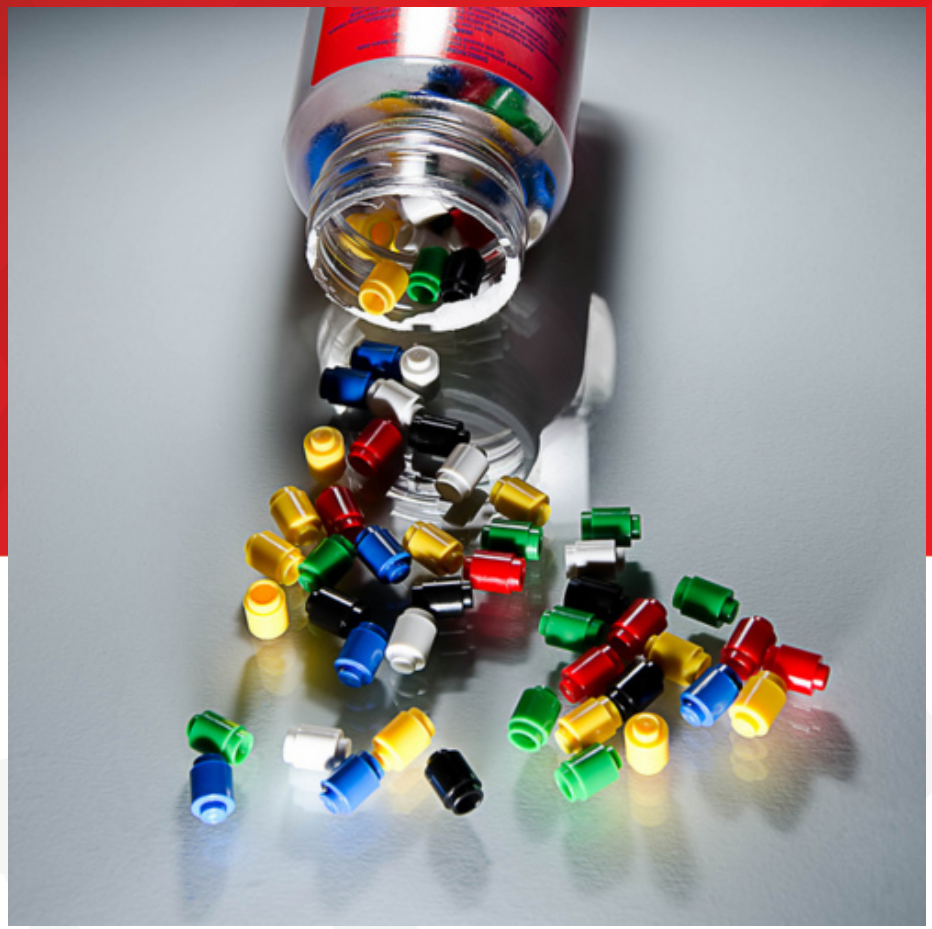
Design drugs with a computer: Background, concepts and teaching material.

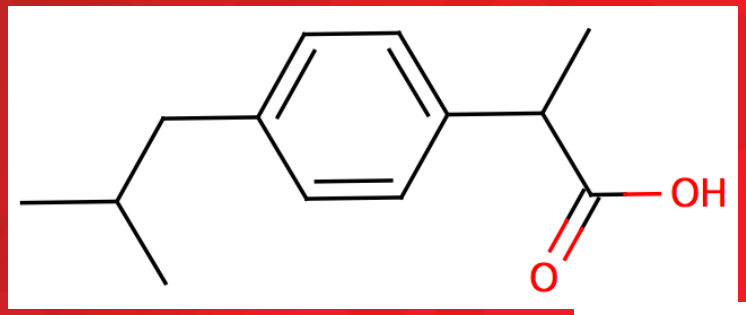
Antoine Daina & Vincent Zoete
Molecular Modeling Group – Prof. O. Michielin
SIB | Swiss Institute of Bioinformatics | Lausanne

