Docking and virtual screening

- What is virtual screening?
- Pharmacophore searching
- Shape-based searching
- Docking
- Estimating model quality

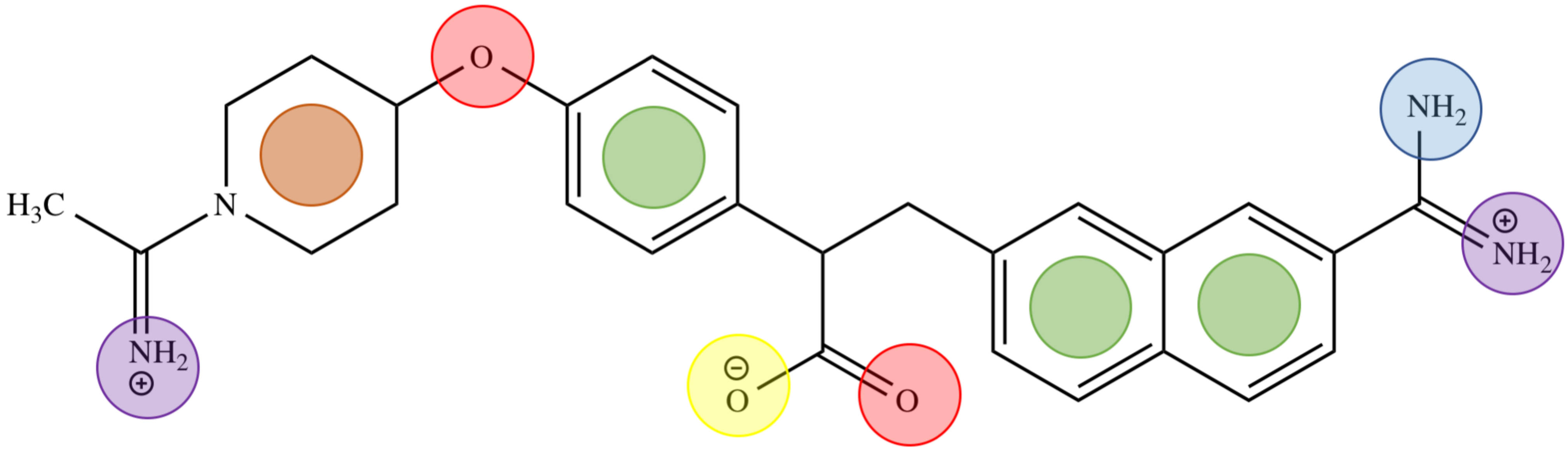
What is virtual screening (VS)?

- Identification of interesting molecules out of a database of (virtual) molecules
- Ligand-based VS
  - Chemo-informatics
  - Pharmacophore searching
  - Shape-based searching
- Protein structure-based VS
  - Docking

Docking and virtual screening

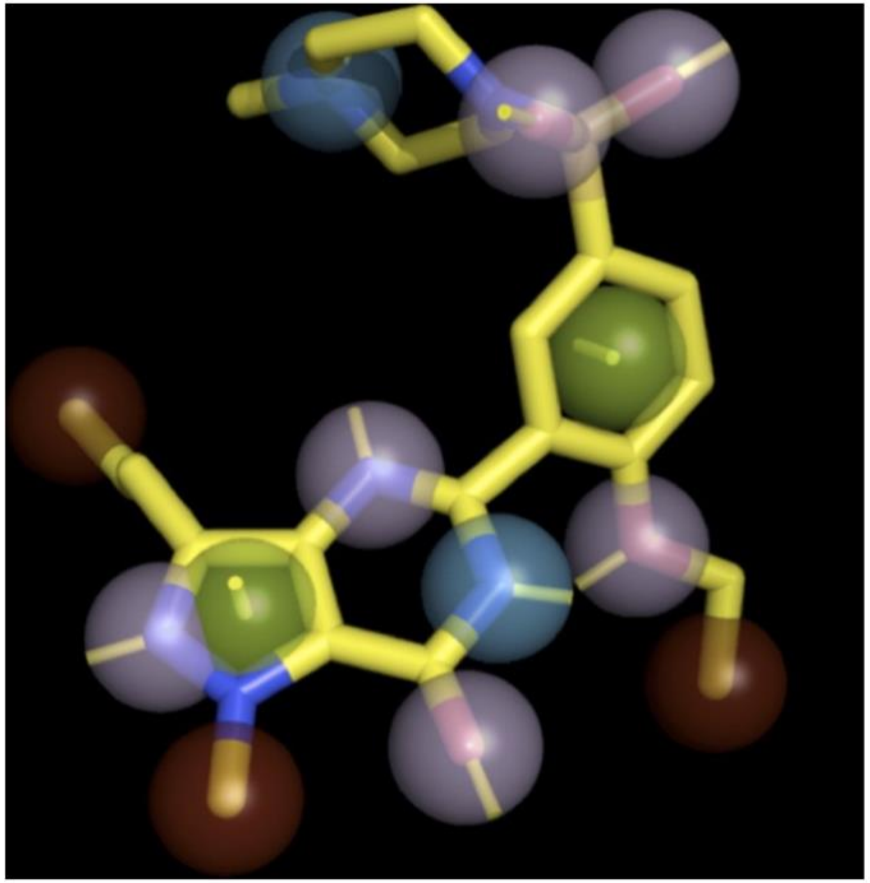
- What is virtual screening?
- Pharmacophore searching
- Shape-based searching
- Docking
- Estimating model quality

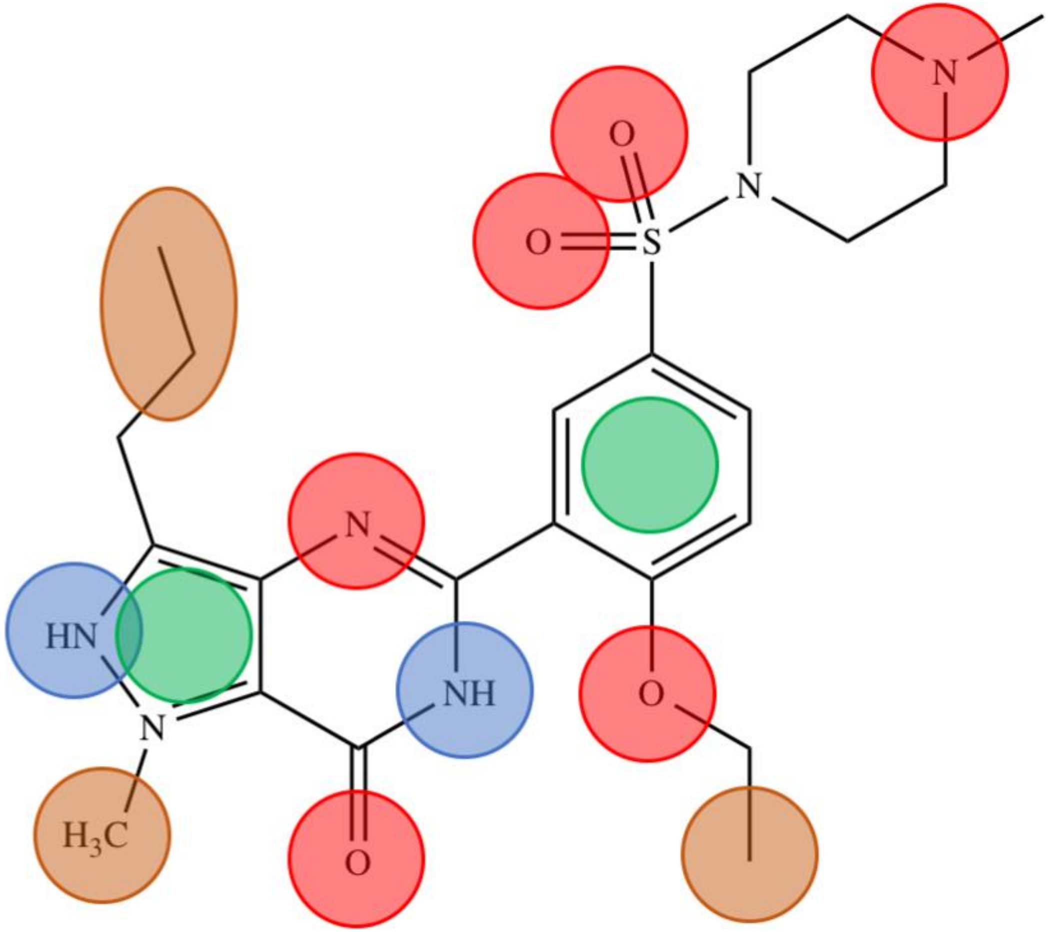
What is a pharmacophore?



Pharmacophore types

Code	Description	Normal
AROM	Aromatic ring	Yes
HDON	Hydrogen bond donor	Yes
HACC	Hydrogen bond acceptor	Yes
LIPO	Lipophilic (hydrophobic) region	No
POSC	Positive charge center	No
NEGC	Negative charge center	No
HYBH	Hydrogen bond donor and hydrogen bond acceptor	Yes
HYBL	Aromatic and lipophilic ring	Yes
EXCL	Exclusion sphere	No





O
O
S
N
N
N
N
HN
NH
N
H3C
O
O

Gaussian representation of points

$$V = \int p \, e^{\left(-\frac{|m-r|^2}{\sigma}\right)} dr$$

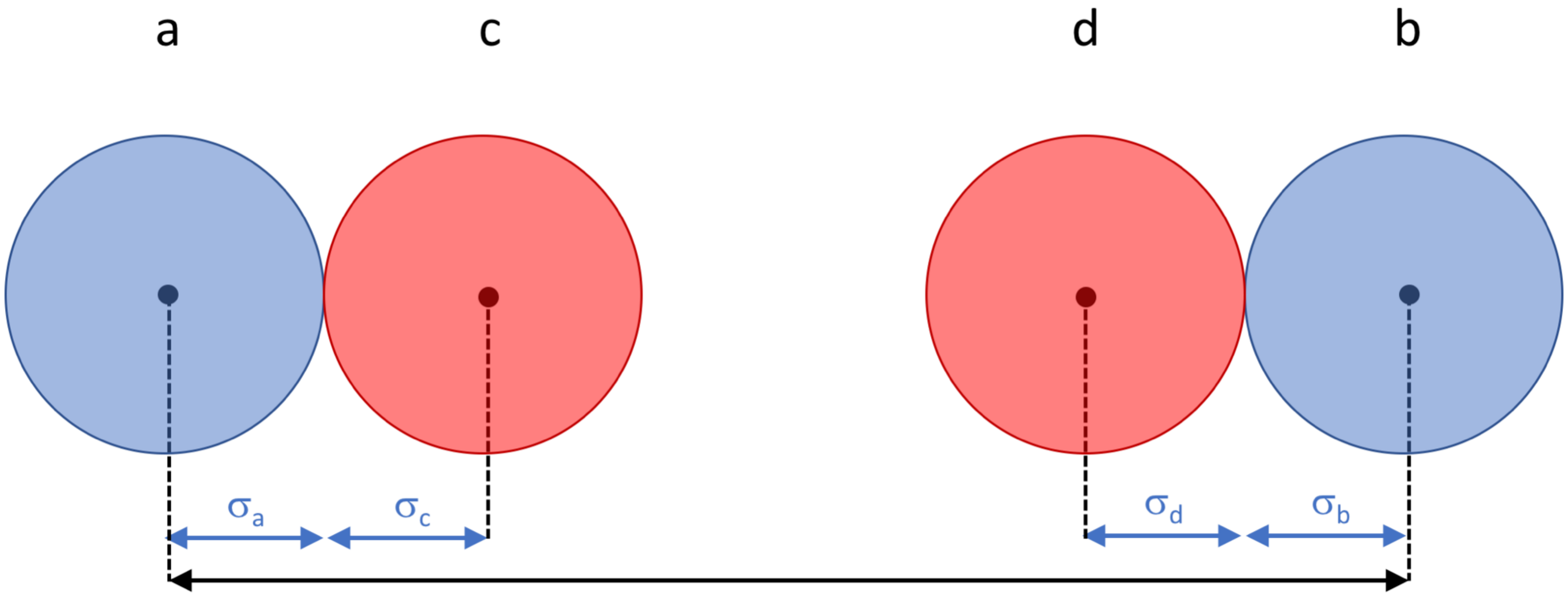
With:

