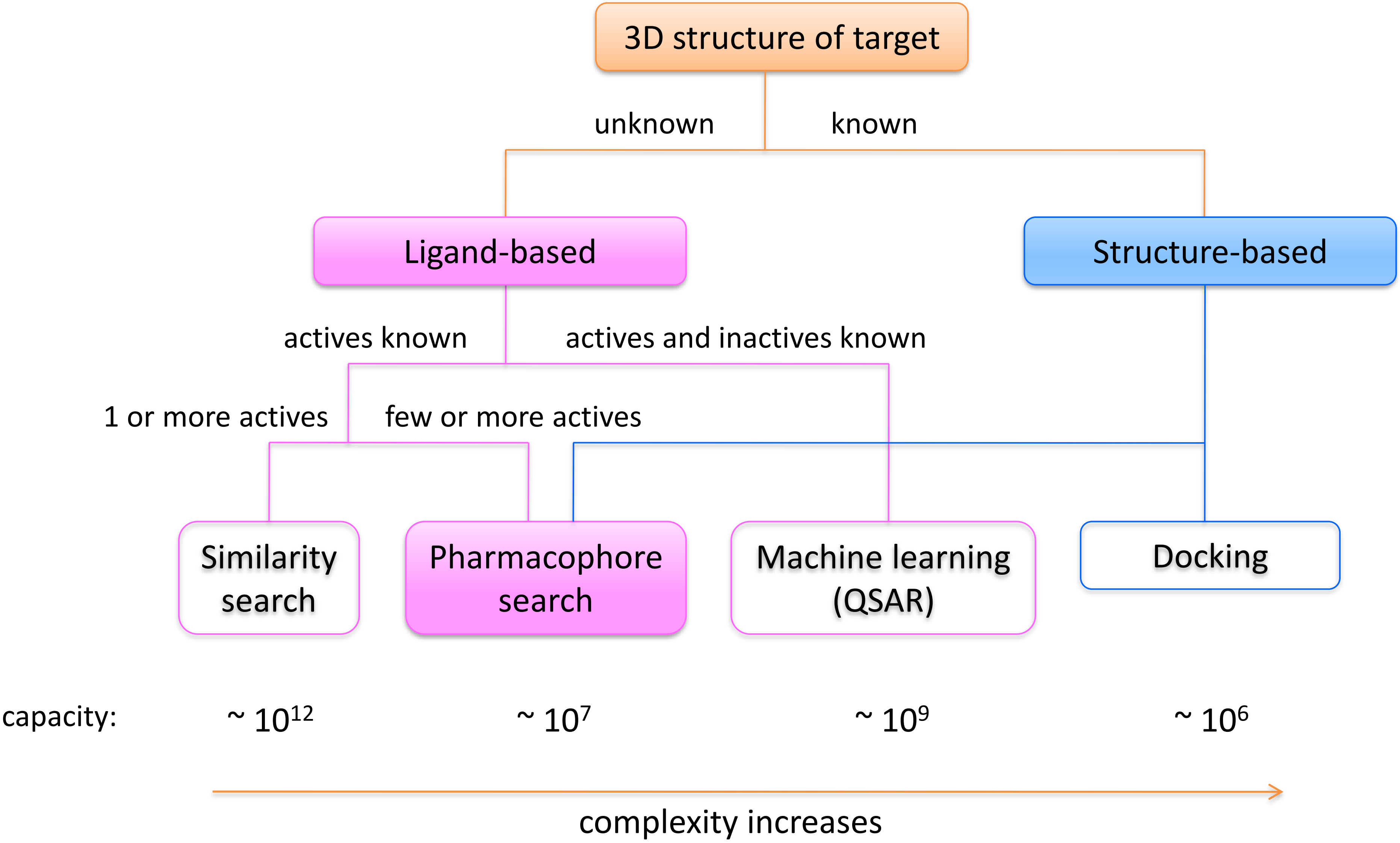
Kellenberger, E., et al., Identification of nonpeptide CCR5 receptor agonists by structure-based virtual screening. *Journal of Medicinal Chemistry* **2007**, 50, 1294–1303.

Similarity search: conclusions

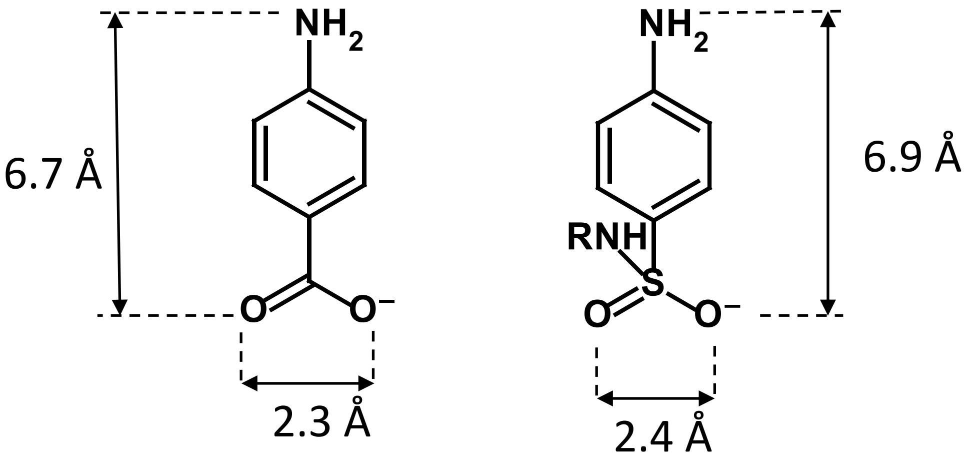
+ Little information is required to start searching
+ Different chemotypes can be retrieved
+ Ultra fast screening

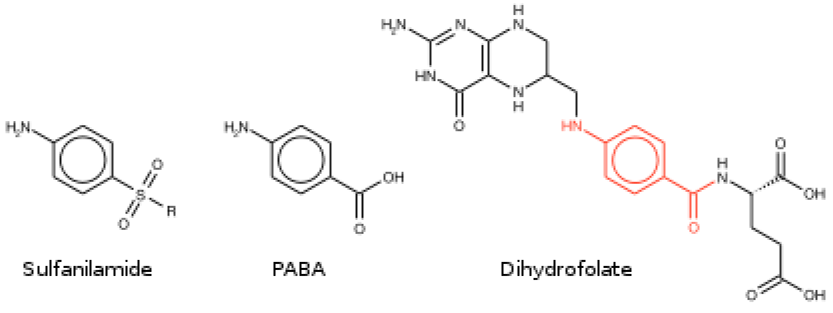
- Hits may share common substructures with reference structures that may reduce their patentability
- Results depend on chosen descriptors and similarity measure
- Structural similarity is not always followed by biological one

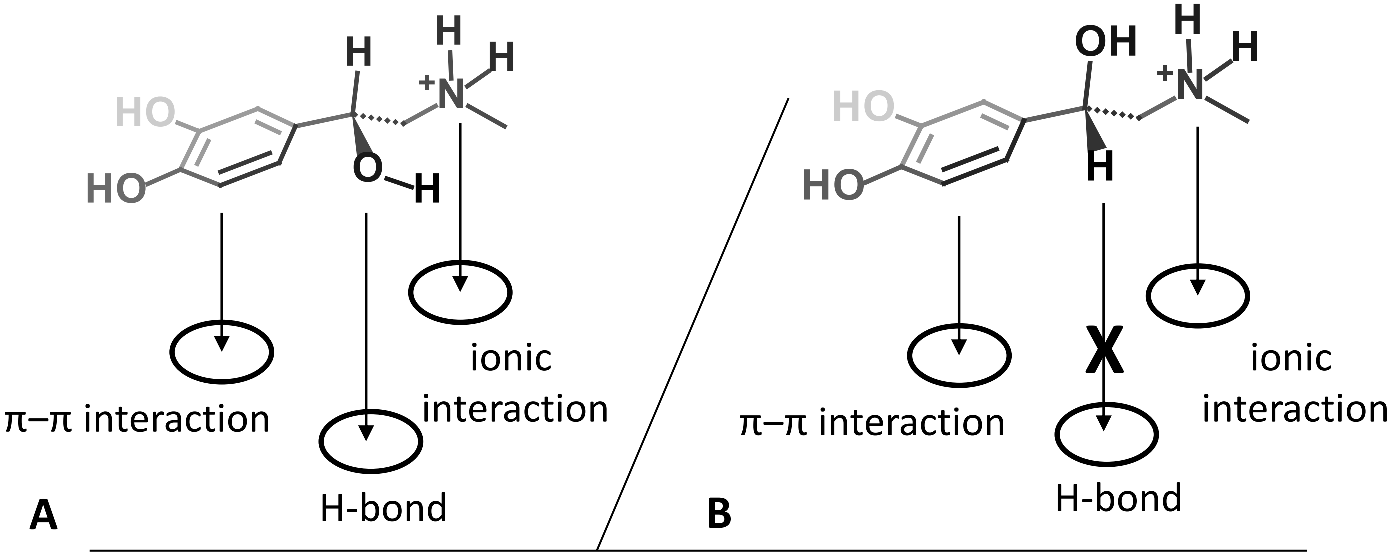
Pharmacophore search



Early pharmacophore hypothesis





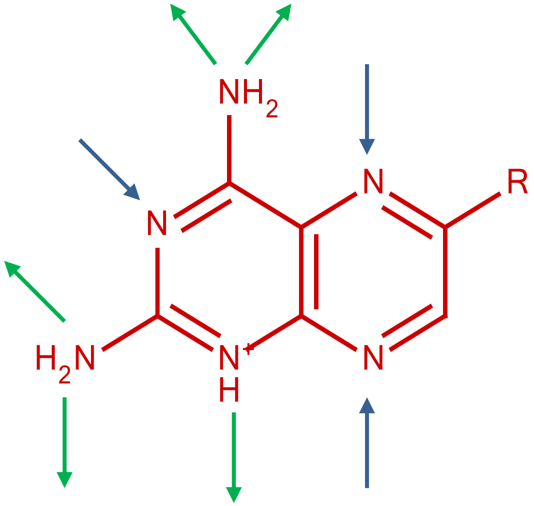


(R)-(-) –Epinephrine
(Adrenalin)