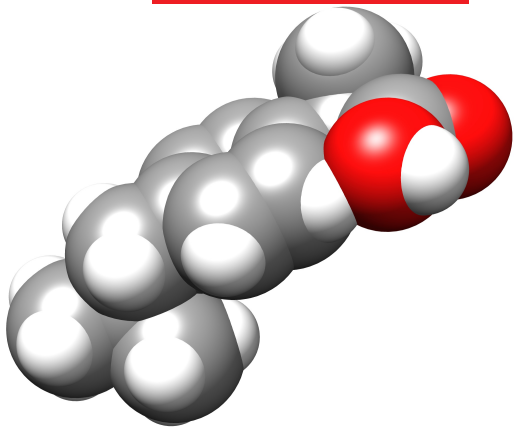


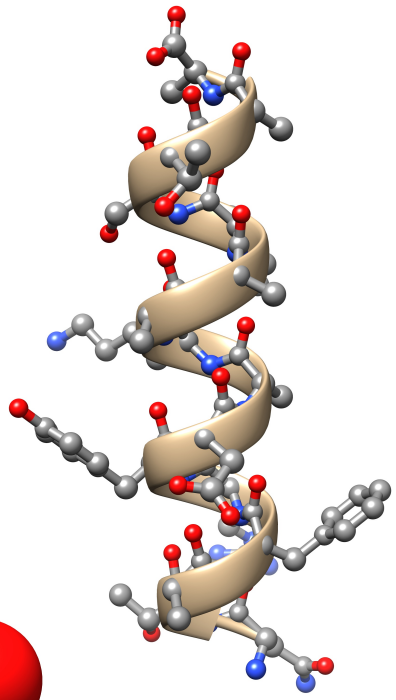
January 22, 2016.

SIB
Swiss Institute of Bioinformatics

Prologue: molecular representations







SIB
Swiss Institute of Bioinformatics

Molecular representations – “small” molecules

Organic molecules of less than ~ 100 atoms are often referred to as “small” molecules, as opposed to biological macromolecules (i.e. proteins, DNA, etc.)

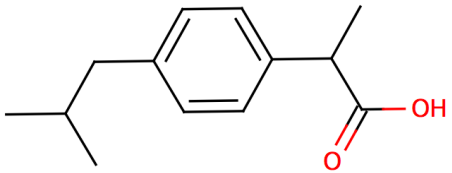
Small molecules can be represented in 1D, 2D or 3D:

Example of Ibuprofen

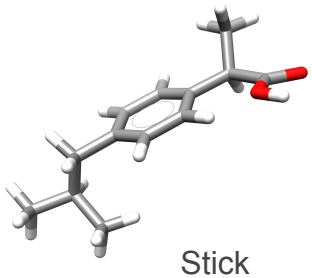
1D

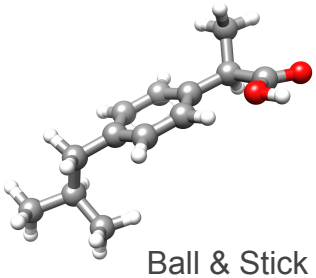
SMILES: CC(C)CC1=CC=C(C=C1)C(C)C(O)=O
InChI: 1S/C13H18O2/c1-9(2)8-11-4-6-12(7-5-11)10(3)13(14)15/h4-7,9-10H,8H2,1-3H3,(H,14,15)

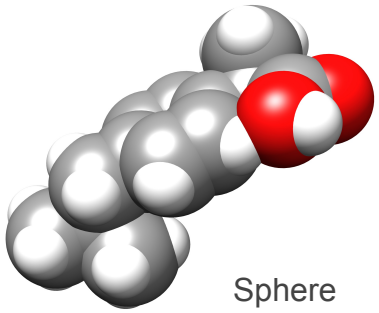
2D

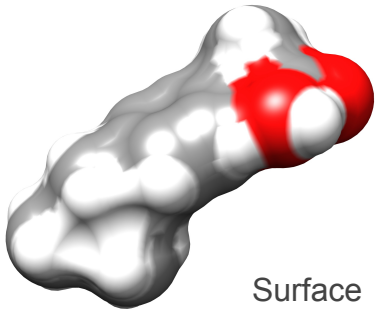


3D


Stick


Ball & Stick


Sphere


Surface

Molecular representations – biological macromolecules

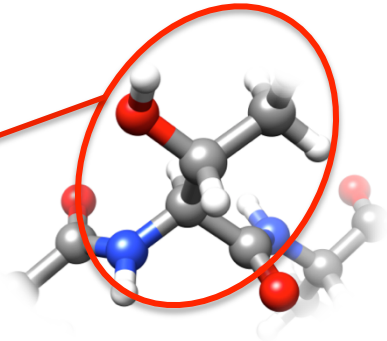
Biological macromolecules can also be represented in 1D, 2D or 3D:

Example of proteins

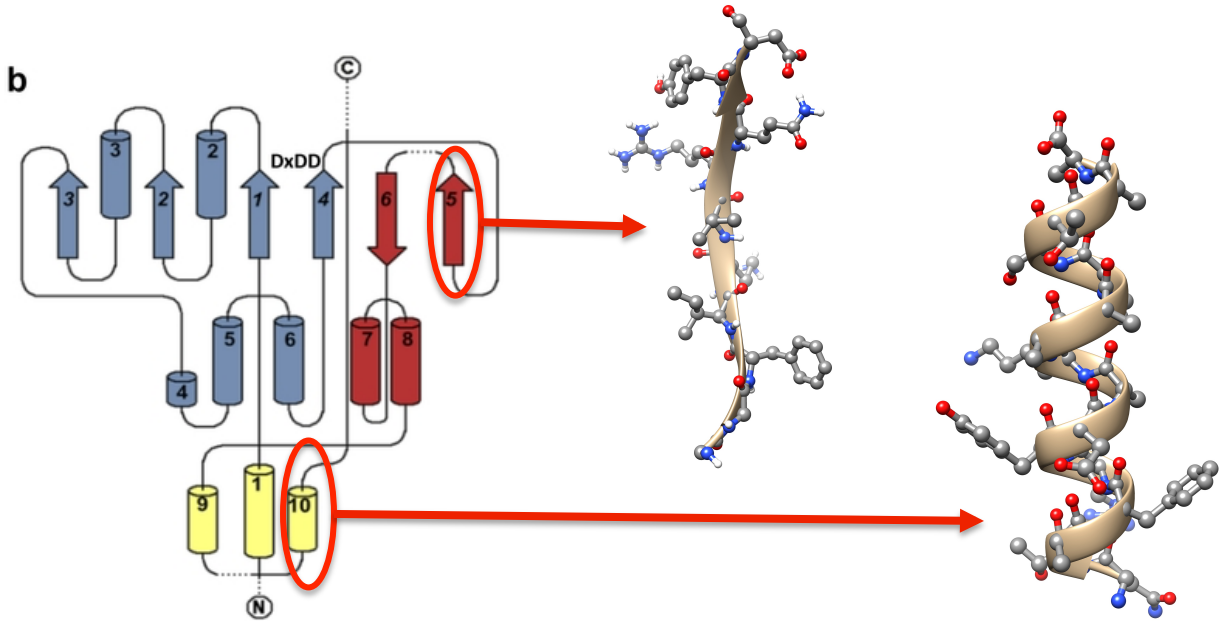
1D

Sequence if amino-acids:

```
MGCTLSAEDKAAVERSKMIDRNLREDGEKAAREVKLLLLGAGESGKSTIVKQMKIIHEAG
YSEEECKQYKAVVYSNTIQSIIAIIRAMGRLKIDFGDSARADDARQLFVLAGAAEEGFMT
AELAGVIKRLWKDSGVQACFNRSREYQLNDSAAYYLNDLDRIAQPNYIPTQQDVLRTRVK
TTGIVETHFTFKDLHFKMFDVGGQRSERKKWIHCFEGVTAIIFCVALSDYDLVLAEDEEM
NRMHESMKLFDSICNNKWFTDTSIILFLNKKDLFEEKIKKSPLTICYPEYAGSNTYEEAA
AYIQCQFEDLNKRKDTKEIYTHFTCATDTKNVQFVFDAVTDVIIKNNLKDCGLF
```