p: scaling constant

m: position in space

σ: spread

Feature mapping



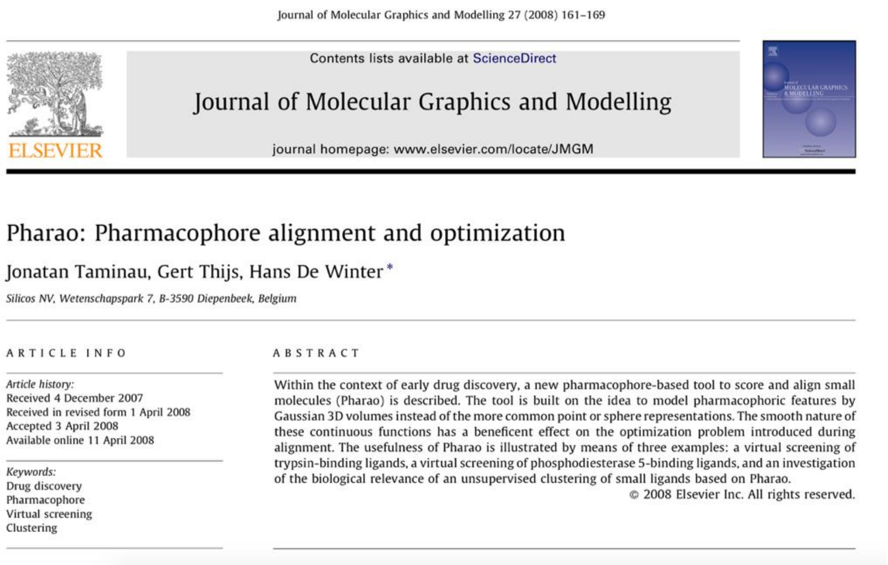
$$\varepsilon = \frac{|d_{ab} - d_{cd}|}{\sigma_a + \sigma_b + \sigma_c + \sigma_d}$$

Calculating the overlap

- The pharmacophore spheres are represented by Gaussian spheres, hence easy to calculate the overlap
- $TANIMOTO = \frac{V_{overlap}}{V_1 + V_2 - V_{overlap}}$
- $TVERSKY = \frac{V_{overlap}}{V_1}$

Popular pharmacophore searching programs

- Open source: [Pharao]


Journal of Molecular Graphics and Modelling 27 (2008) 161–169

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

ELSEVIER

Pharao: Pharmacophore alignment and optimization

Jonatan Taminau, Gert Thijs, Hans De Winter *

Silicos NV, Wetenschapspark 7, B-3590 Diepenbeek, Belgium

ARTICLE INFO

Article history:
Received 4 December 2007
Received in revised form 1 April 2008
Accepted 3 April 2008
Available online 11 April 2008

Keywords:
Drug discovery
Pharmacophore
Virtual screening
Clustering

ABSTRACT

Within the context of early drug discovery, a new pharmacophore-based tool to score and align small molecules (Pharao) is described. The tool is built on the idea to model pharmacophoric features by Gaussian 3D volumes instead of the more common point or sphere representations. The smooth nature of these continuous functions has a beneficent effect on the optimization problem introduced during alignment. The usefulness of Pharao is illustrated by means of three examples: a virtual screening of trypsin-binding ligands, a virtual screening of phosphodiesterase 5-binding ligands, and an investigation of the biological relevance of an unsupervised clustering of small ligands based on Pharao.

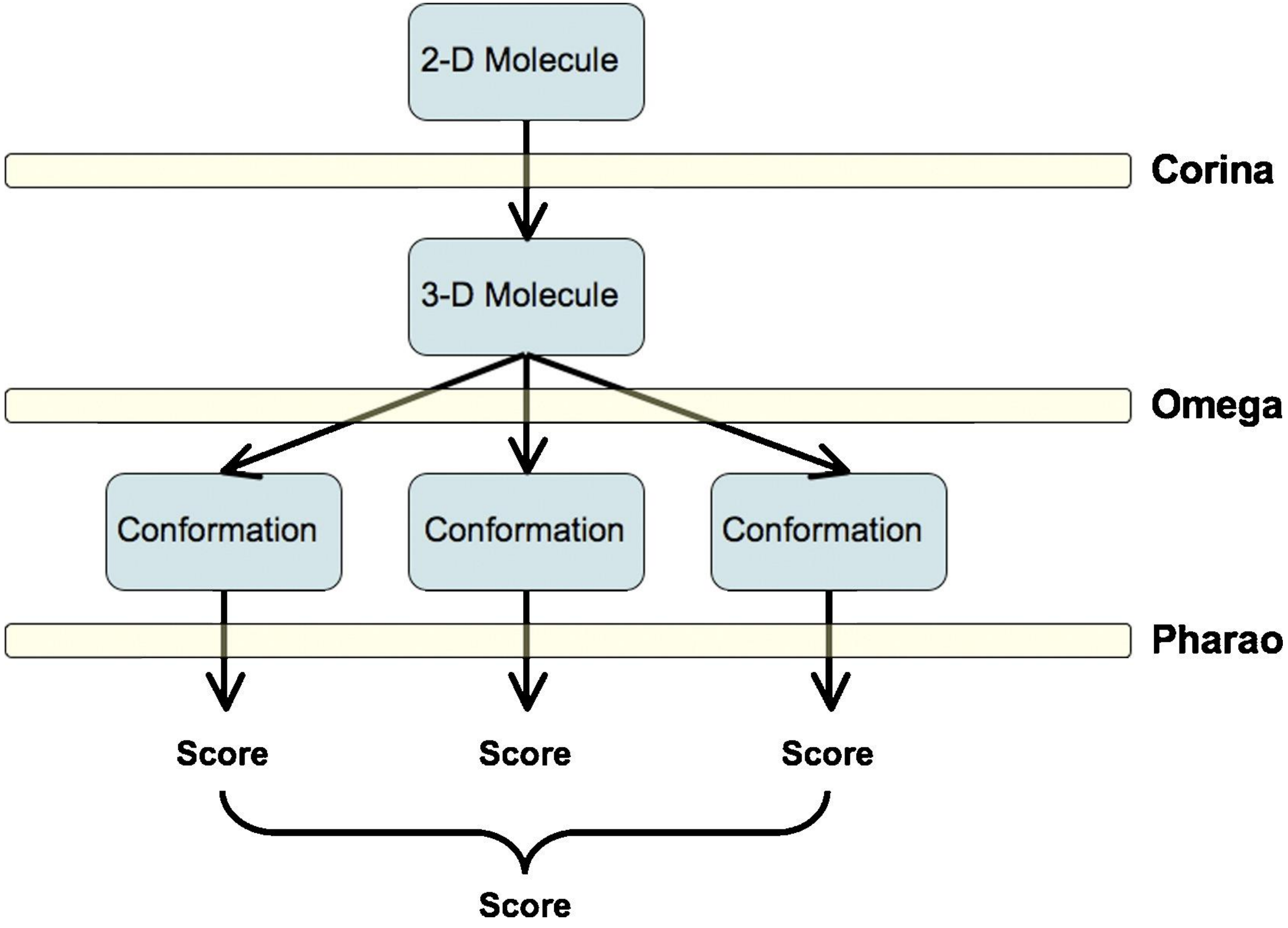
© 2008 Elsevier Inc. All rights reserved.

Taminau, J.; Thijs, G. & De Winter, H. (2008) *J. Mol. Graph. Model.* **27**, 161-169.

- Commercial: [Rocs]

- Commercial: [Phase (Schrödinger)]

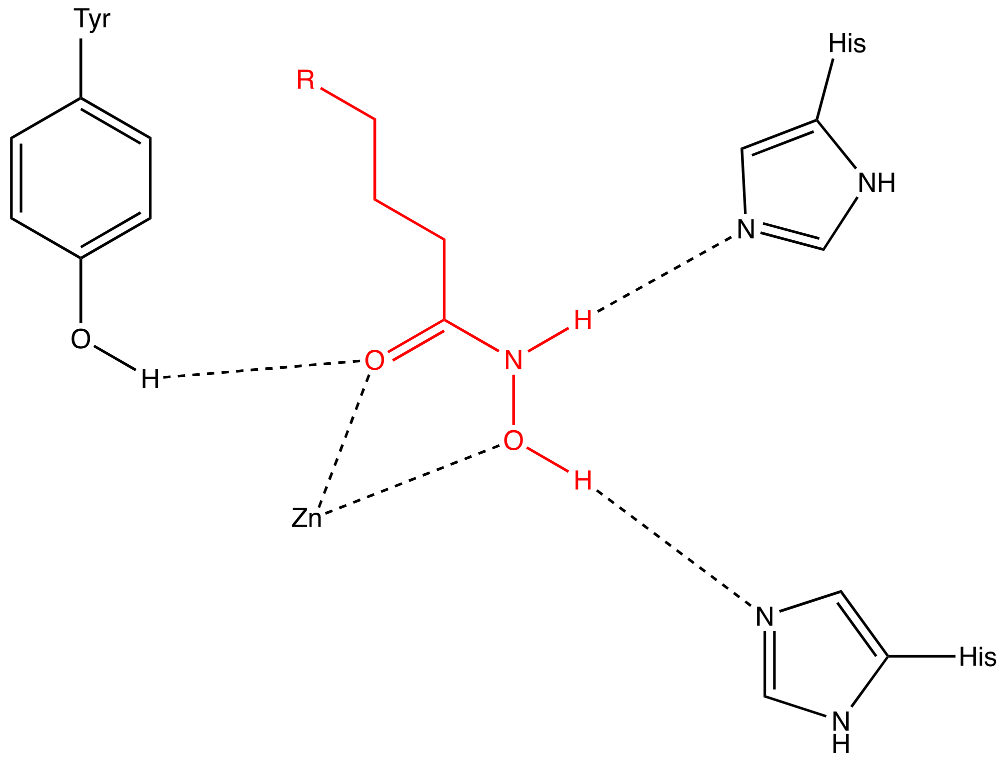
Pharao workflow



Case study

- HDAC inhibitors

  Crystal structure with SAHA



https://github.com/UAMCAntwerpen/2040FBDBIC/blob/master/Topic_03/HDAC-SAHA.pdb

Docking and virtual screening

- What is virtual screening?
- Pharmacophore searching
- Shape-based searching
- Docking
- Estimating model quality