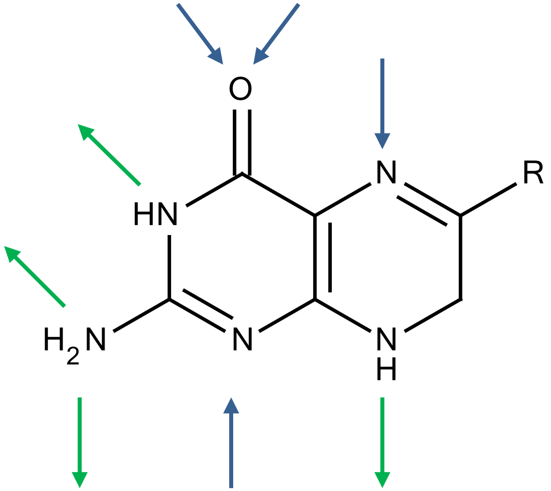
(S)-(+) –Epinephrine

Atom- and pharmacophore-based alignment

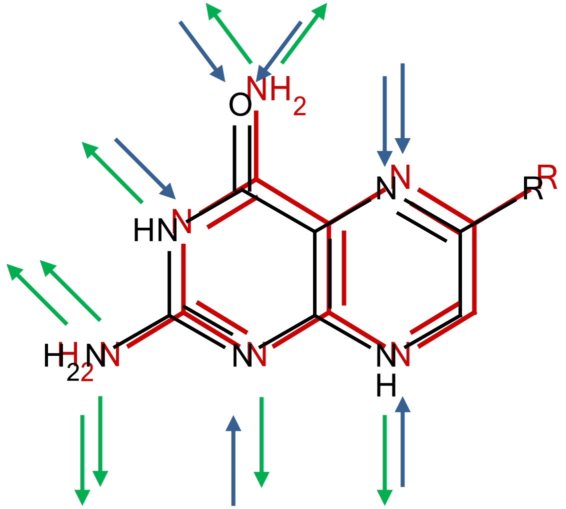
Methotrexate



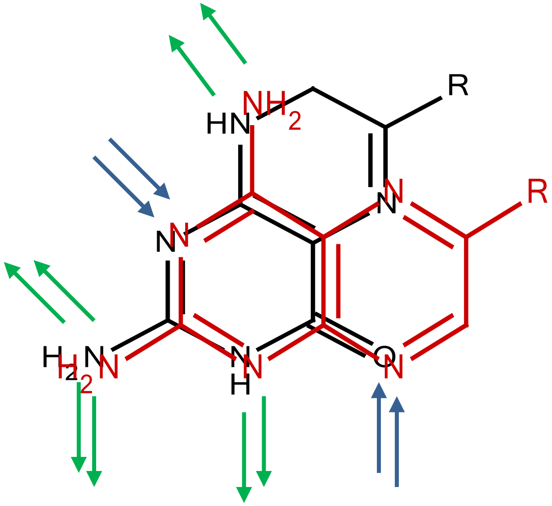
Dihydrofolate



Hydrogen bonding patterns



Atom-based alignment



Pharmacophore alignment

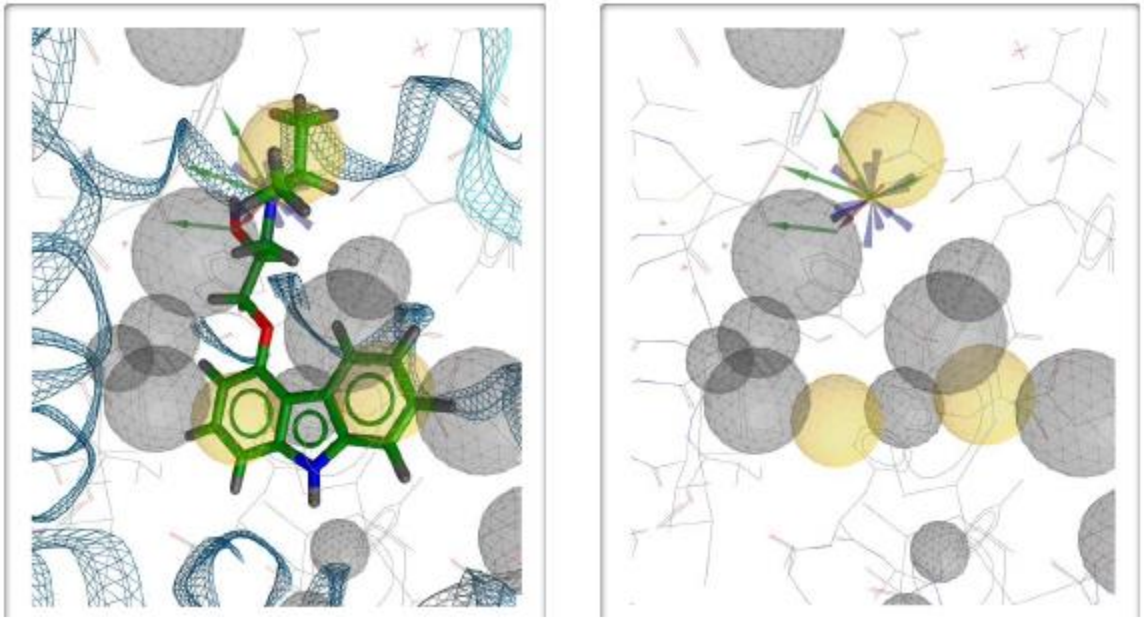
Pharmacophore definition

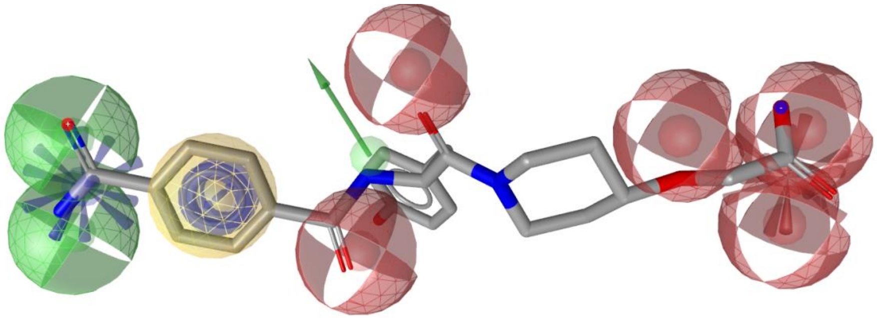
A **pharmacophore** is the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interaction with a specific biological target structure and to trigger (or block) its biological response.

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

Feature-based pharmacophore models

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



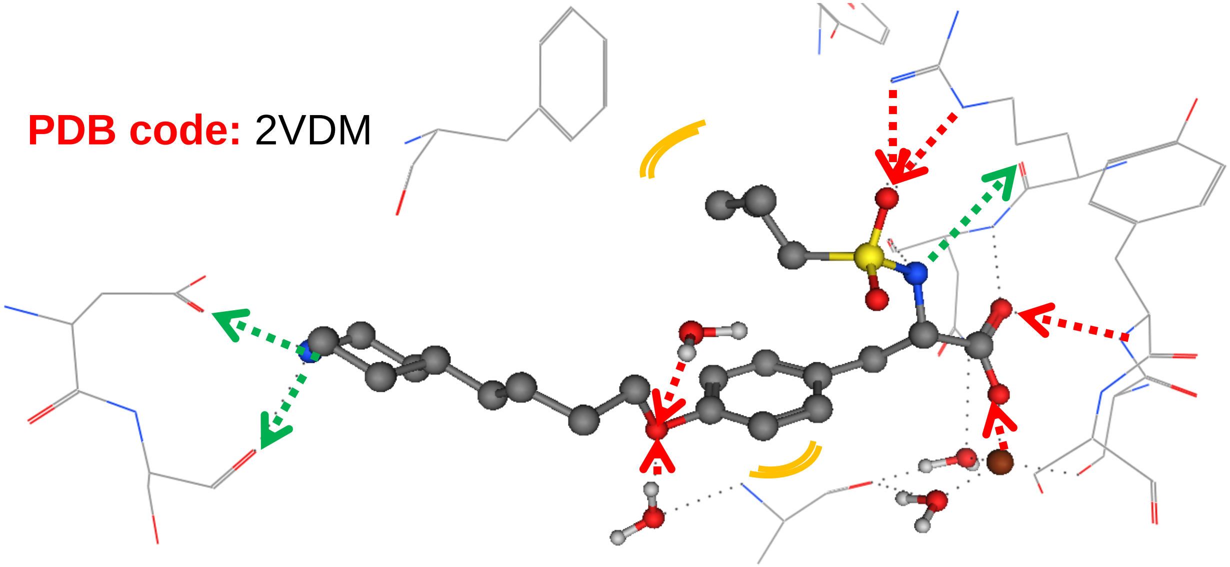


Pharmacophore features

- 🟩 *H-bond donor*
- 🟥 *H-bond acceptor*
- ✳ *Positive ionizable*
- ✳ *Negative ionizable*
- ● *Hydrophobic*
- Aromatic ring