2D

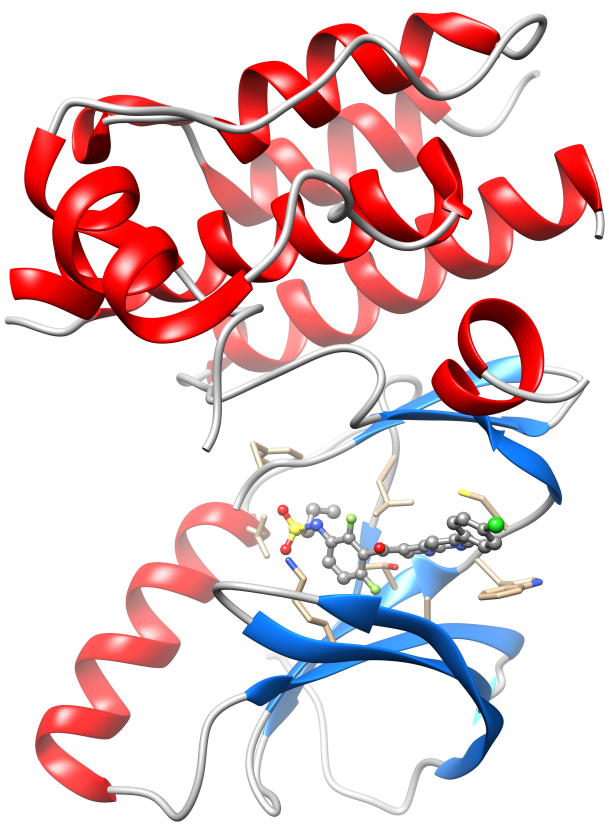


Molecular representations – biological macromolecules

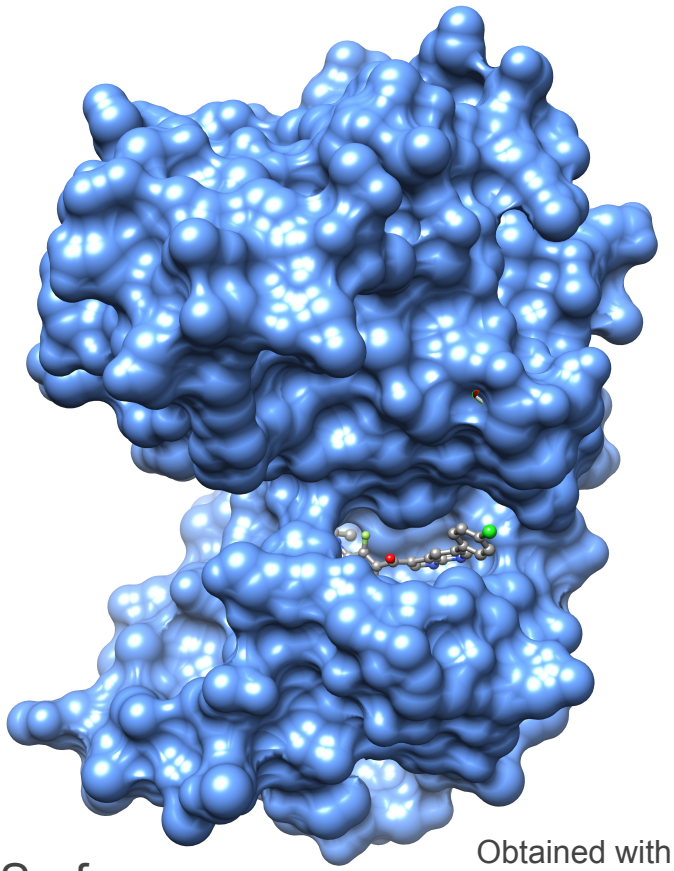
Biological macromolecules can also be represented in 1D, 2D or 3D:

Example of proteins

3D



Ribbon / cartoon



Surface

Obtained with UCSF Chimera

Molecular representations – biological macromolecules

Biological macromolecules can also be represented in 1D, 2D or 3D:

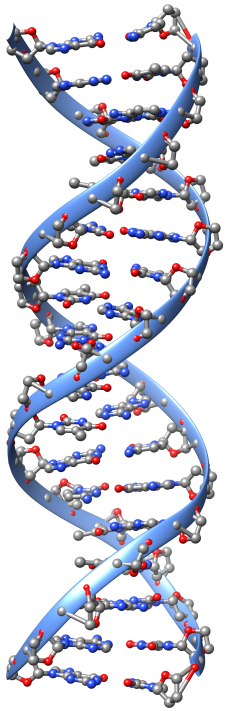
Example of DNA

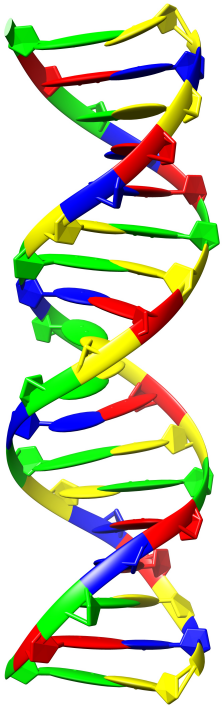
1D

Sequence if nucleotides:

GAGTAGCACGTGCTACTC

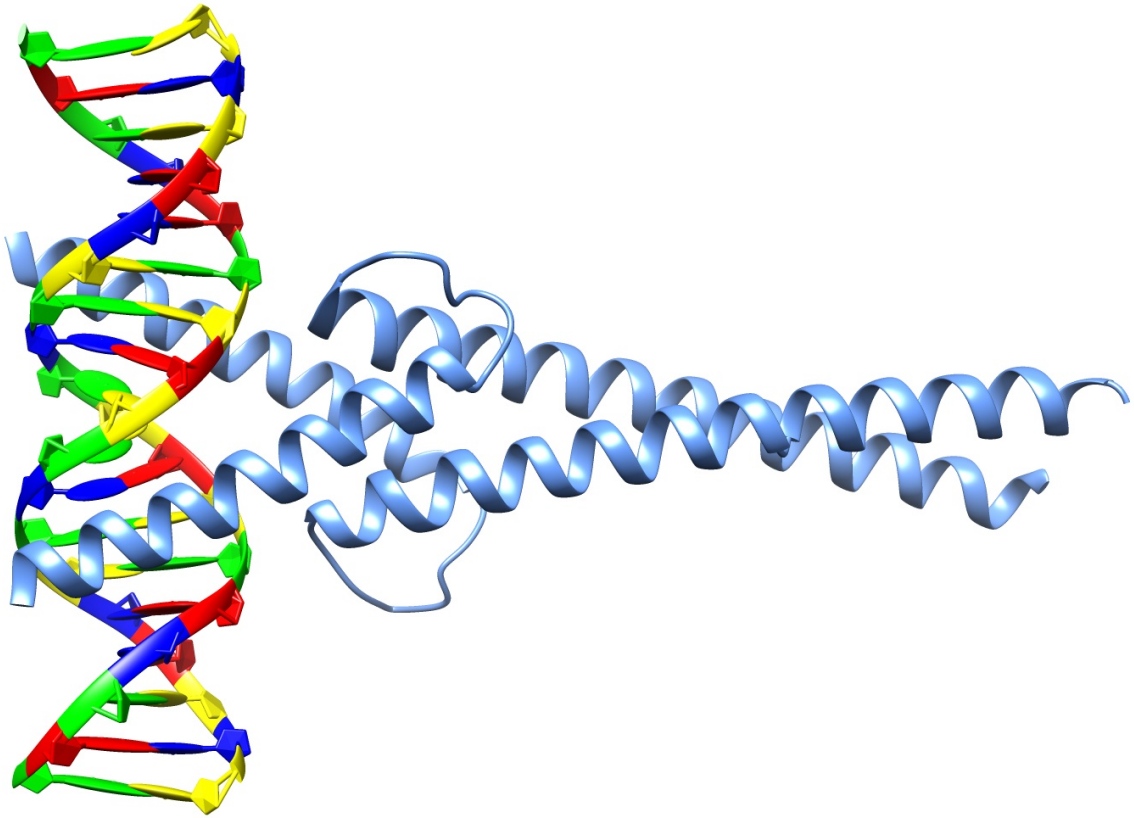
3D





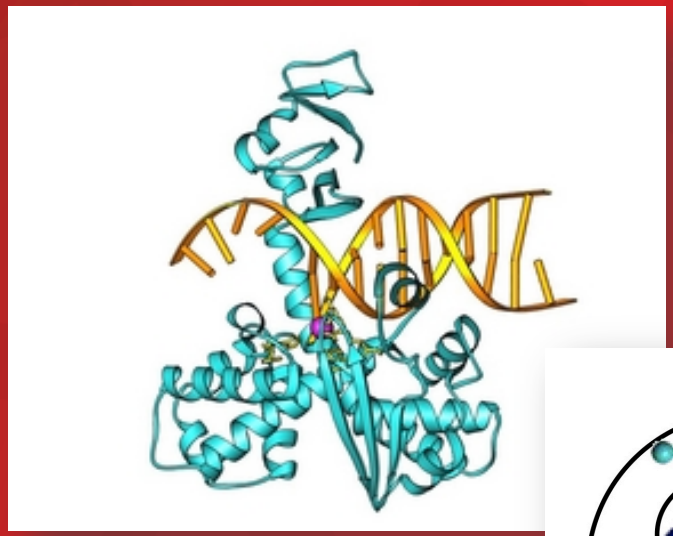
Molecular representations – biological macromolecules