Gaussian representation of points

$$V = \int p\, e^{\left(-\frac{|m-r|^2}{\sigma}\right)} dr$$

With:

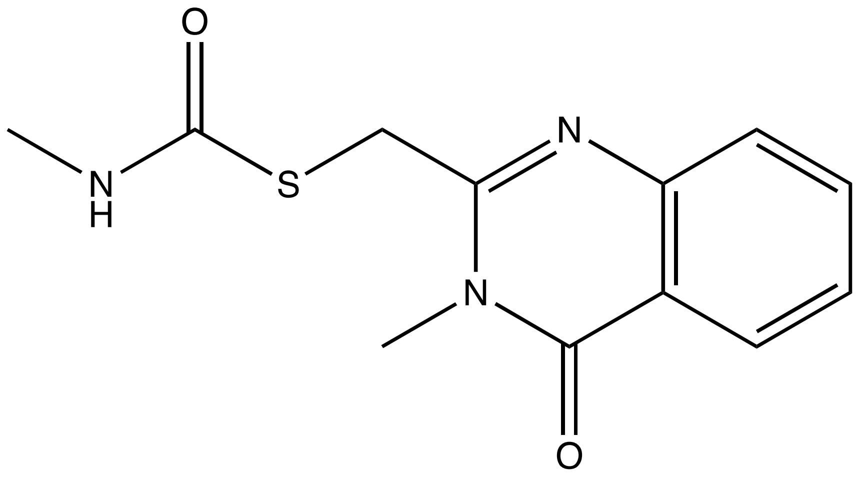
p: scaling constant

m: position in space

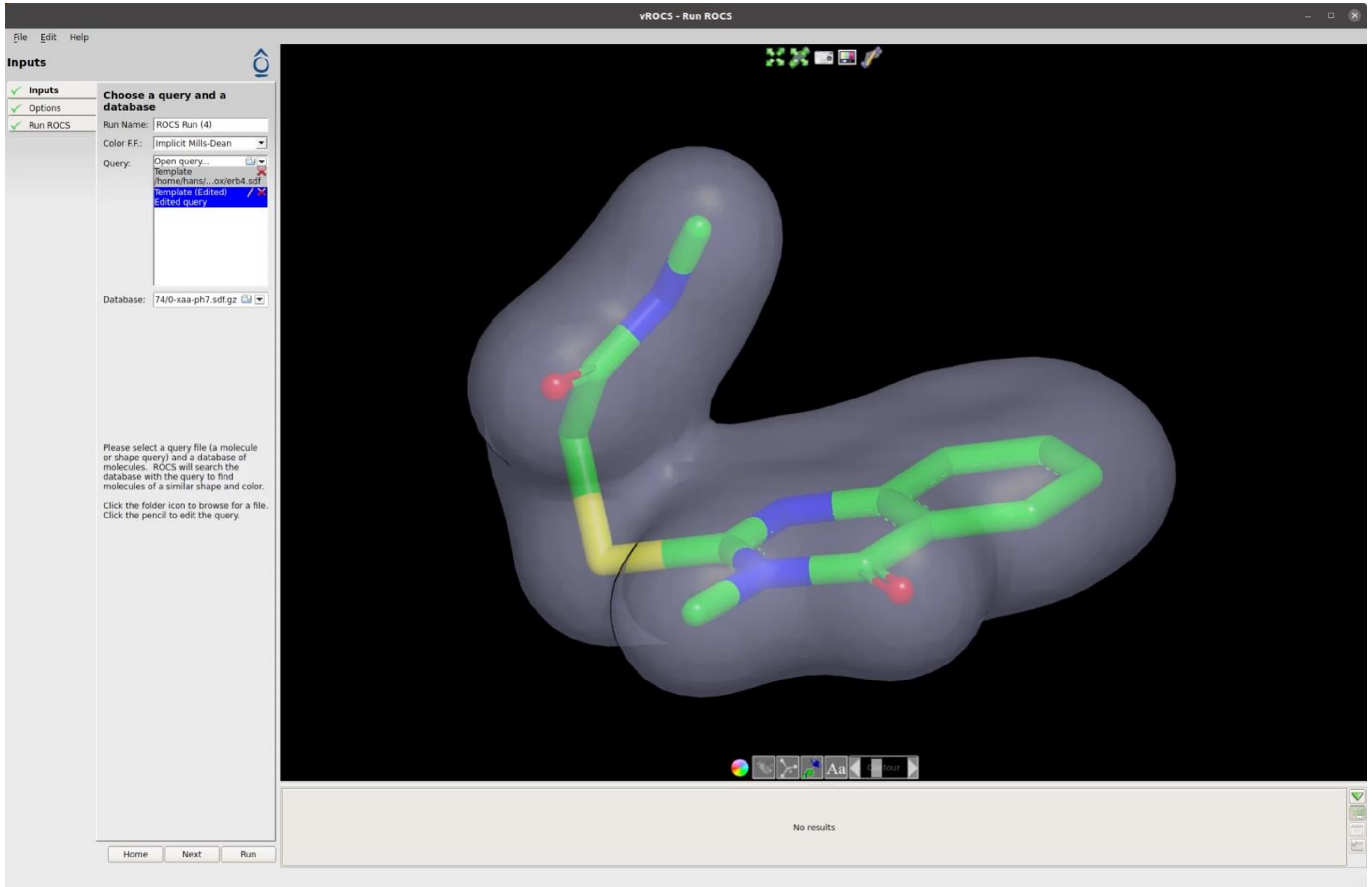
σ: spread

Case study

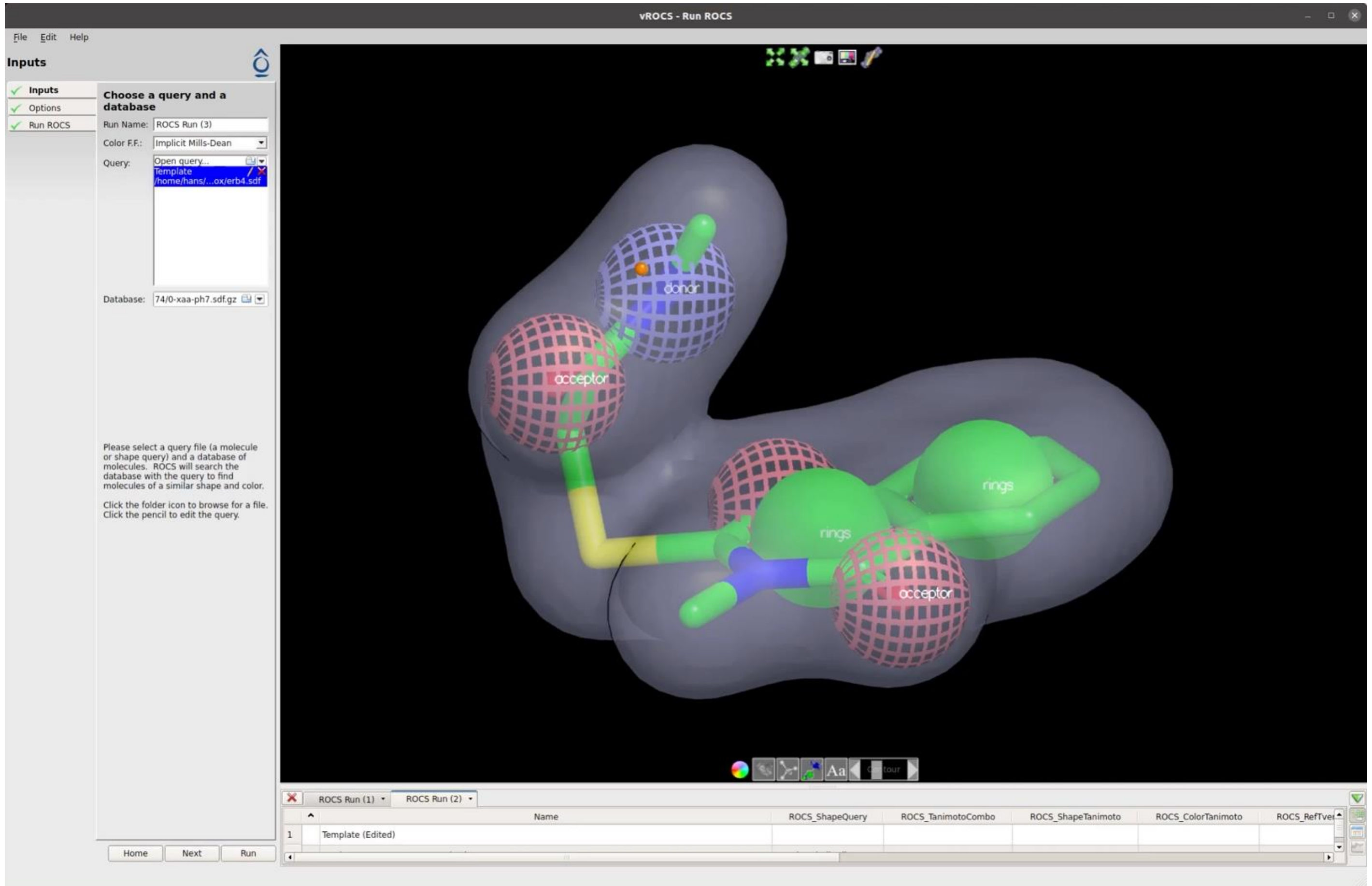
- Erb4 activators



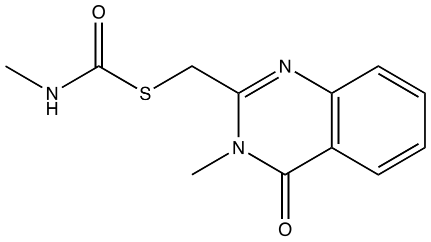
Using only the shape...



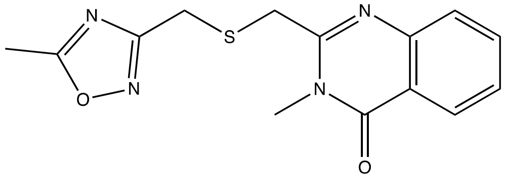
Or the shape with pharmacophoric points...