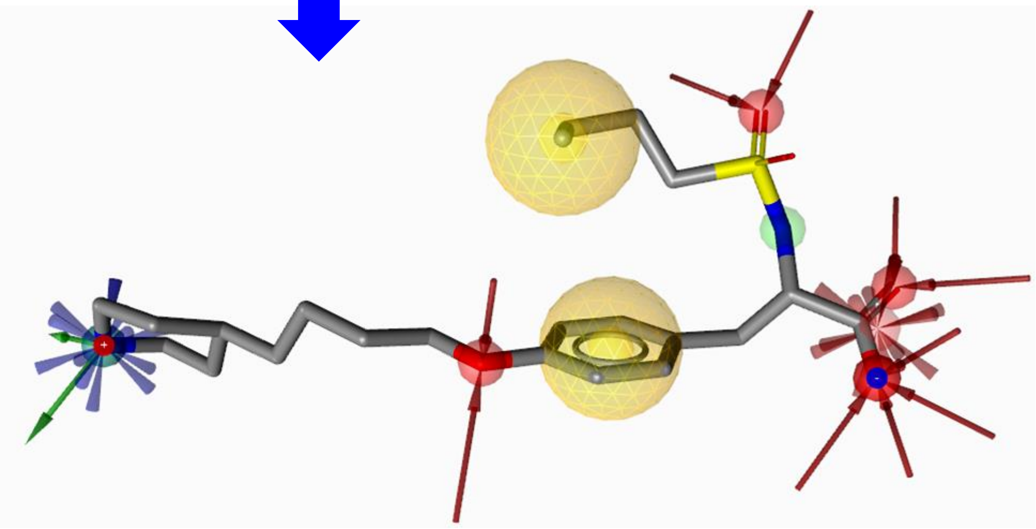
Structure-based pharmacophores

PDB code: 2VDM



H-bonds formed by the ligand

Hydrophobic interaction

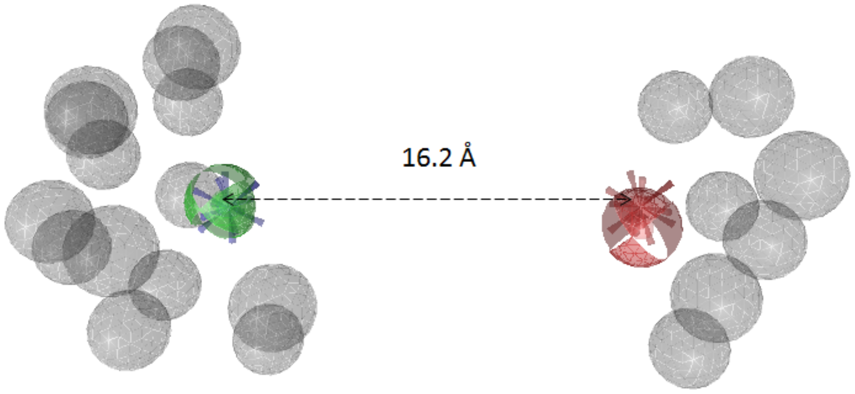


Pharmacophore features

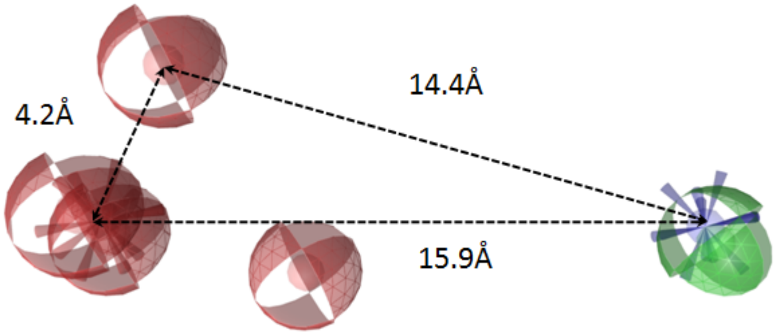
- H-bond donor
- H-bond acceptor
- Positive ionizable
- Negative ionizable
- Hydrophobic

Ligand-based pharmacophores

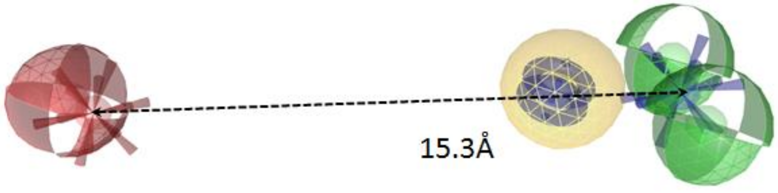
Shared model on 83 antagonists of fibrinogen receptor



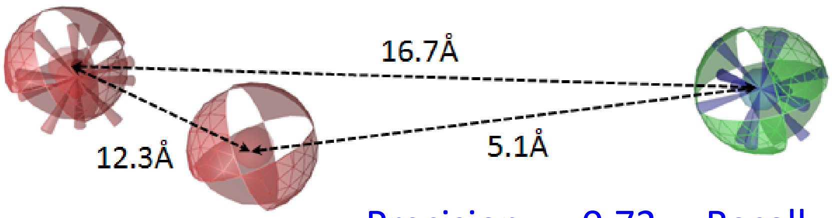
Pharmacophore models obtained for clusters of compounds



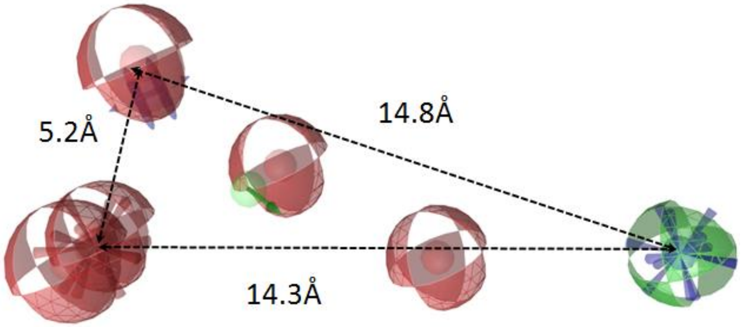
Precision: 0.77 Recall: 0.27



Precision: 0.67 Recall: 0.29



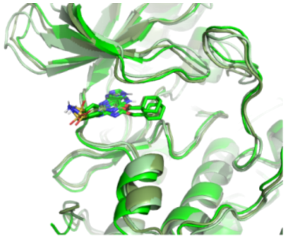
Precision: 0.72 Recall: 0.77

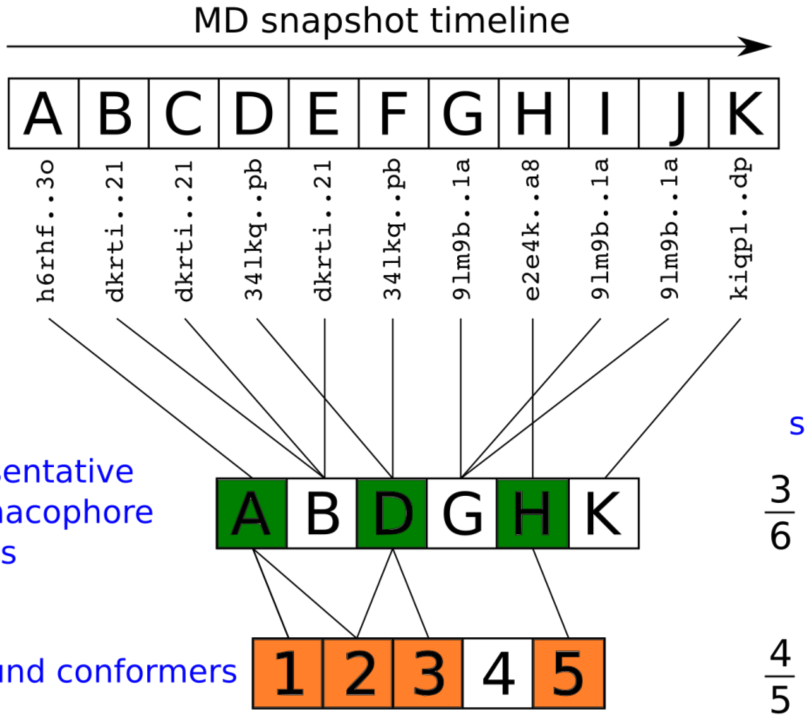


Precision: 0.90 Recall: 0.04

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

MD pharmacophores





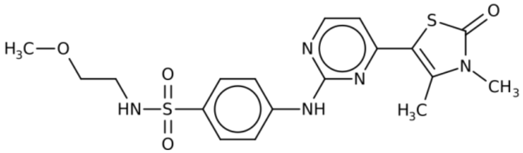
MD pharmacophores

3D pharmacophore hashes

representative pharmacophore models

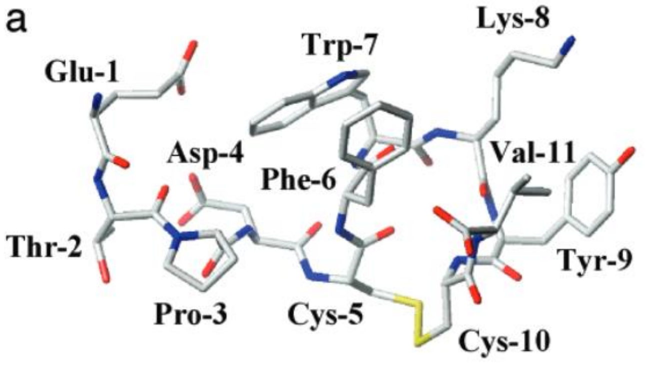
compound conformers

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



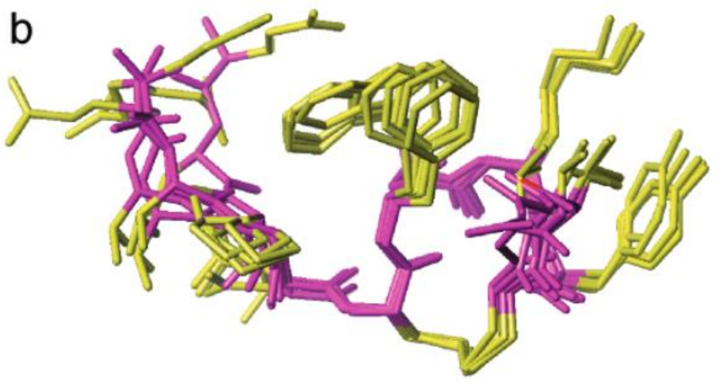
Polishchuk, P. et al. Virtual Screening Using Pharmacophore Models Retrieved from Molecular Dynamic Simulations. *International Journal of Molecular Sciences* **2019**, 20, (23), 5834.