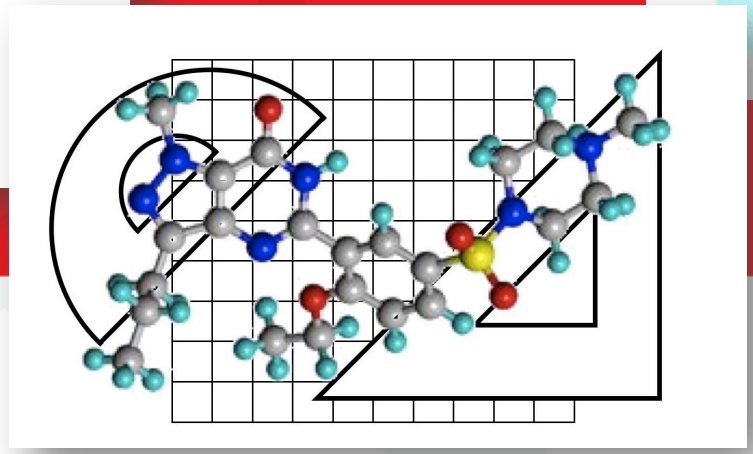
Biological macromolecules can also be represented in 1D, 2D or 3D:



Myc-Max transcription factor

Introduction: Overview of the Drug Design Pipeline





Swiss Institute of Bioinformatics

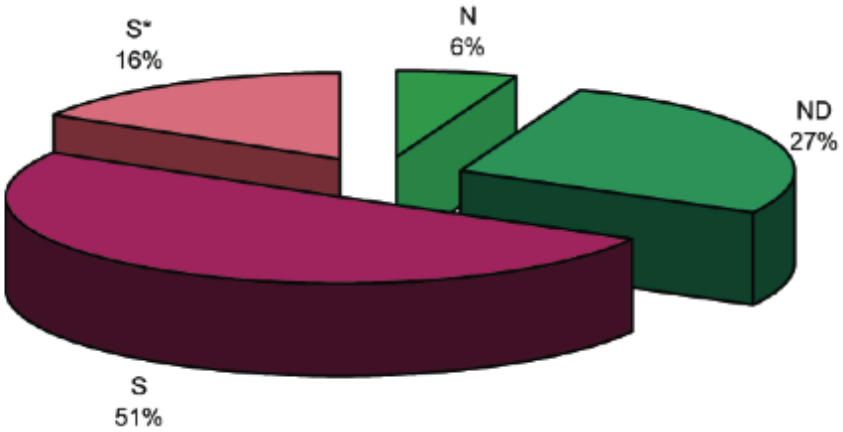
What is a drug ?

Drug:

- An **substance** administered to a patient with possibly various objective:
  - ➔ a **therapeutic** objective (treatment): to cure a **disease**, or
  - a **prophylactic** objective (prevention): to avert the emergence of a **disease**, or
  - a **diagnostic** objective: to identify and monitor a **disease**.

- In the context of **Drug Design**, the substance is a chemical **"small" molecule**.

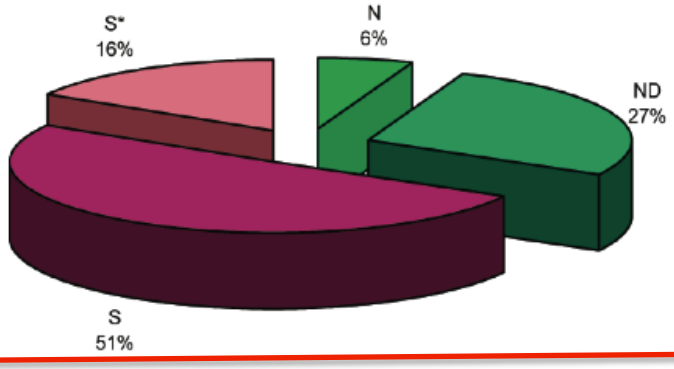
- Where do these drug molecules come from ?



synthetic drug (S 51%)
natural product (N 6%)
derived from N (ND 27%)
inspired by N (S* 16%)

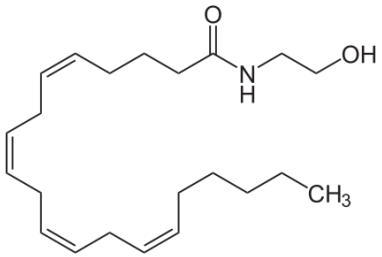
What is a drug ?