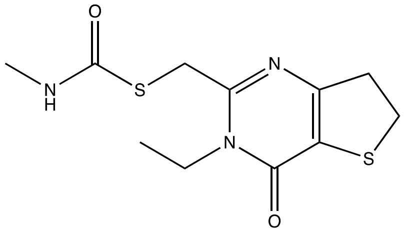


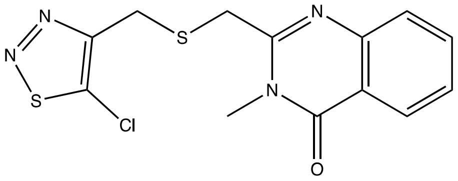
ROCS results



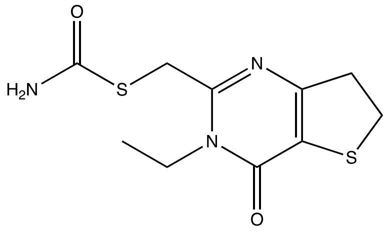
Using only the shape...

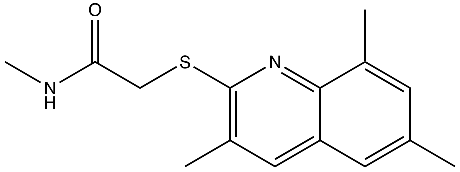


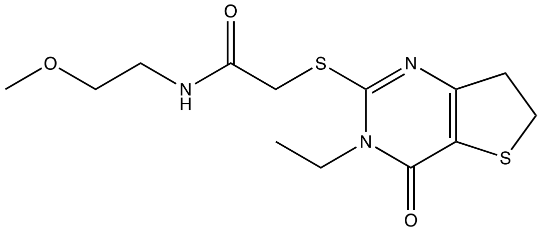




...or with pharmacophore info



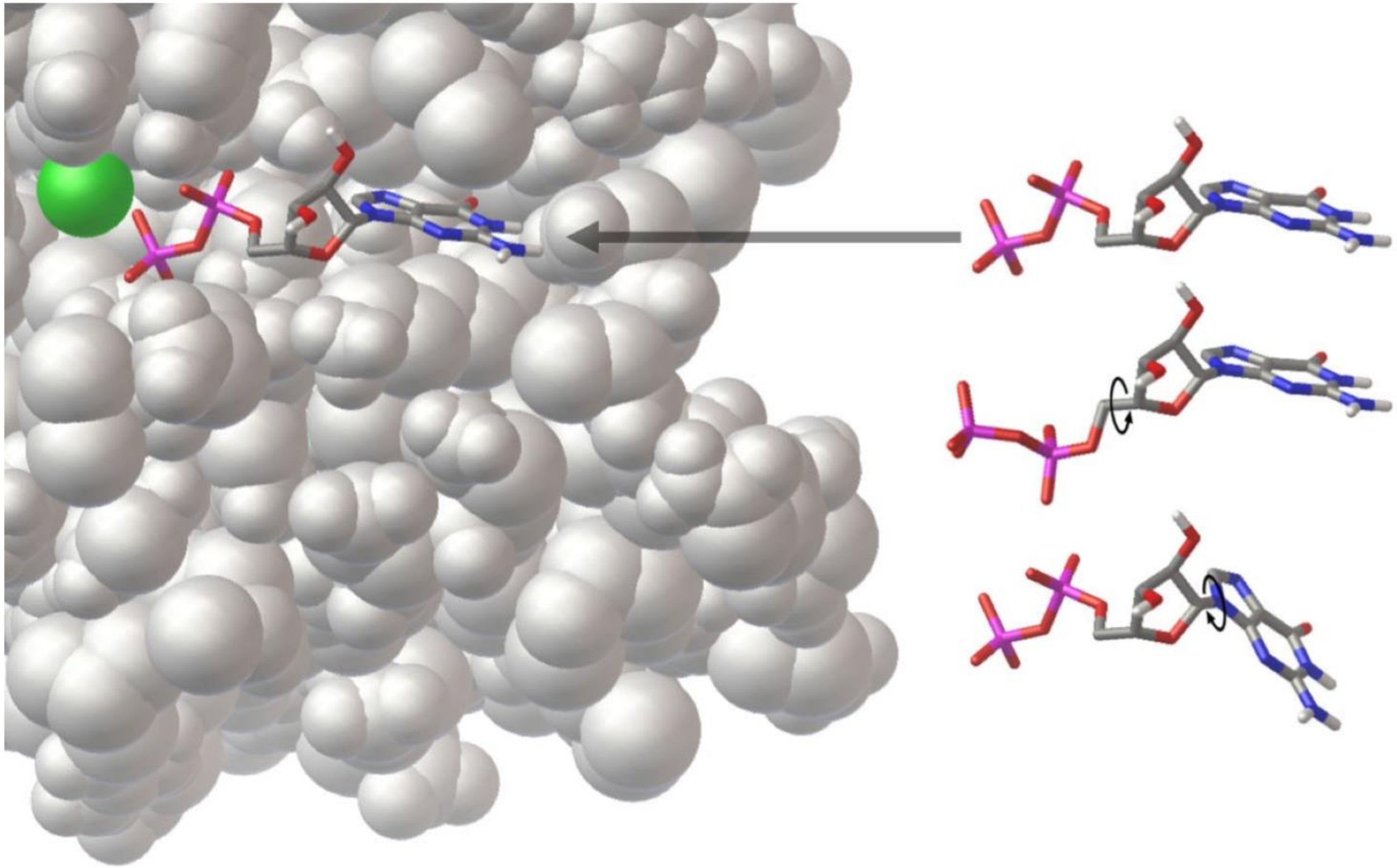




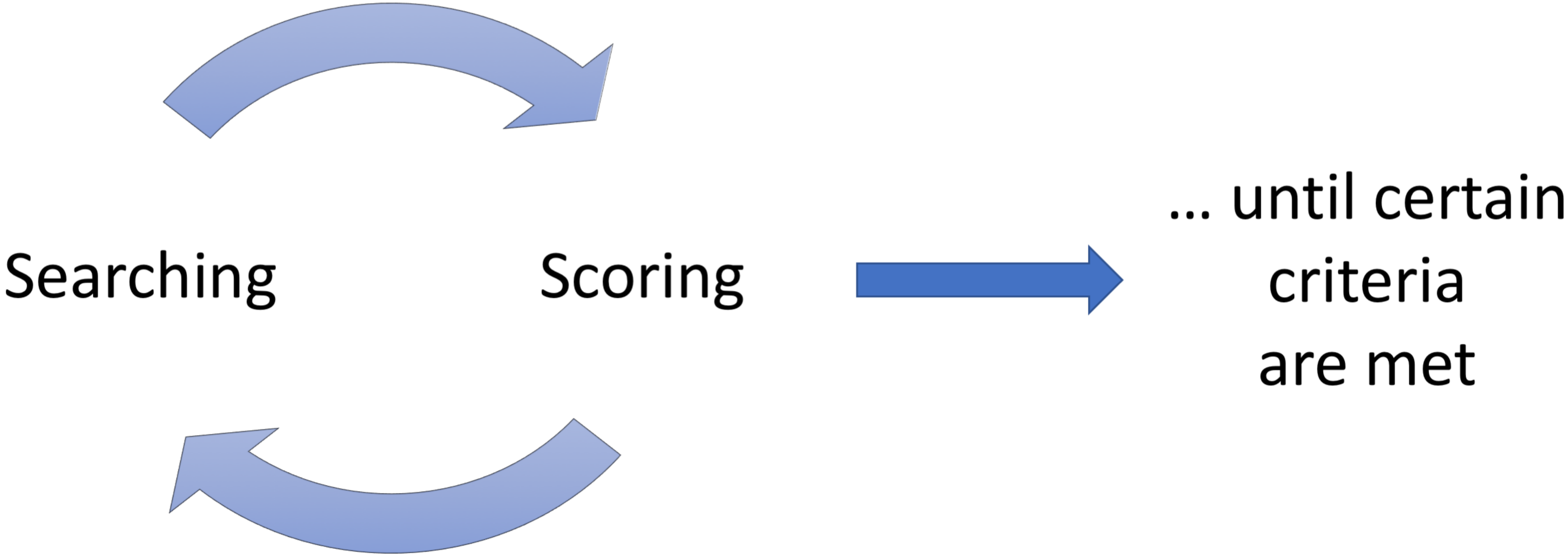
Docking and virtual screening

- What is virtual screening?
- Pharmacophore searching
- Shape-based searching
- Docking
- Estimating model quality

Docking



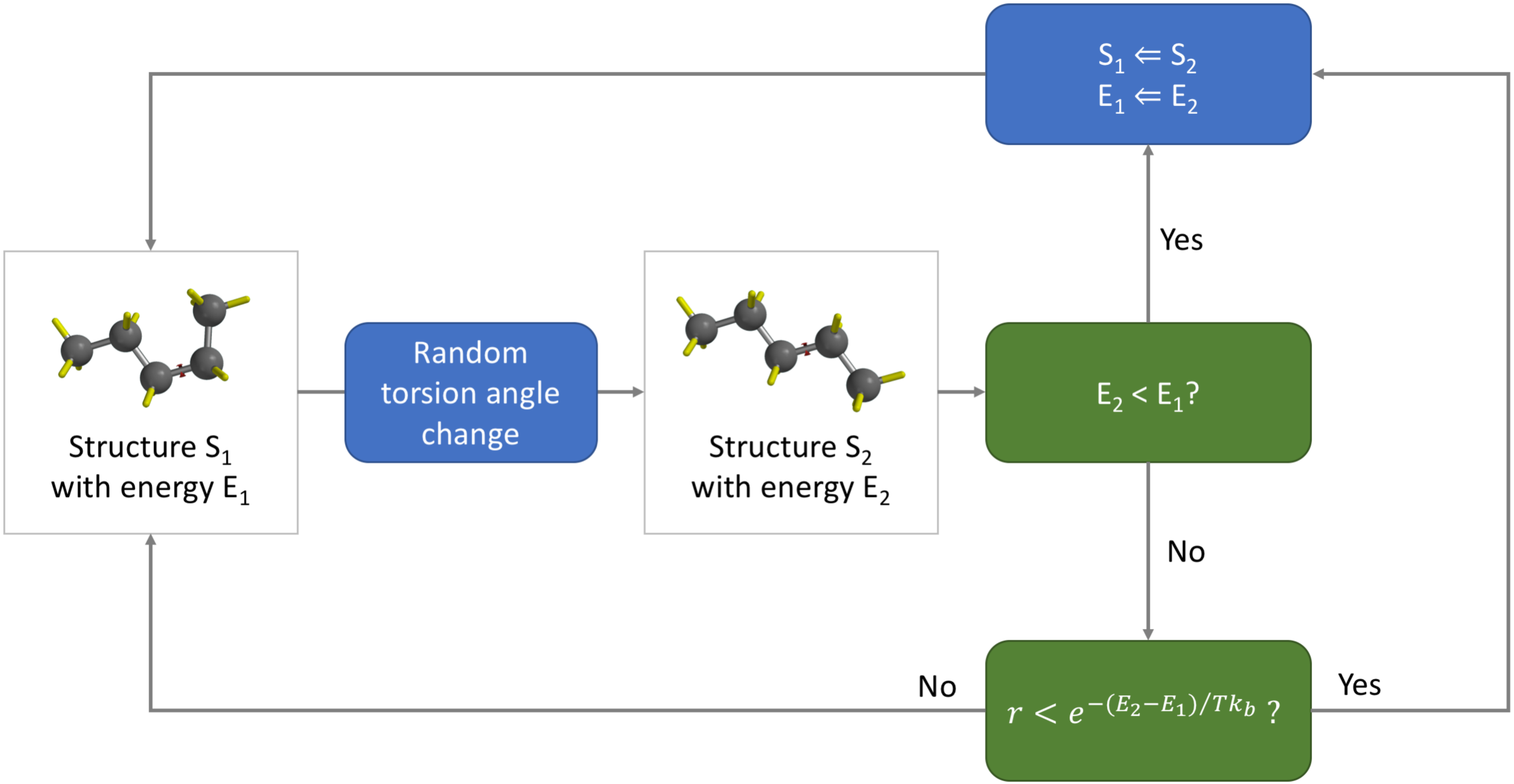
The repeated process of searching and scoring