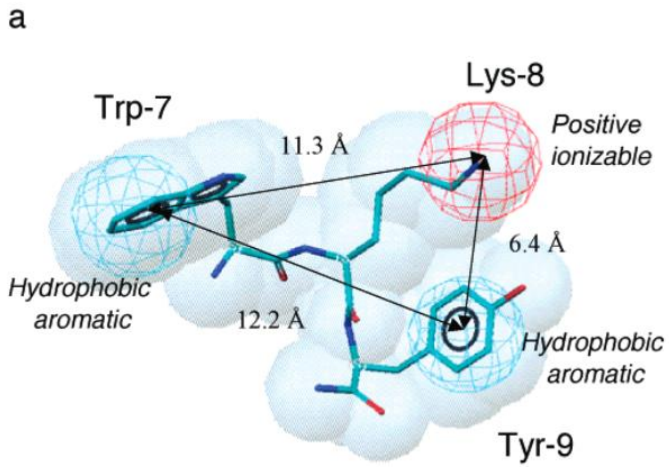
Pharmacophore example



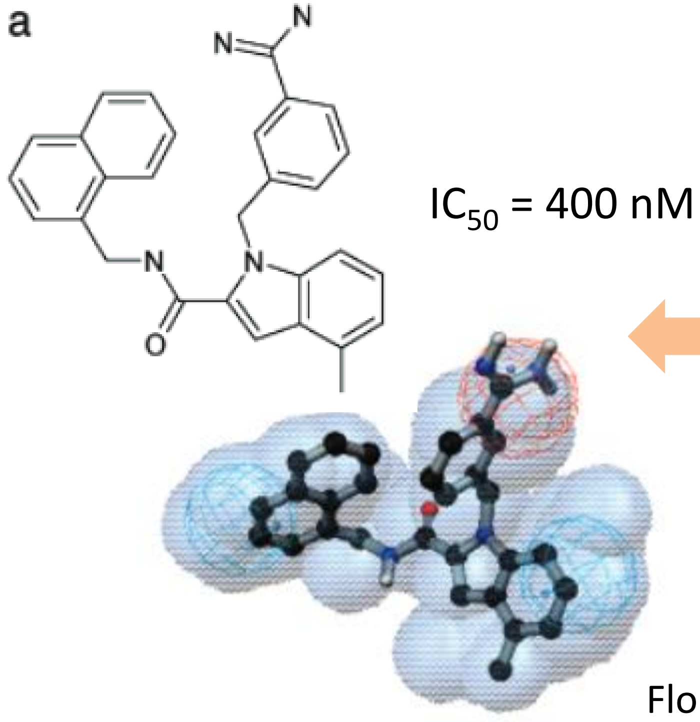
Urotensin II - ETPDc[CFWKYCV]
potent vasoconstrictor

Ala scan
NMR
MD





500 hits



IC_{50} = 400 nM

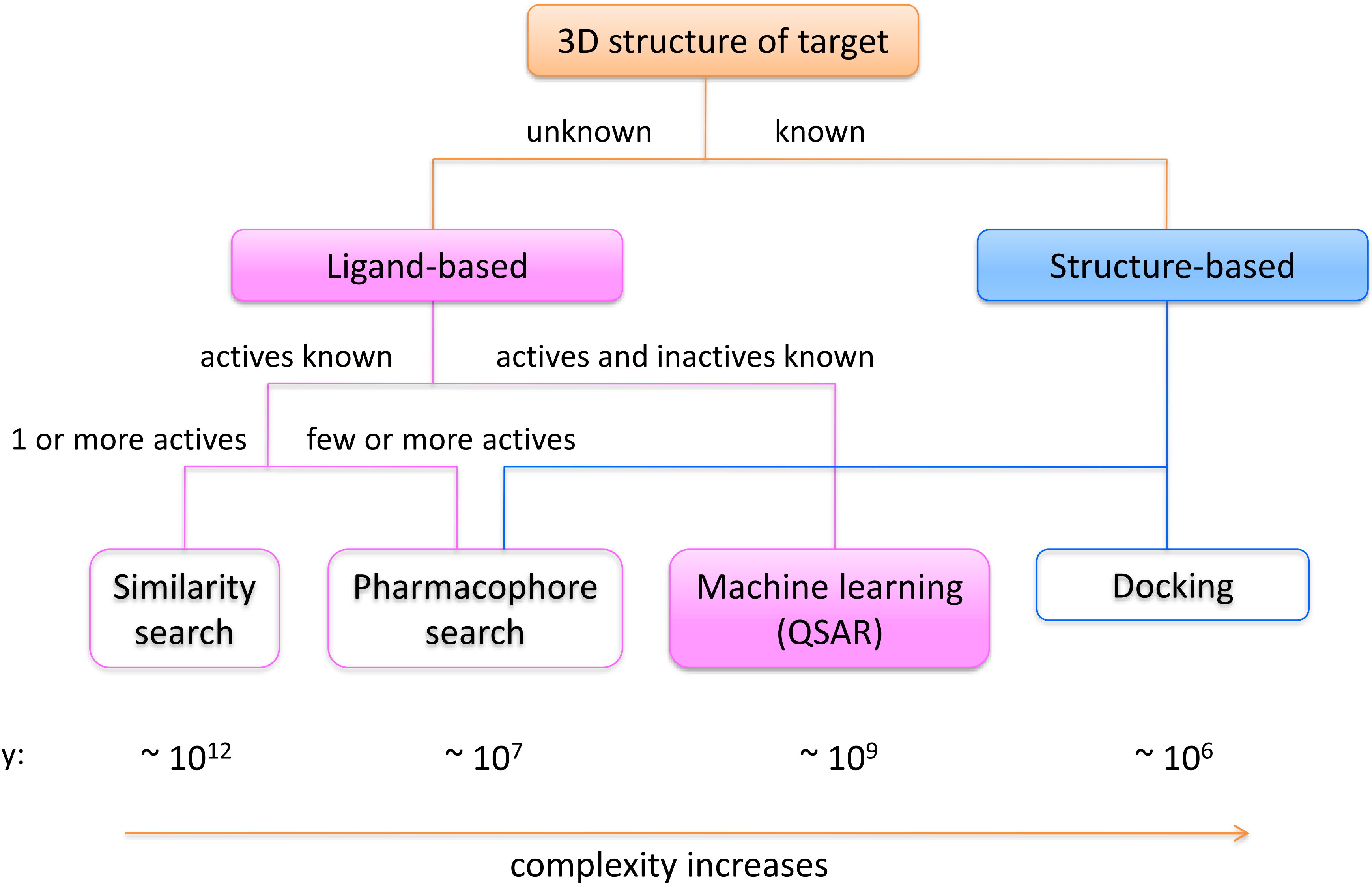
Flohr S. et al., J. Med. Chem., **2002**, 45 (9), pp 1799–1805

Pharmacophores: conlusions

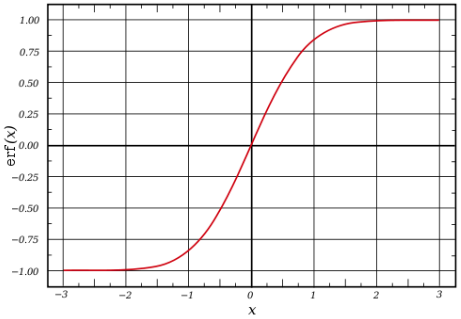
+ Universal representation of binding pattern
+ Qualitative output
+ Very fast screening
+ Scaffold hopping

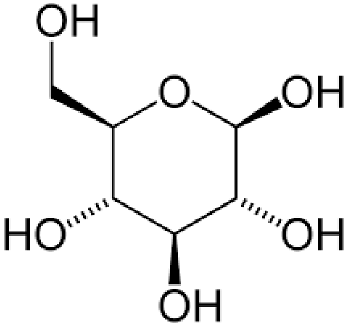
- Structure-based models can be very specific
- Ligand-based models depend on conformational sampling

Machine learning (QSAR)

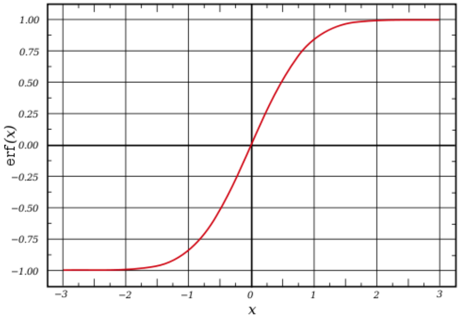


Modeling of compound properties





Activity = F(structure)