- Where do these drug molecules come from ?

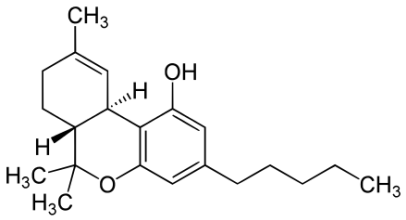


- **synthetic drug (S 51%)**
- **natural product (N 6%)**
- **derived from N (ND 27%)**
- **inspired by N (S* 16%)**

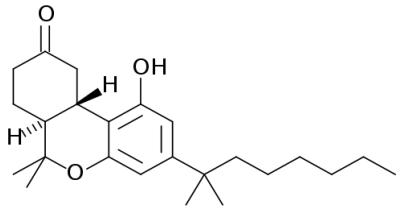
Anandamide. Natural, endogenous, ligand of cannabinoid receptors



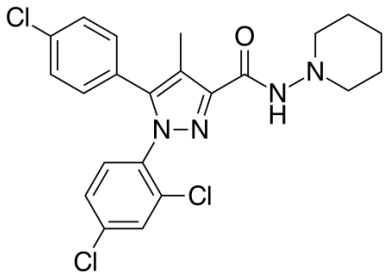
Tetrahydrocannabinol (THC). Natural ligand of cannabinoid receptors, from plant. Analgesic, antiemetic



Nabilone. Synthetic ligand, derived from THC. Analgesic, antiemetic

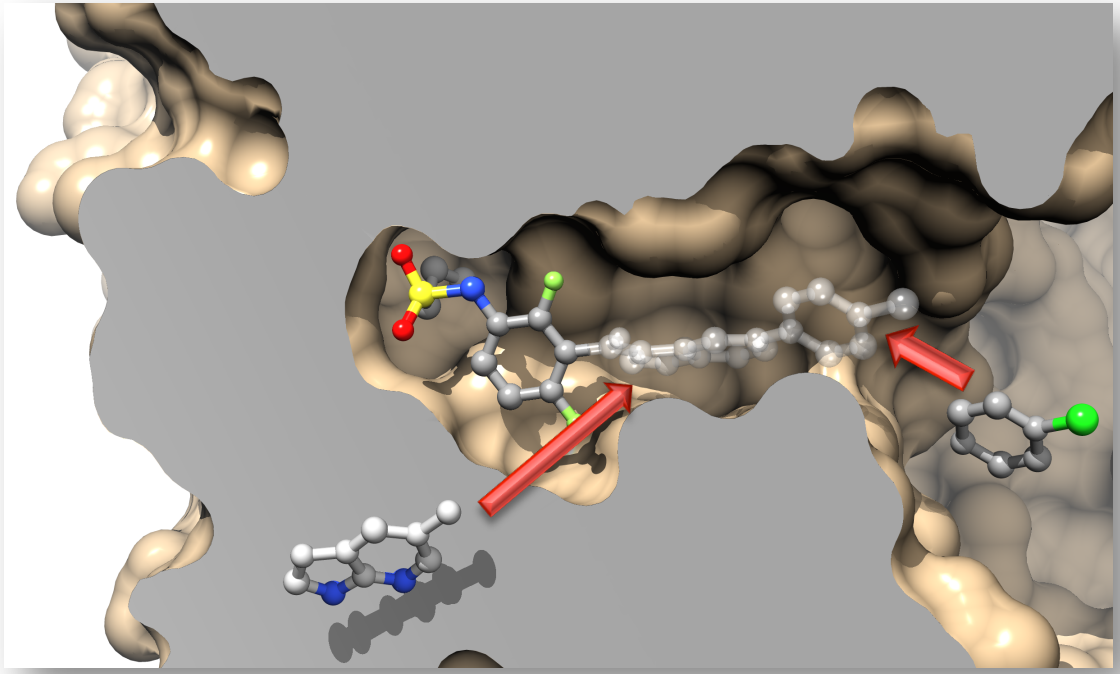


Rimonabant. Synthetic ligand. Anorectic anti-obesity.