Searching methods

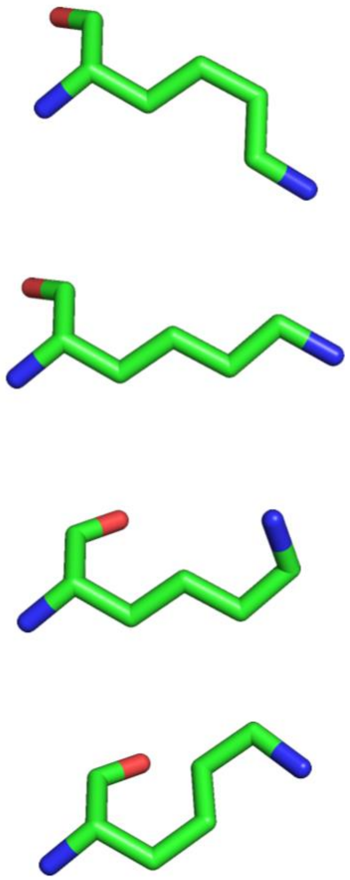
- Molecular dynamics or Monte Carlo simulations
  - $\boldsymbol{F} = m\,\boldsymbol{a}$
- Genetic algorithms
  - Gold
  - Autodock
- Shape-based methods
  - DOCK
  - FRED
  - Glide (Schrödinger)
  - SURFLEX

Monte Carlo searching

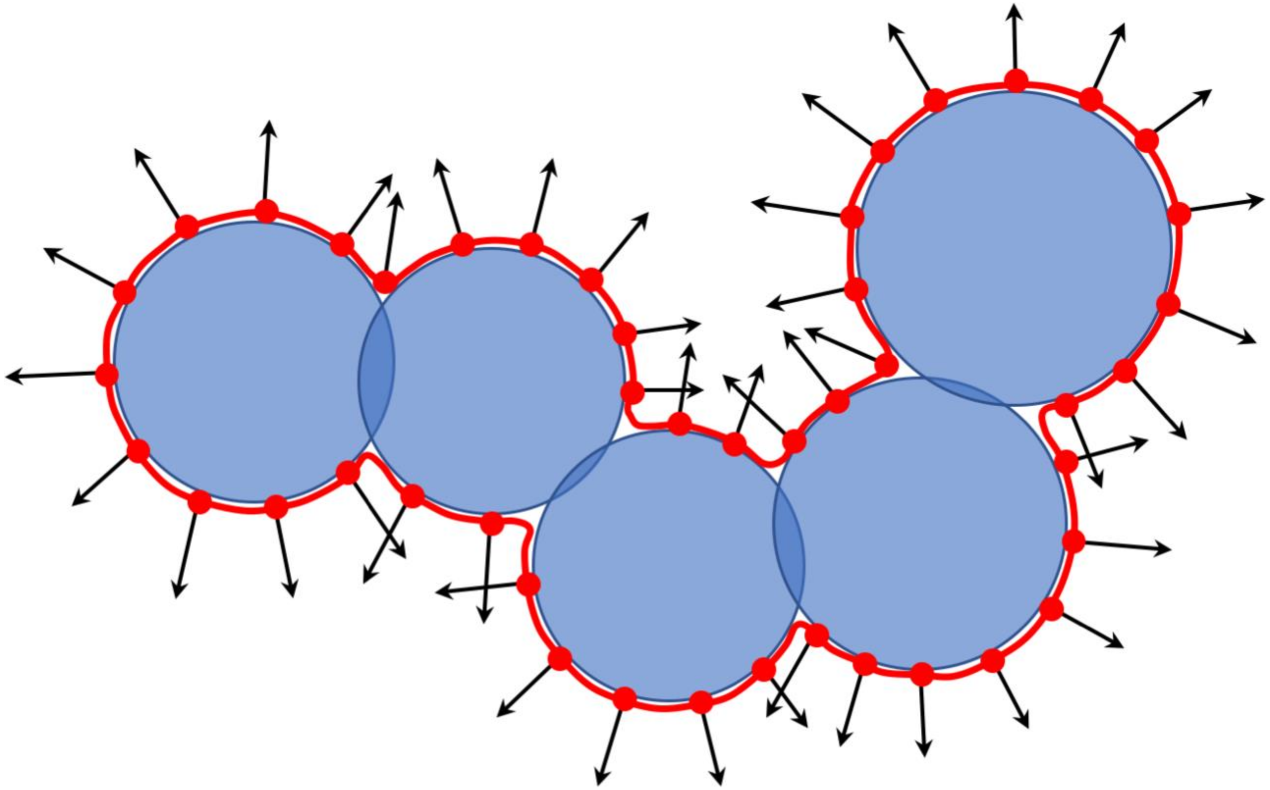


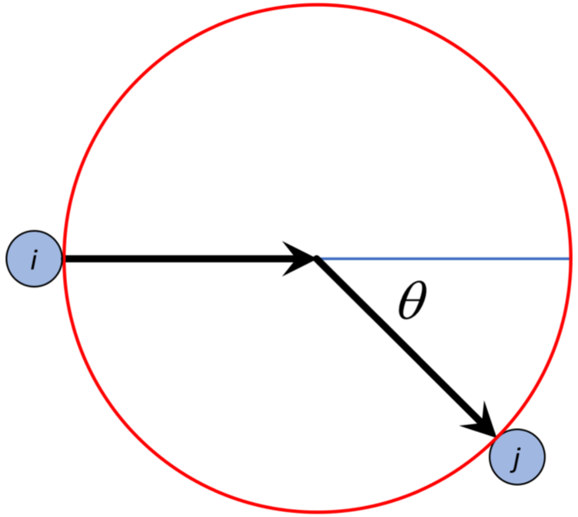
Genetic algorithms

	T1	T2	T3	T4	T5
C1	151	50	180	187	96
C2	150	181	189	180	14
C3	90	273	70	180	88
C4	94	0	50	180	210



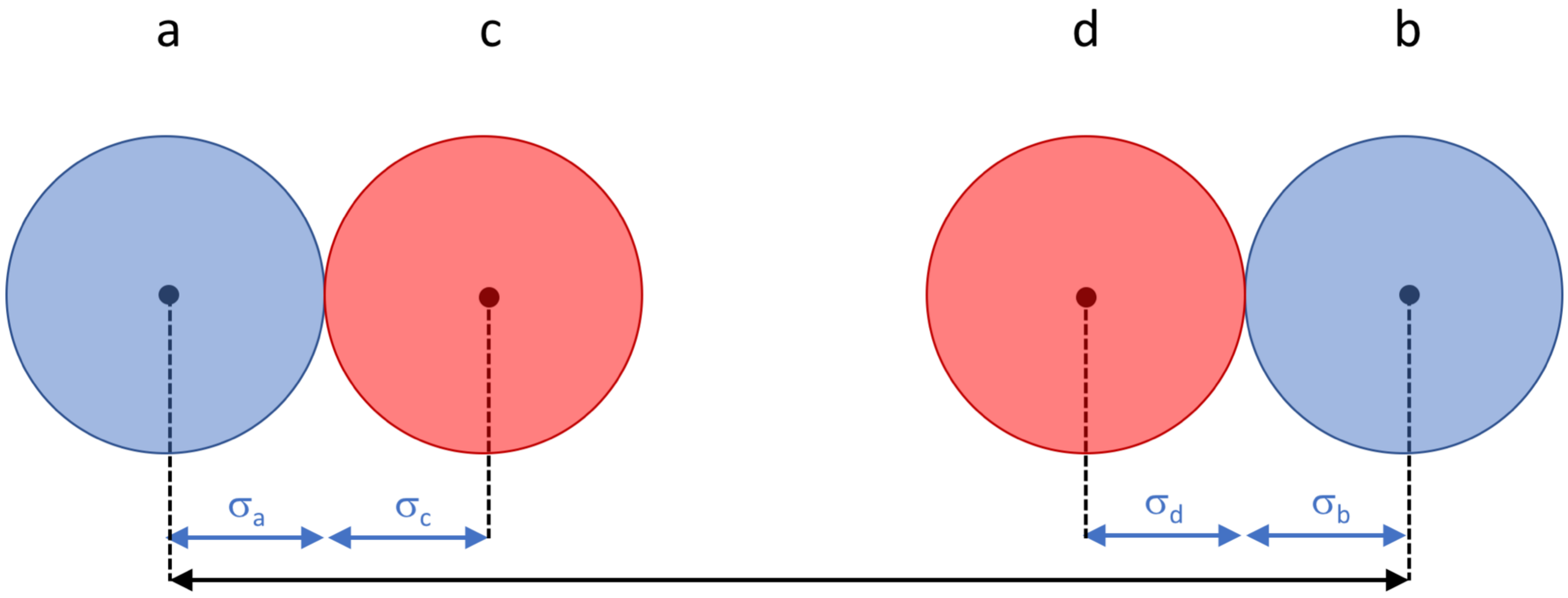
Shape-based searching:
- step 1: representation





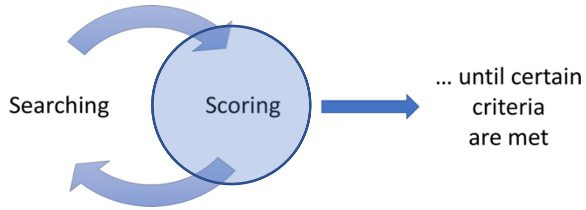
Kuntz et al. (1982) ‘A geometric approach to macromolecule-ligand interactions’, *J. Mol. Biol.* **161**, 269-288.