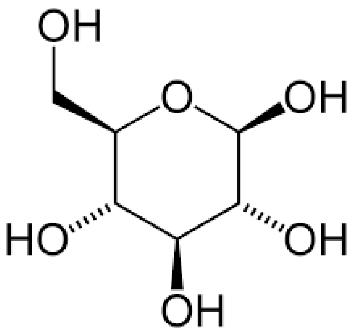
X_1	X_2	X_3	X_4	X_5	X_6	…	X_N
1	0	9	0	11	1	…	1
4	0	1	0	0	0	…	1
0	0	0	0	0	4	…	6
0	2	3	6	0	0	…	3
…	…	…	…	…	…	…	…
4	0	0	0	1	2	…	1

Activity = M(E(structure))

M – mapping function
E – encoding function

QSAR modeling workflow

Structure



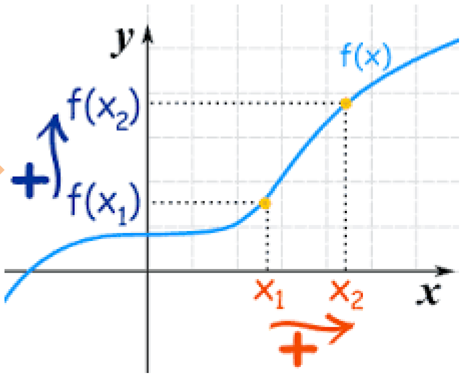
Descriptors (features)

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

Bioassays

Databases

Preprocessing

Data normalization & curation

Feature extraction

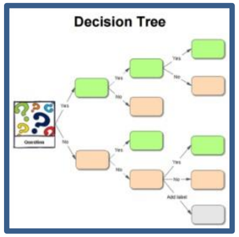
Feature engineering

$$x_i' = \frac{x_i - \bar{x}}{\sum_j z_j}$$

Feature selection

Feature combination

Model training



Classification

Regression

Clustering

Model validation

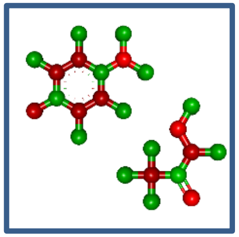
Cross-validation

Bootstrap

Test set

Applicability Domain

Interpretation



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

Examples of QSAR models

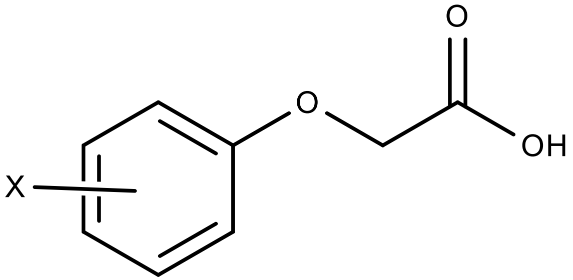
Hansch equation

plant growth inhibition activity of phenoxyacetic acids

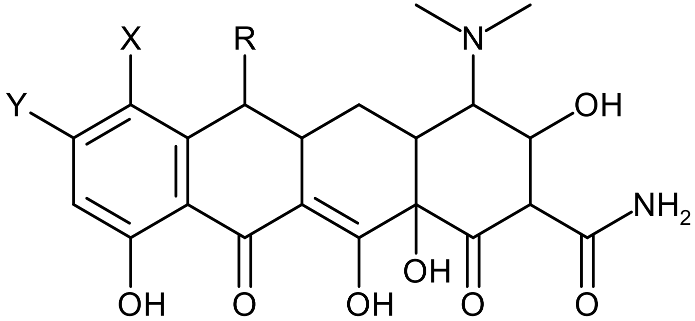
$$1/C = 4.08\pi - 2.14\pi^2 + 2.78\sigma + 3.38$$

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

σ - Hammet constant



Free-Wilson models



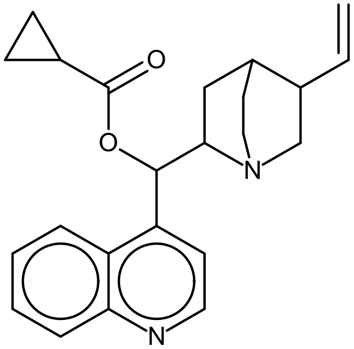
Inhibition activity of compounds against *Staphylococcus aureus*

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

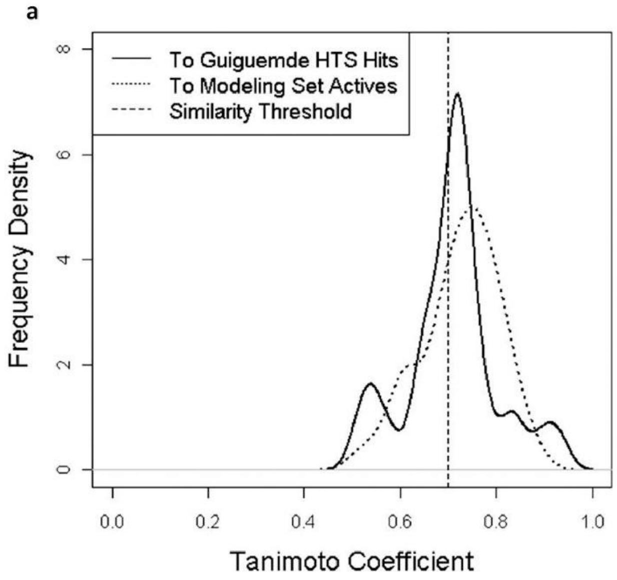
$$Act = 75R_H - 112R_{CH3} + 84X_{Cl} - 16X_{Br} - 26X_{NO2} + 123Y_{NH2} + 18Y_{NHC(=O)CH3} - 218Y_{NO2}$$

QSAR: example

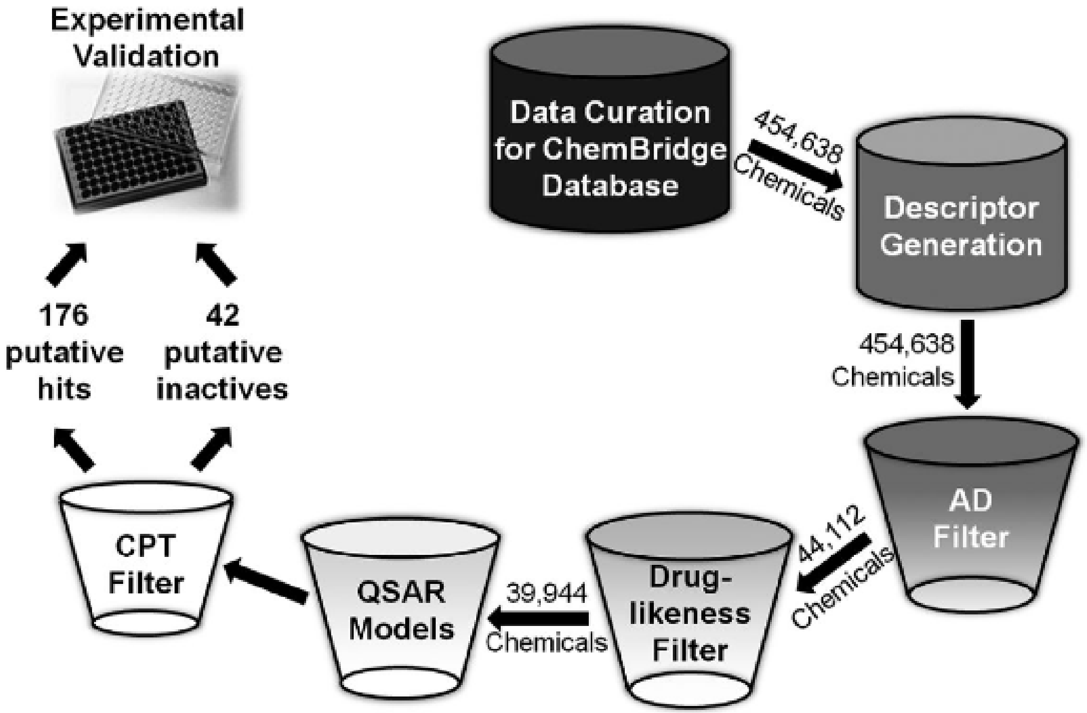
Antimalarial activity



EC_{50} = 95 nM



7 hits, EC_{50} < 2µM



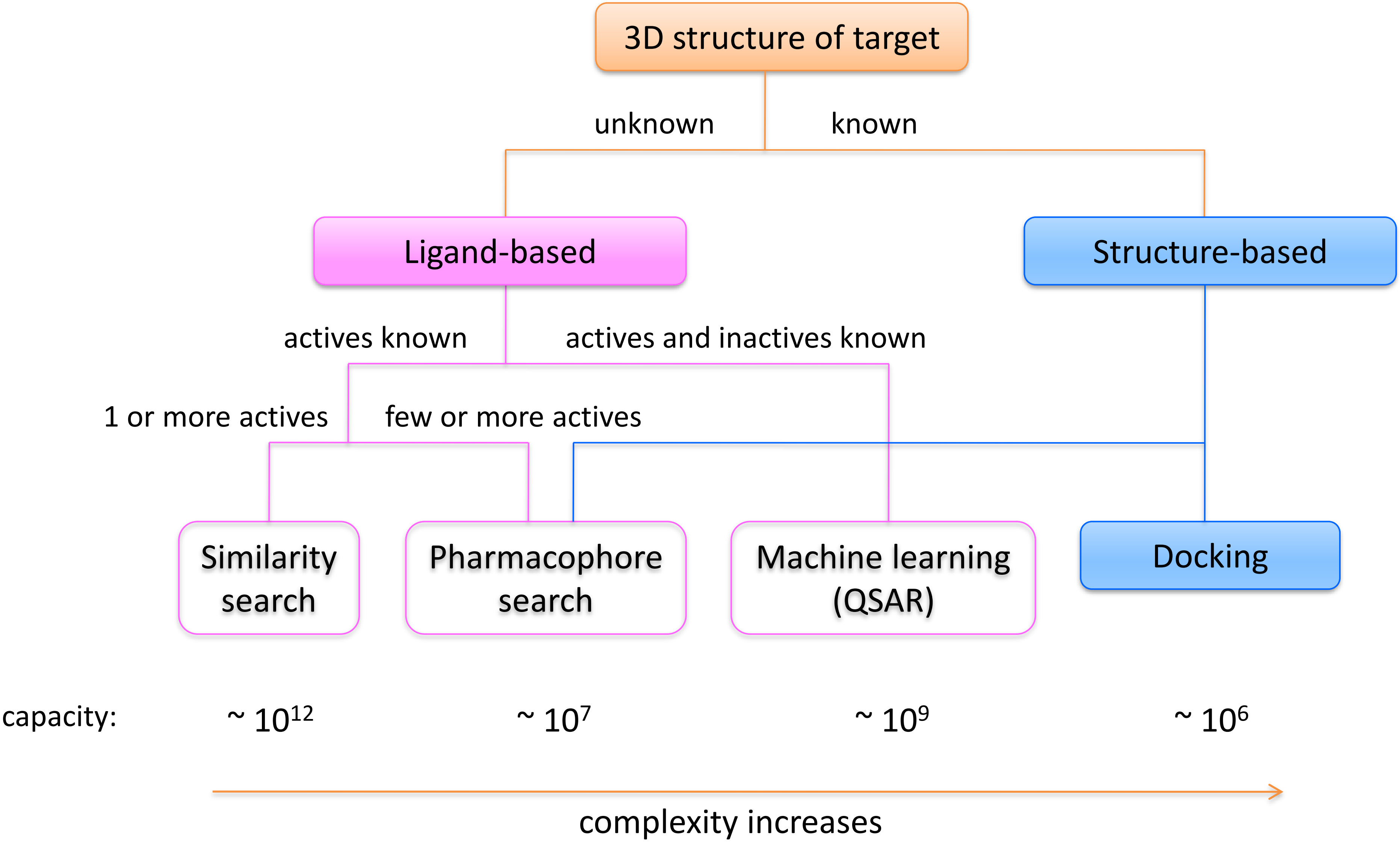
Zhang L. et al., J. Chem. Inf. Model., **2013**, 53 (2), pp 475–492

QSAR: conclusions

+ 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



3D structure of target

unknown

known

Ligand-based

Structure-based

actives known

actives and inactives known

1 or more actives

few or more actives

Similarity search

Pharmacophore search

Machine learning (QSAR)

Docking

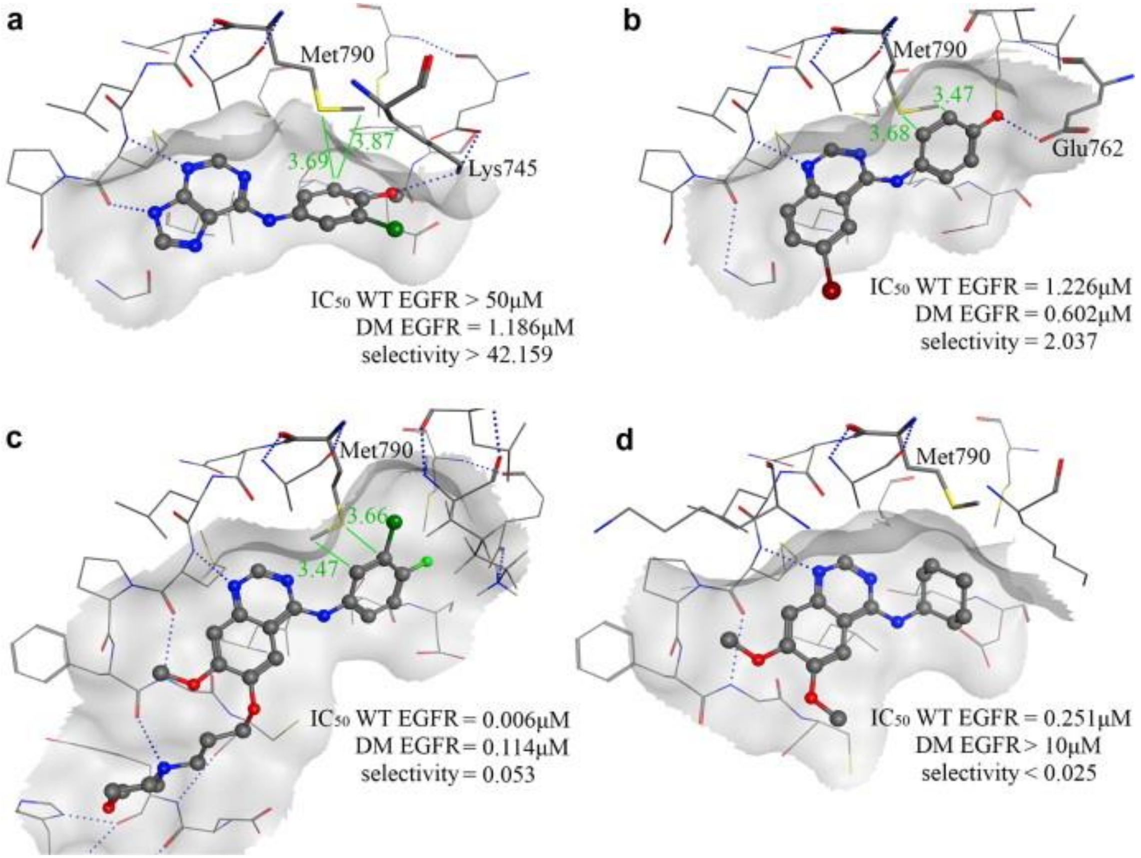
capacity: ~ 10^{12} ~ 10^{7} ~ 10^{9} ~ 10^{6}

complexity increases

Molecular docking predictions

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 complex?

Complex 3D jigsaw puzzle

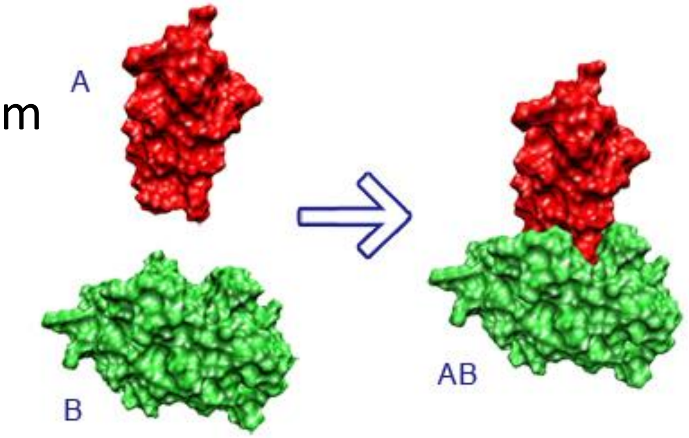
Conformational flexibility – many degrees of freedom

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.”

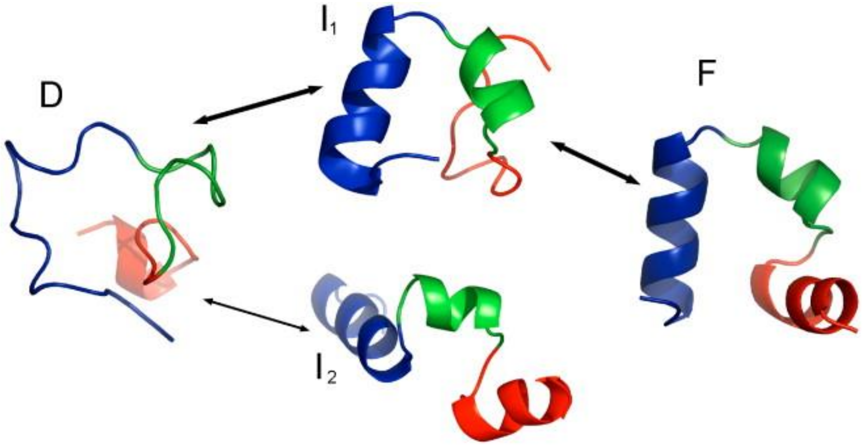
(Leach A.R., Shoichet B.K., Peishoff C.E..
J. Med. Chem. 2006, *49*, 5851-5855)

Sampling and scoring

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

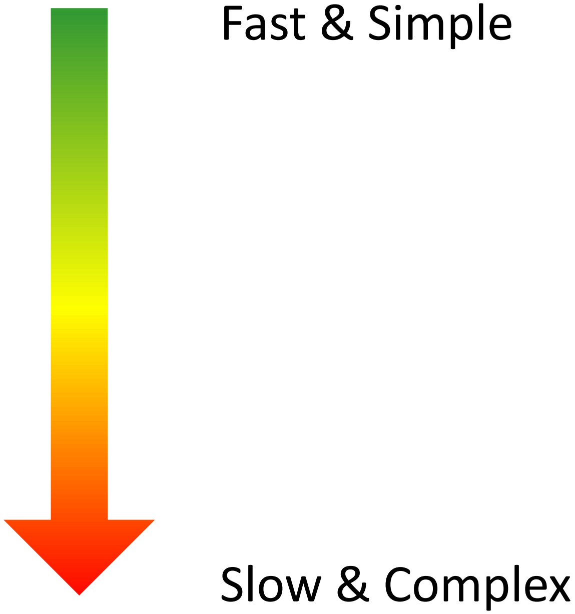
1. **Search algorithm (sampling)** - generates a large number of poses of a molecule in the binding site.