Shape-based searching:
- step 2: matching

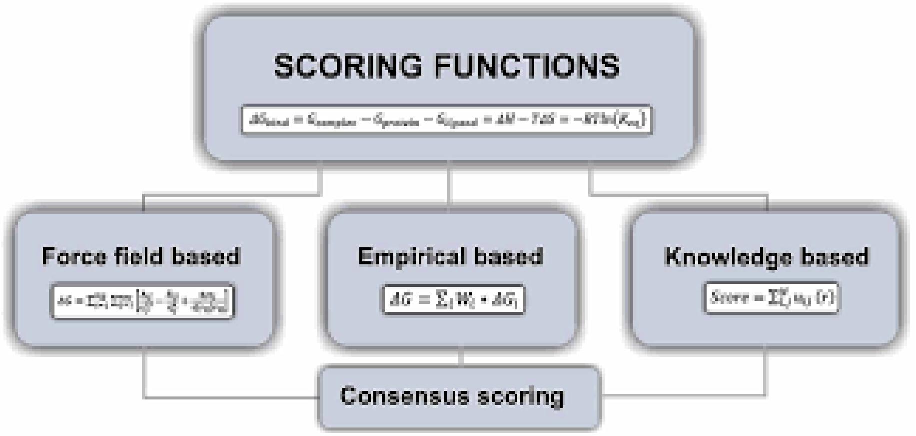


- step 3: optimisation

Scoring methods

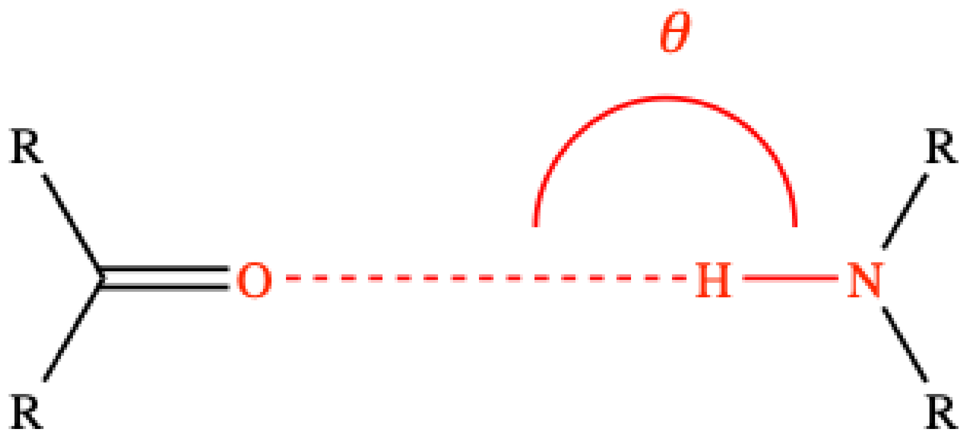


- Force-field based scoring functions
- Empirical scoring function
- Knowledge-based scoring function



Force-field based scoring

$$E = W_{VDW}\sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}}+\frac{B_{ij}}{r_{ij}^{6}}\right)+W_{hbond}\sum_{i,j}p(\theta)\left(\frac{C_{ij}}{r_{ij}^{12}}+\frac{D_{ij}}{r_{ij}^{6}}\right)+W_{elec}\sum_{i,j}\frac{q_i q_j}{r_{ij}}+W_{sol}\sum_{i,j}(S_i V_j+S_j V_i)e^{\left(-r_{ij}^2/2\sigma^2\right)}$$



Empirical scoring functions

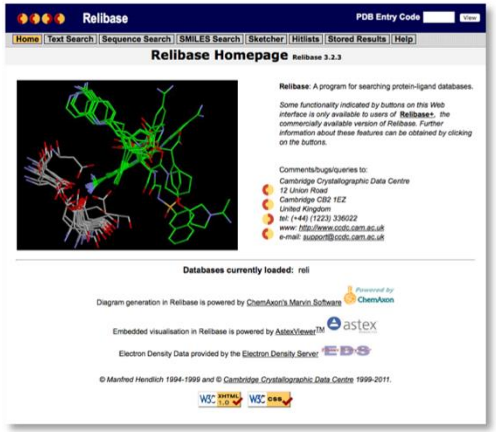
$$\Delta G = f_{hbonds}\Delta G_{hbonds} + f_{polar-apolar}\Delta G_{polar-apolar} + f_{nrot}\Delta G_{nrot} + f_{apolar-apolar}\Delta G_{apolar-apolar}$$

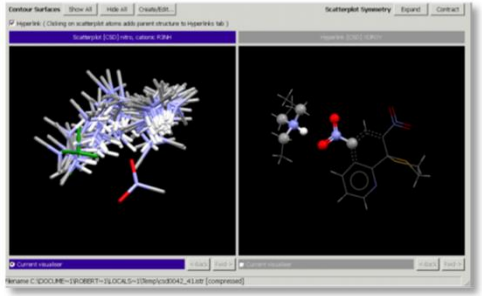
Knowledge-based scoring functions

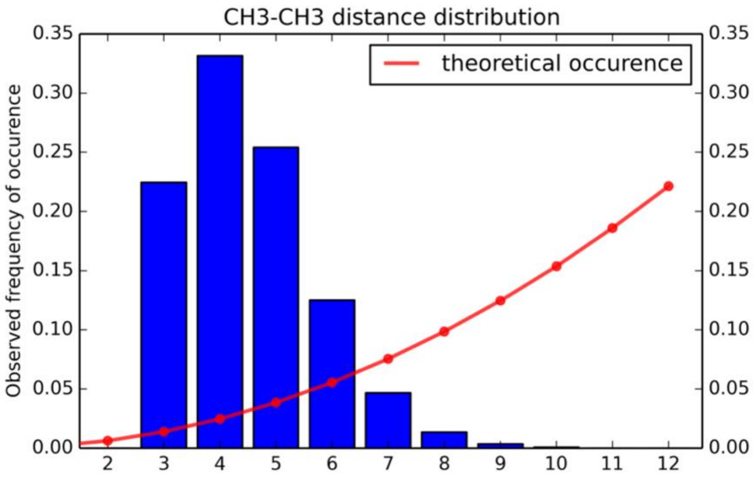
Experimental contact data from **X-ray structures**

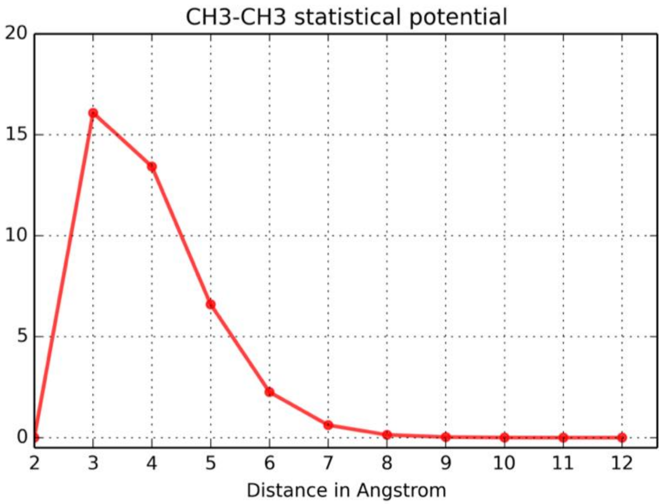
Extract **distance distributions** for each pair of atomtypes

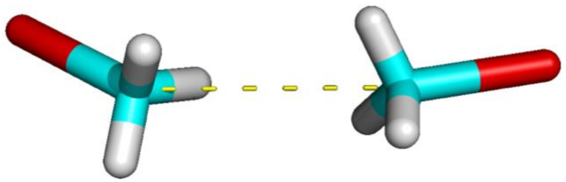
Calculate **statistical potential** for each pair of atomtypes











$$P_{ij} = -\ln \frac{g_{ij}(r)}{g_{ref}}$$

Scope of the different scoring functions

	Pose prediction	**Compound selection**
Forcefield-based	✔	
Empirical		✔
Knowledge-based	✔	✔

Case studies

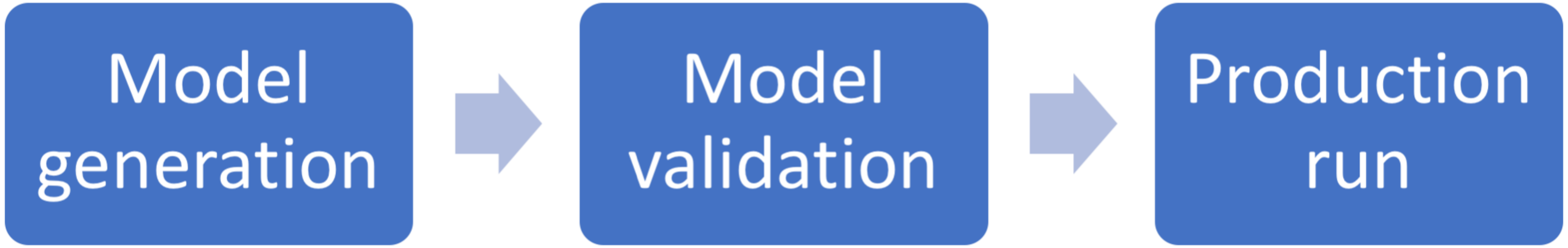
- BACE inhibitors

Docking and virtual screening

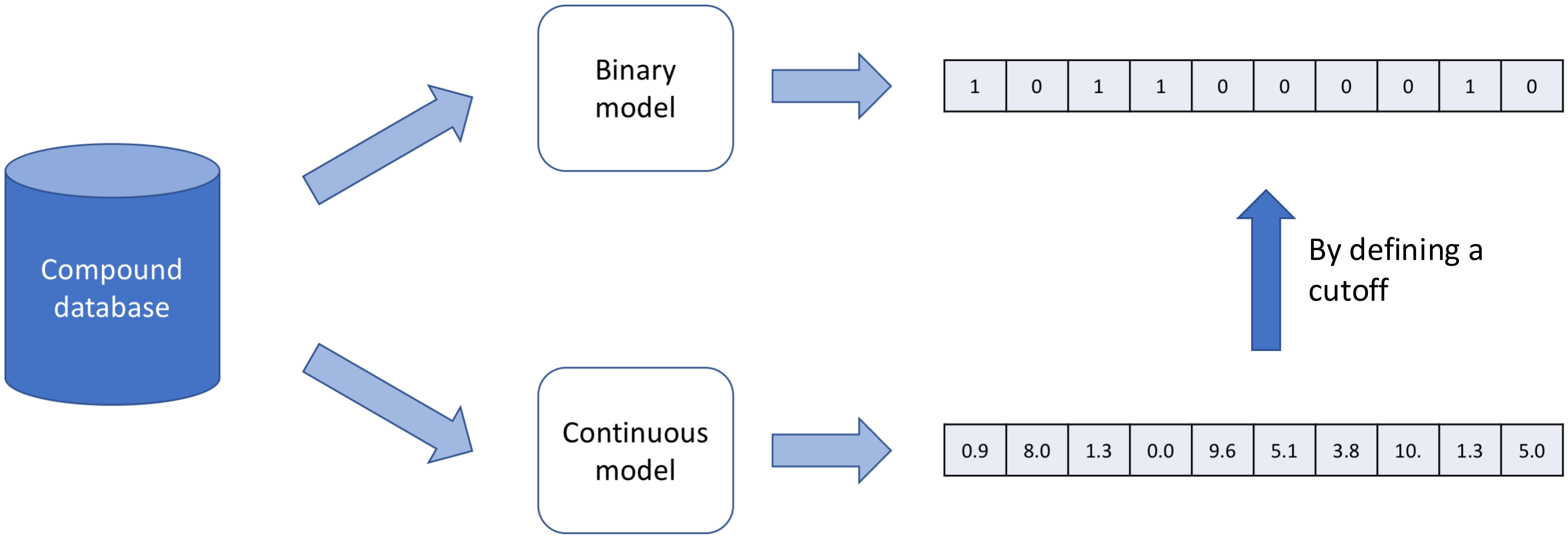
- What is virtual screening?
- Pharmacophore searching
- Shape-based searching
- Docking
- Estimating model quality