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

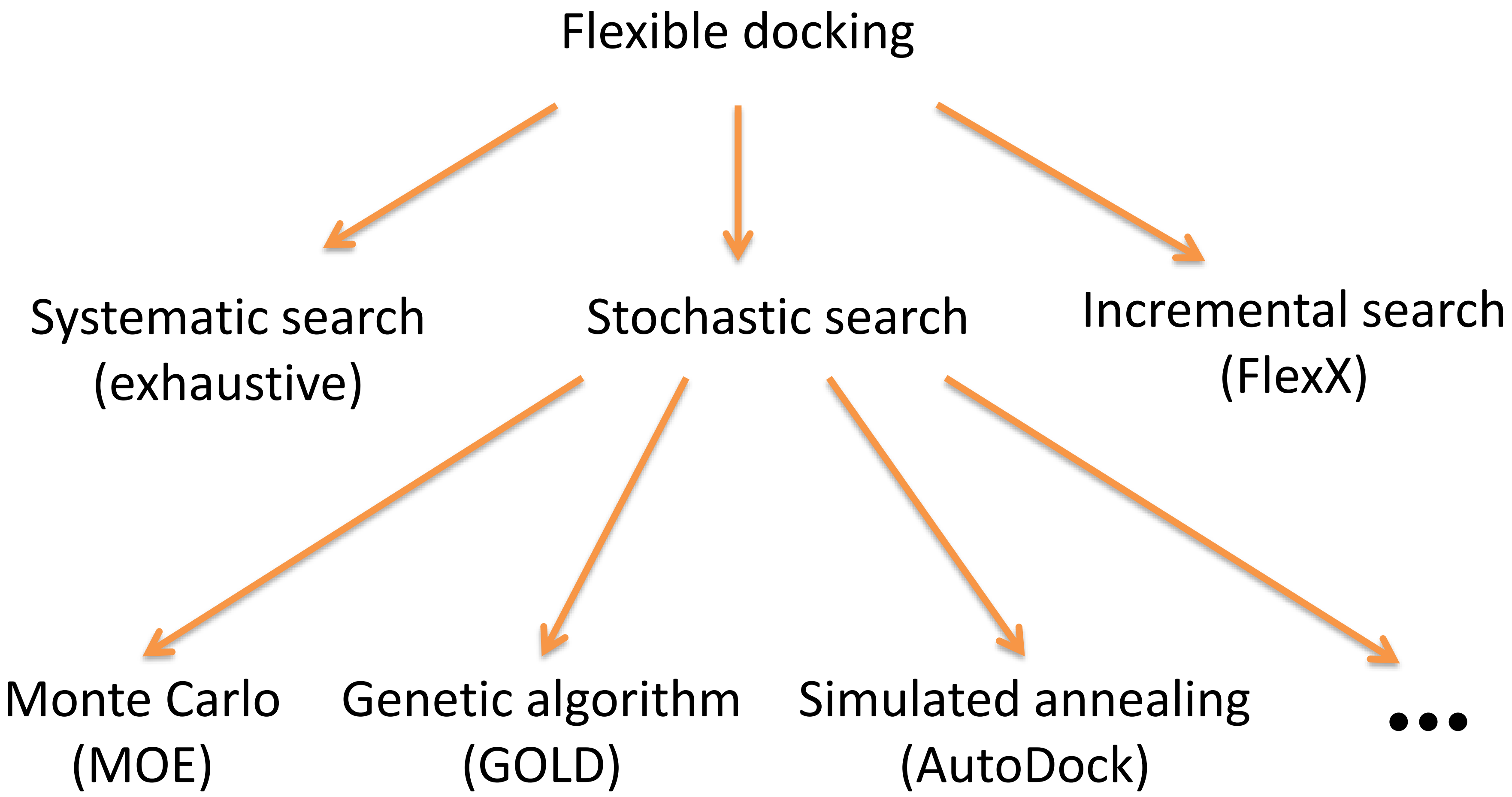


Search algorithms (sampling)

Ligand	Receptor
Rigid	Rigid
Flexible	Rigid
Flexible	Flexible



Search algorithms (sampling)



Flexible docking

Systematic search (exhaustive)

Stochastic search

Incremental search (FlexX)

Monte Carlo (MOE)

Genetic algorithm (GOLD)

Simulated annealing (AutoDock)

…

Classes of scoring functions

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

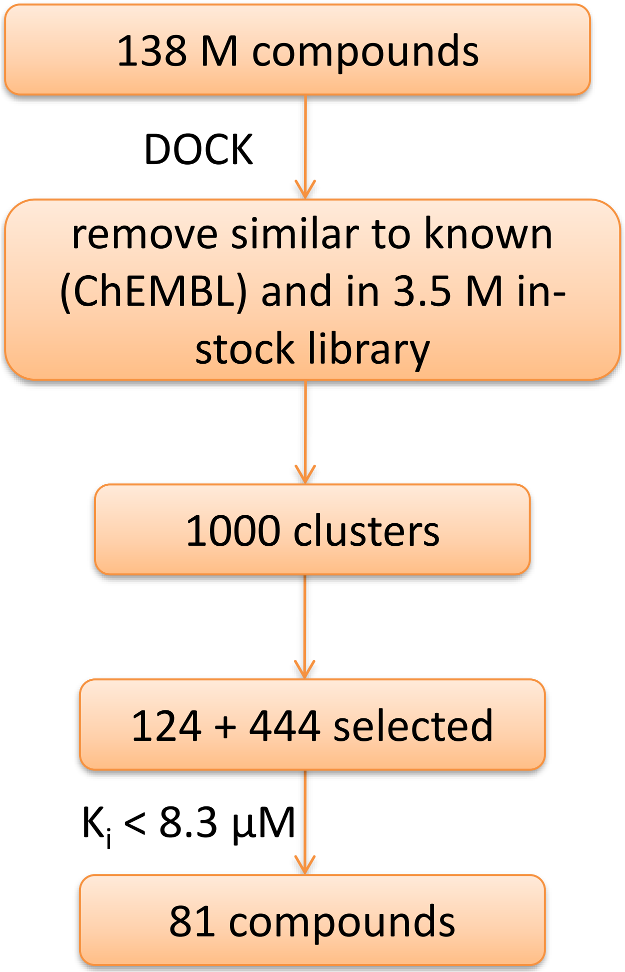
Based on statistical analysis of observed pairwise distributions

PMF, DrugScore, ASP

Molecular docking: example

ligands of D4 receptor

enumerated library



138 M compounds

DOCK

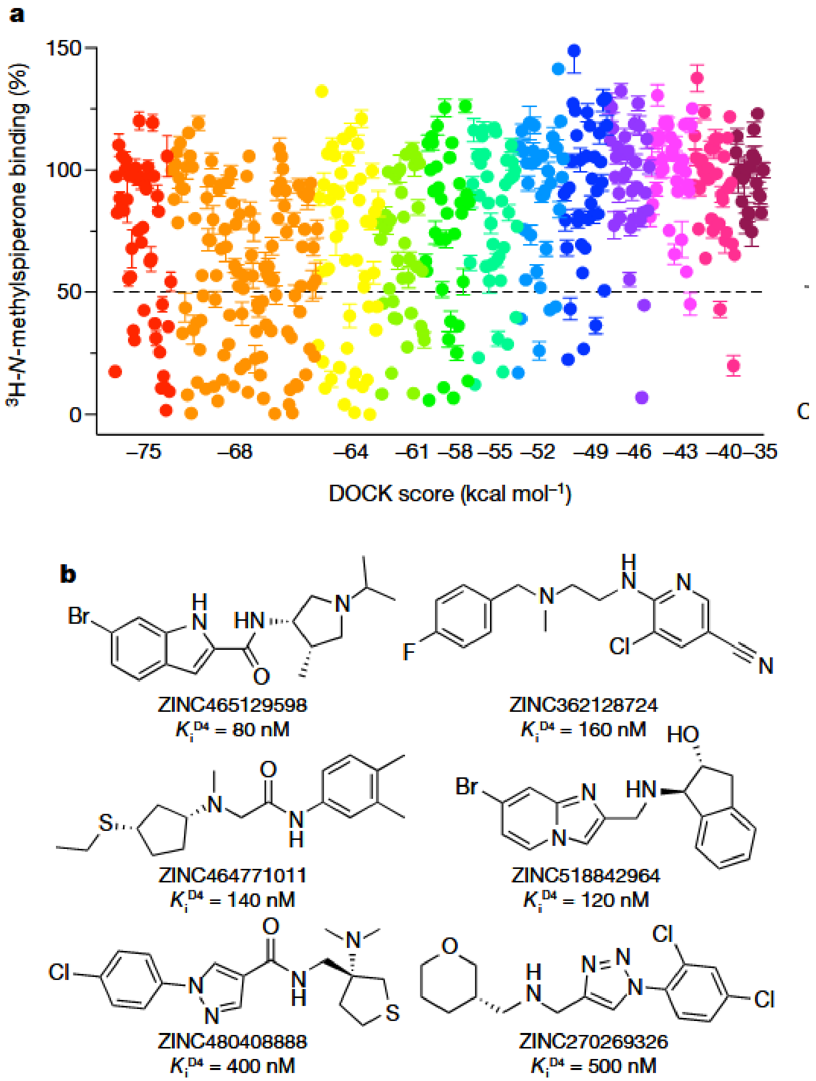
remove similar to known (ChEMBL) and in 3.5 M in-stock library

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: conclusions

+ Relatively fast

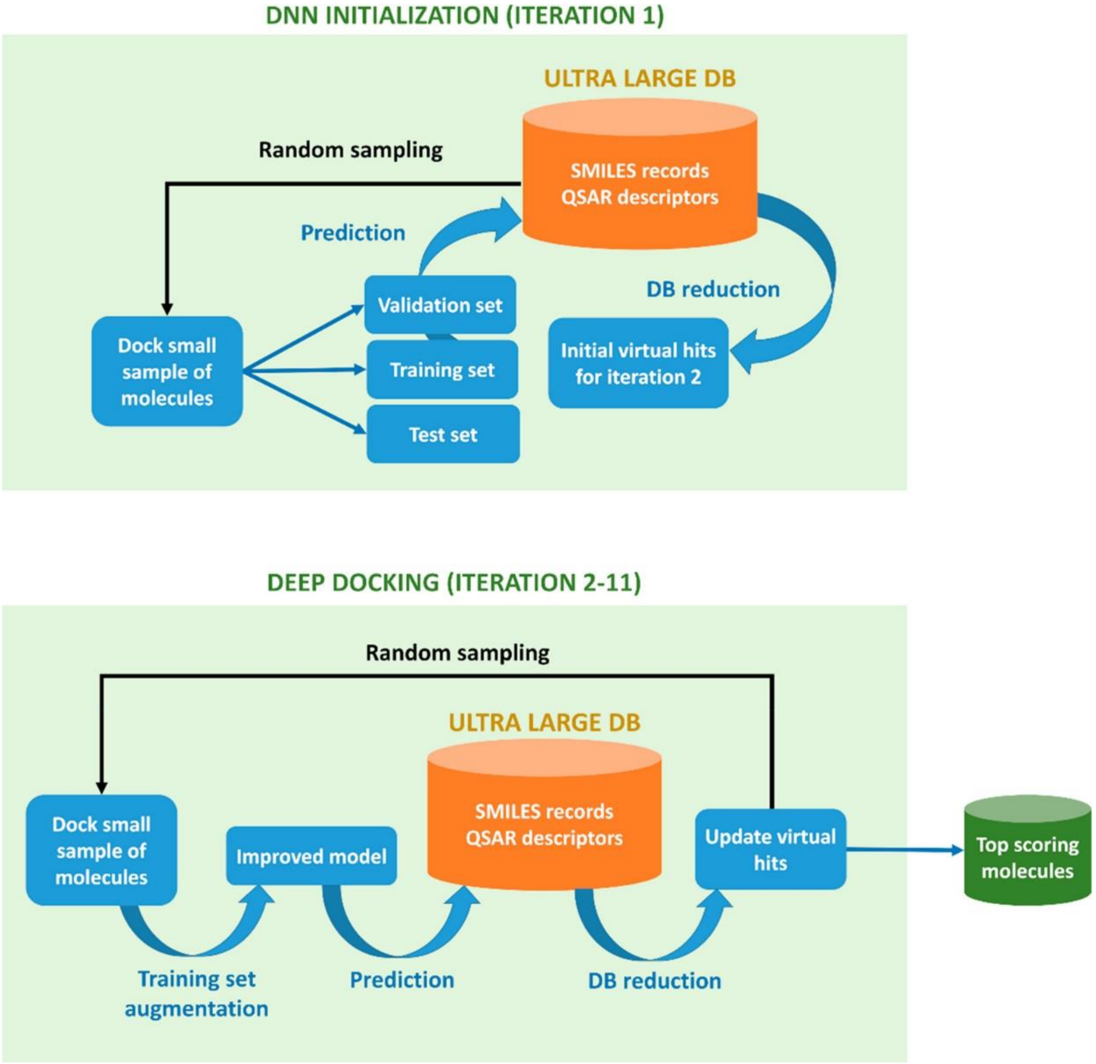
+ Determine binding poses

+ Good in ranking ligands for virtual screening

- Low accuracy of binding energy estimation

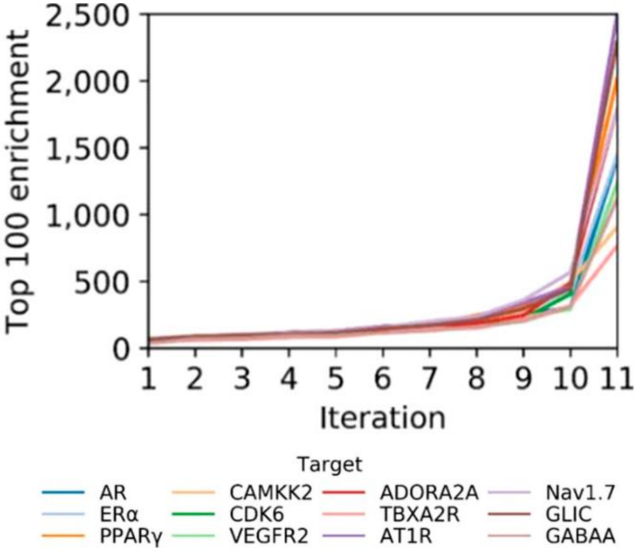
- Require knowledge about binding site

Deep docking (surrogate modeling)



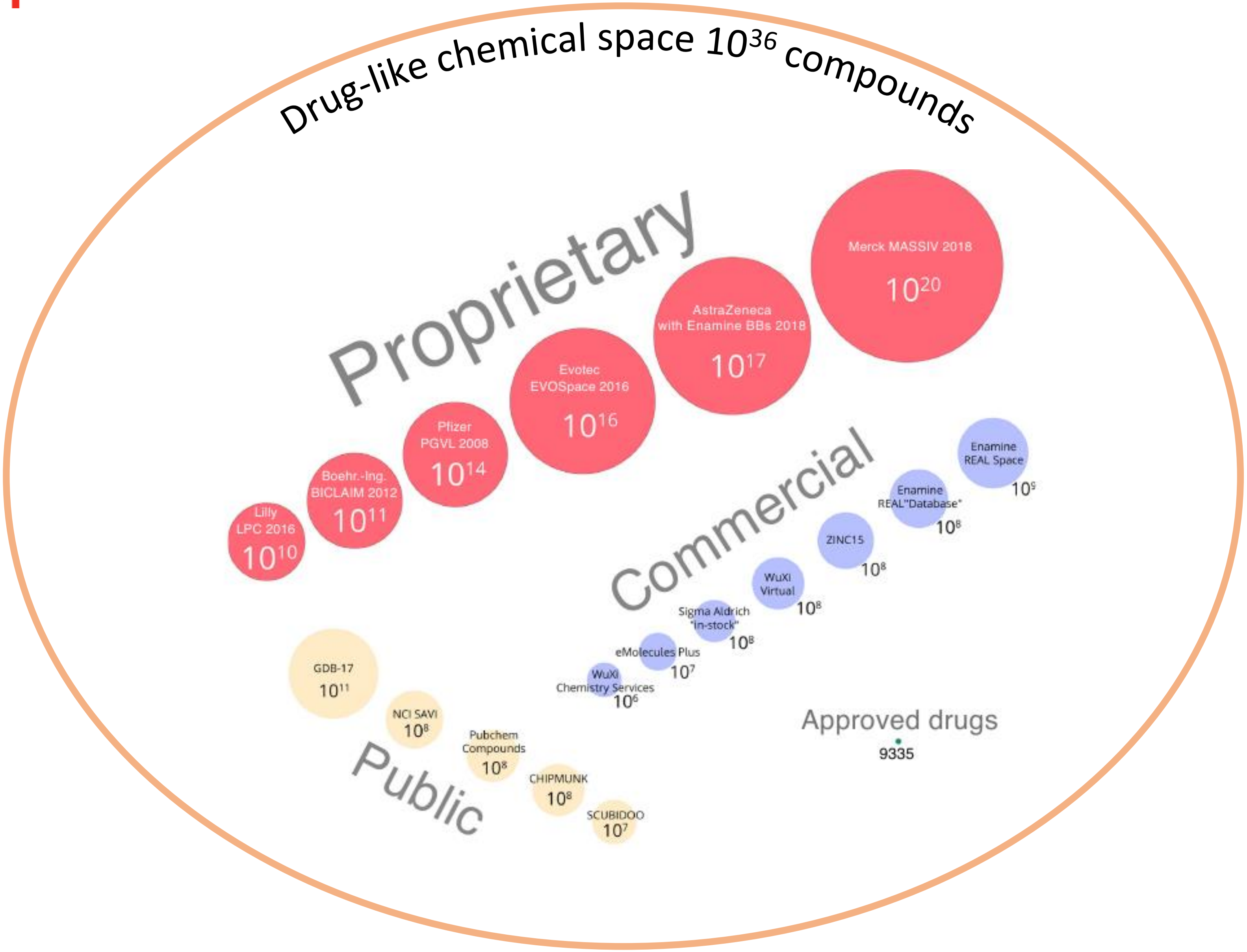
1.38B compounds ZINC15

1M compounds / iteration



Gentile, F.; Agrawal, V.; Hsing, M.; Ton, A.-T.; Ban, F.; Norinder, U.; Gleave, M. E.; Cherkasov, A. Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent. Sci.* **2020**, 6 (6), 939-949

Vastness of chemical space



Hoffmann, T.; Gastreich, M., The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discovery Today* **2019**, 24, 1148-1156