Estimating model quality

- Chemoinformatics-based screening
- Pharmacophore-based screening
- Docking



Ranking and classification



From continuous to binary

-3.5	+2.1	-7.5	+0.1	+9.3	-4.3	-3.2	+5.2	+0.5	+2.1

Rank from ‘good’ to ‘bad’

-7.5	-4.3	-3.5	-3.2	+0.1	+0.5	+2.1	+2.1	+5.2	+9.3

Specify cutoff to classify ‘active’ *versus* ‘inactive’

-7.5	-4.3	-3.5	-3.2	+0.1	+0.5	+2.1	+2.1	+5.2	+9.3

cutoff ≤ -4.0

Everything below cutoff is ‘active’

1	1	0	0	0	0	0	0	0	0

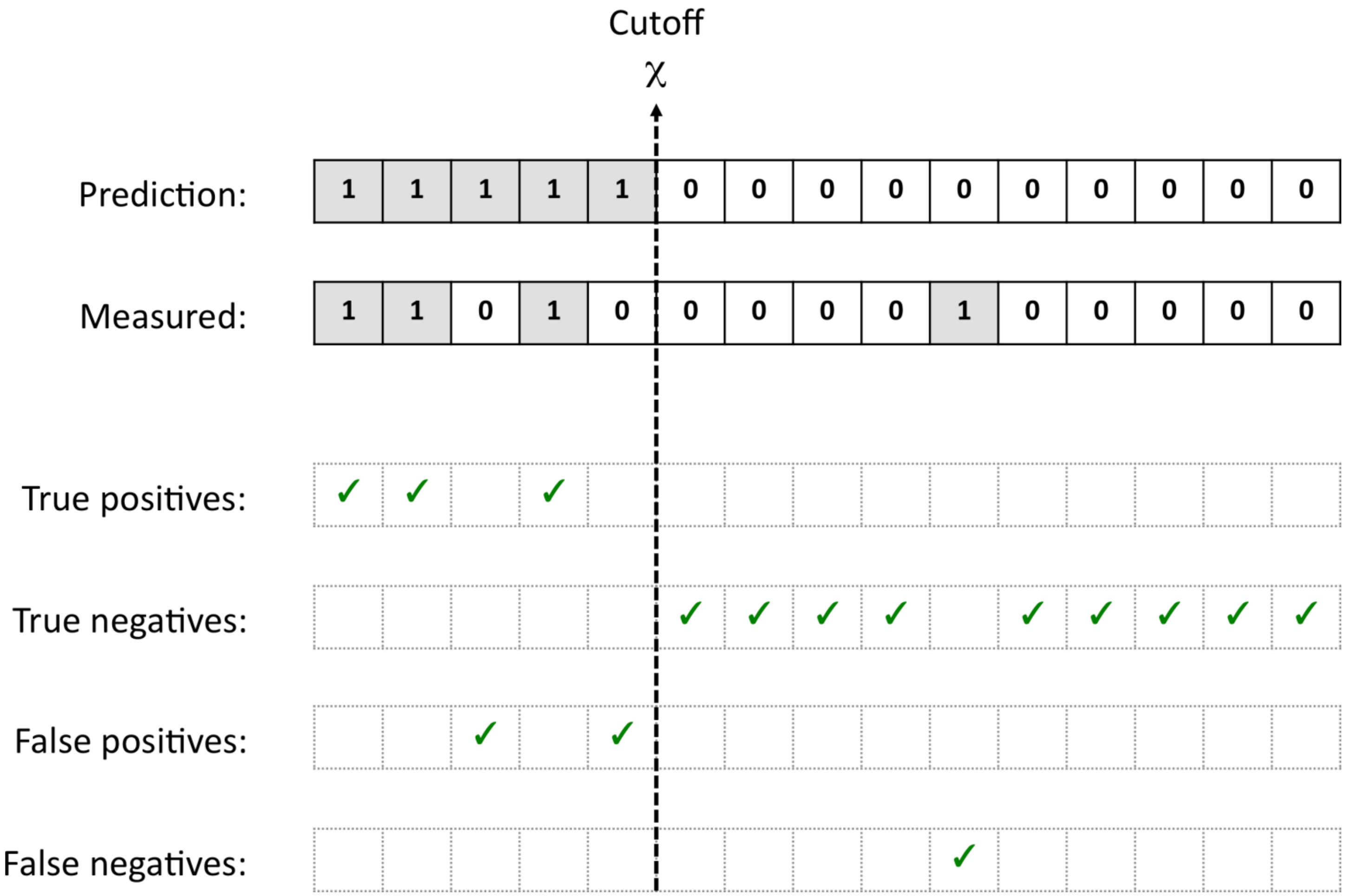
predicted active

predicted inactive

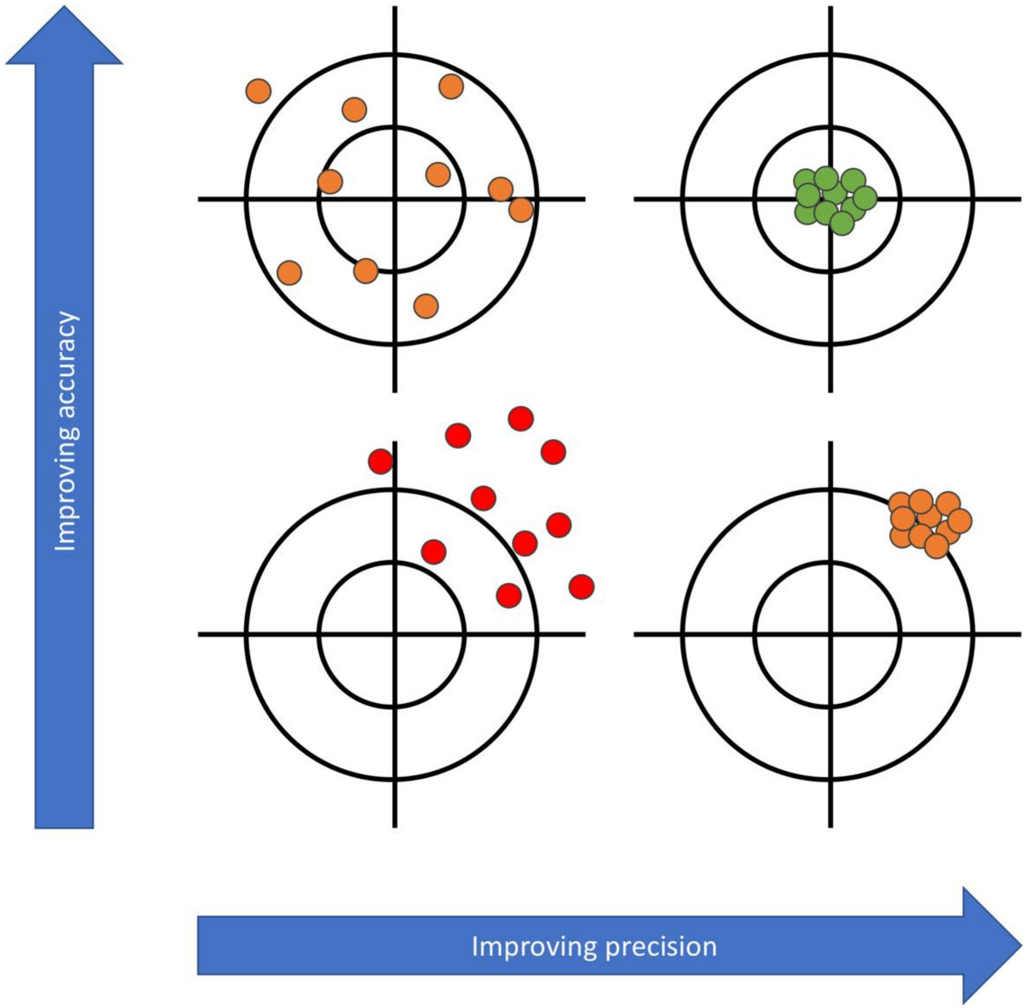
Confusion matrix

		Actual	
		Active (1)	Inactive (0)
Predicted	Active (1)	TP	FP
	Inactive (0)	FN	TN

Confusion matrix and cutoff

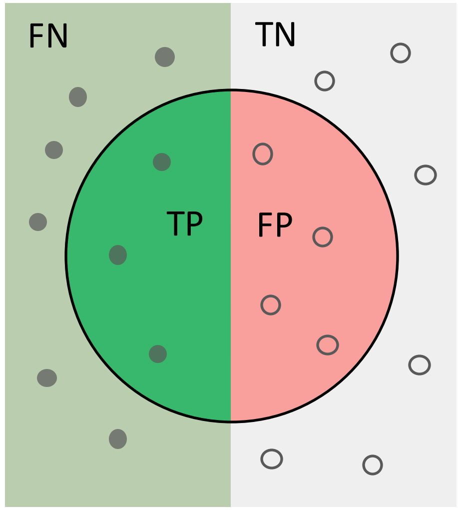


Performance metrics: accuracy & precision



Metrics and the confusion matrix

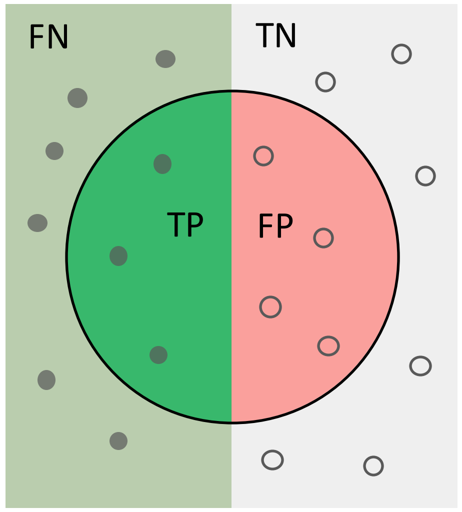
- $ACC = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+FN+TN+FP}$
- $PRE = \frac{TP}{TP+FP}$
- $SEN = \frac{TP}{P} = \frac{TP}{TP+FN}$

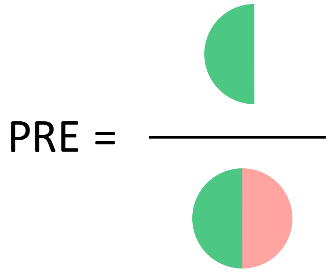


$$\text{Precision} = \frac{TP}{TP+FP}$$

- Useful if you have limited budget and you want to be sure that, if a compound is predicted to be 'active', changes are very likely that the compound is really active.

- A high precision comes at the cost of missing out real actives which are not selected by the method



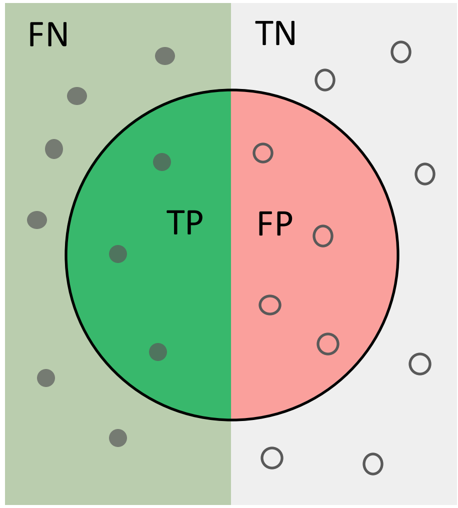


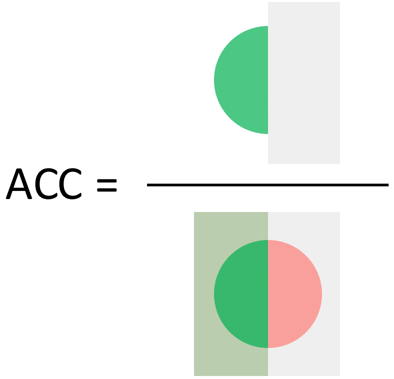
(only looks at the hitlist)

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

- Useful if you have a balanced dataset with balanced number of actives and inactives

- Should *never* be used when there are only a limited number of actives in the dataset.

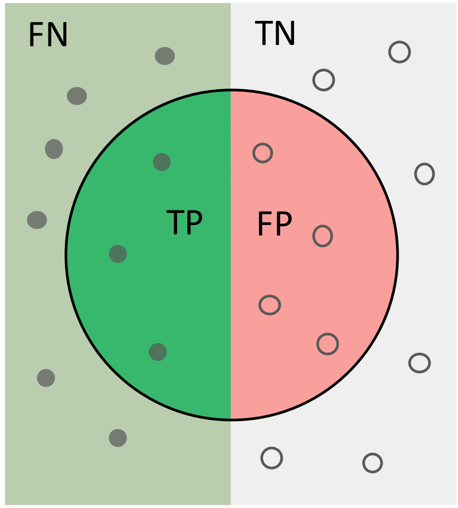




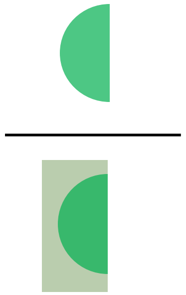
Recall = $\frac{TP}{TP+FN}$ = Sensitivity

- Useful if you want to retrieve as many actives as possible from the database *("you don't want to miss actives")*
- Comes at the risc of retrieving many false positives
- Optimising for recall is only useful if the precision is also taken into account:
    - When you screen the entire database you will always get 100% recall...

F1-score

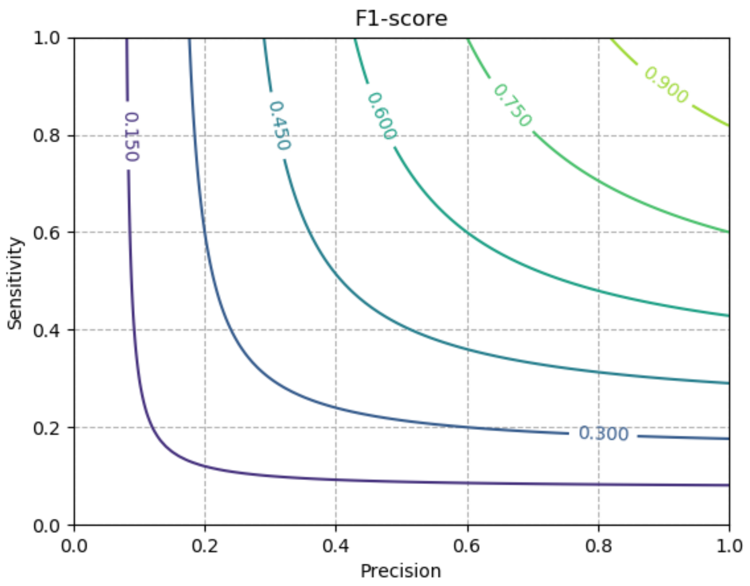


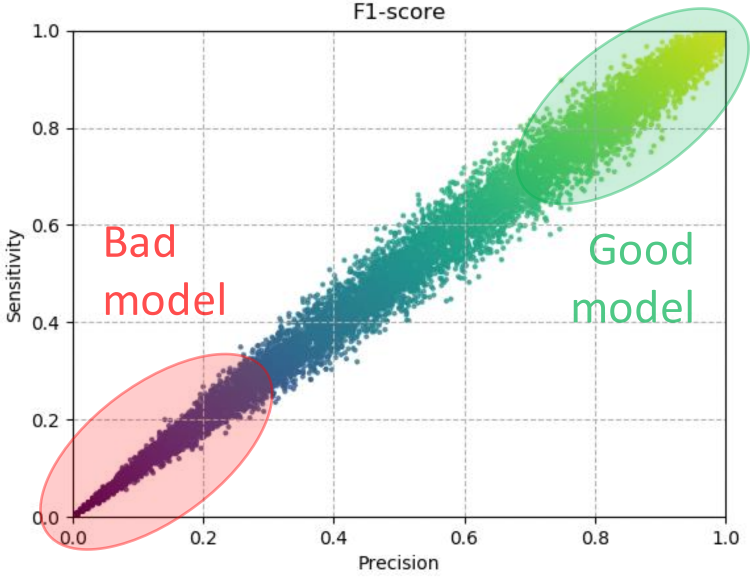
SEN =



$$\text{F1-score} = \frac{2\,TP}{2\,TP + FP + FN}$$

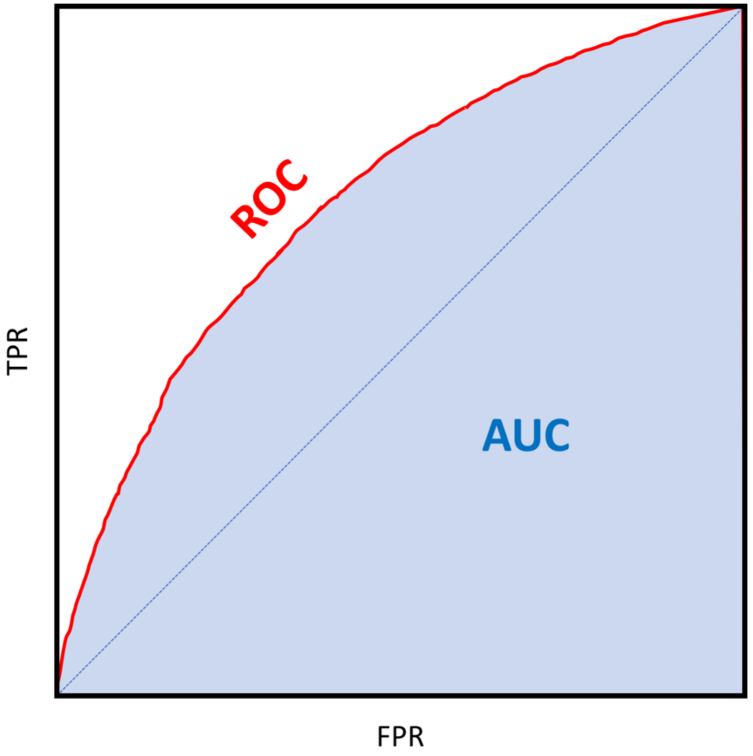
- The harmonic mean of precision and sensitivity:
  - $F1 = \dfrac{2 * PRE * SEN}{PRE + SEN}$
- Represents a good trade-off between identifying all actives versus a good likelihood of being truly active





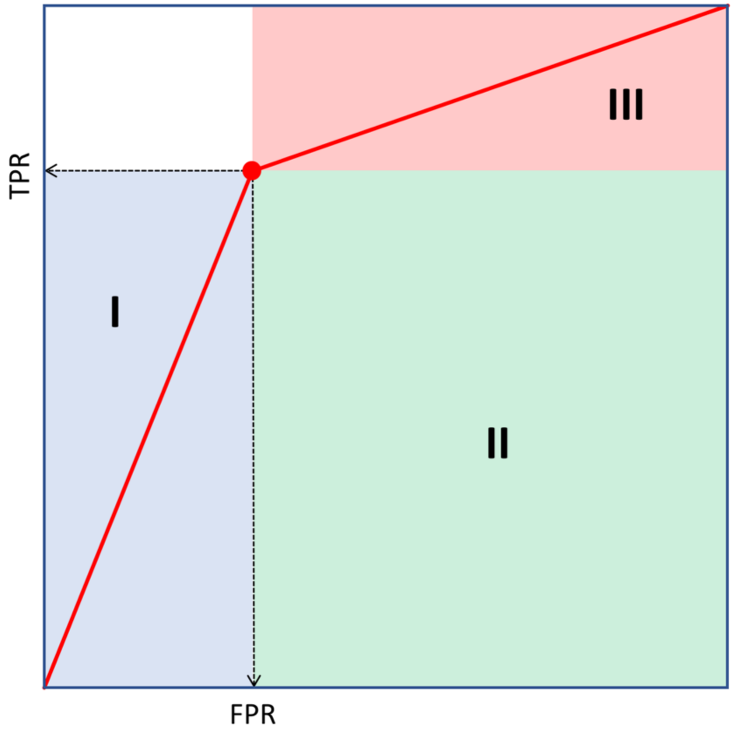
AUC-ROC curve

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$



$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN}$$

$$AUC = \frac{TPR - FPR + 1}{2}$$



Metrics, model quality and cutoff

- The confusion matrix metrics are influenced by:
  - The quality of the *model*
  - Selection of the *cutoff* in case of a continuous model

- The quality of the model is influenced by:
  - The model itself:
    - Machine learning algorithm and parameters
    - Docking method and parameters
    - Pharmacophore selection and method
  - The quality of the training data

Good and bad models *versus* cutoff

Bad model:

1	0	0	0	1	0	1	1	1	0

Good model:

1	1	1	0	1	0	0	0	1	0

10% 40% 70%

Metrics:

Bad model:

	TP	TN	FP	FN	ACC	PRE
10% cutoff:						
50% cutoff:						
70% cutoff:						

Good model:

	TP	TN	FP	FN	ACC	PRE
10% cutoff:						
50% cutoff:						
70% cutoff:						

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + FN + TN + FP}$$

$$PRE = \frac{TP}{TP + FP}$$

Model validation: cross-fold approach

Step 1: Divide the dataset into *k* folds, here *k* is 10

Step 2: Use one fold for validating the model that has been built on all other folds

Step 3: Repeat the model building and validation for each of the data folds (10 times)

Step 4: Calculate the avarege of all of the *k* validation performance values