Drug design: Aims

Goal: Create new chemical entities (NCE, “molecules”) that activates or inhibits the function of a therapeutically relevant biomolecule (e.g. enzymes, receptor, mainly protein).



To address:

Molecular recognition; i.e. “Lock and key” (E. Fischer)

→ Potency, Selectivity

But also **ADMET**,

- Absorption
- Distribution
- Metabolism
- Excretion
- Toxicity

Drug design: pipeline

Goal: Create new chemical entities (NCE) that activates or inhibits the function of a therapeutically relevant biomolecule (e.g. enzymes, receptor)

- Hit: **molecule** showing a **signal of activity** for the target.
- Hit finding: process to discover hits, generally **using Molecular Screening** (HTS).
- Hit-to-lead: process to select hits for further steps. **Activity confirmation**, re-testing for dose-response, functional assays, **filters** e.g. drug-likeness, toxicity, promiscuous hit, synthetic accessibility.
- Lead: molecule showing **promising and confirmed properties** ➔ can be used as a starting point for optimization.
- Lead optimization: Modest and targeted **chemical modifications** of the lead **to refine** (and possibly understand) pharmacodynamic (PD) and pharmacokinetic (PK) behaviors.
- Preclinical development: **animal** pharmacology/toxicology **testing** ➔ define the **indication** and decide whether it is reasonably **safe to proceed with human clinical trials**.
- Clinical development: **human trials** to evaluate the **safety**, the **dosage** range, the **efficacy** and the **side-effects** of the drug

Drug design: some figures

Globally:

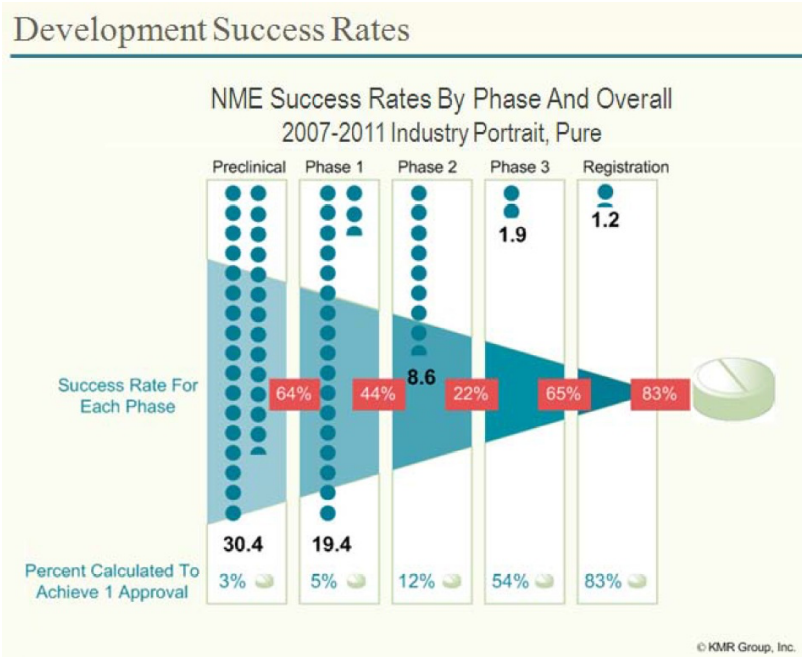
- ~ 40 new active ingredients approved for the market by the FDA each year,
- including 10 'first in class', i.e. drugs with new mode of action.

For each project:

- Thousands of molecules synthesized and tested experimentally during drug discovery ('hit finding' to 'lead optimization')
- Millions of chemical structures ("virtual molecules") created and evaluated *in silico.*
- ~ 3 to 10 molecules tested in preclinical trials.
- 1 to 3 molecules to enter in clinical trials.

Outcome, duration and costs:

- 3 to 10% of the molecules entering preclinical trials will become drugs
- 5 to 17% of the molecules entering clinical trials will become drugs
- 8 - 12 years in total, including 6 - 7 years of clinical trials
- Total cost: ~ 1 billion dollars for a complete project

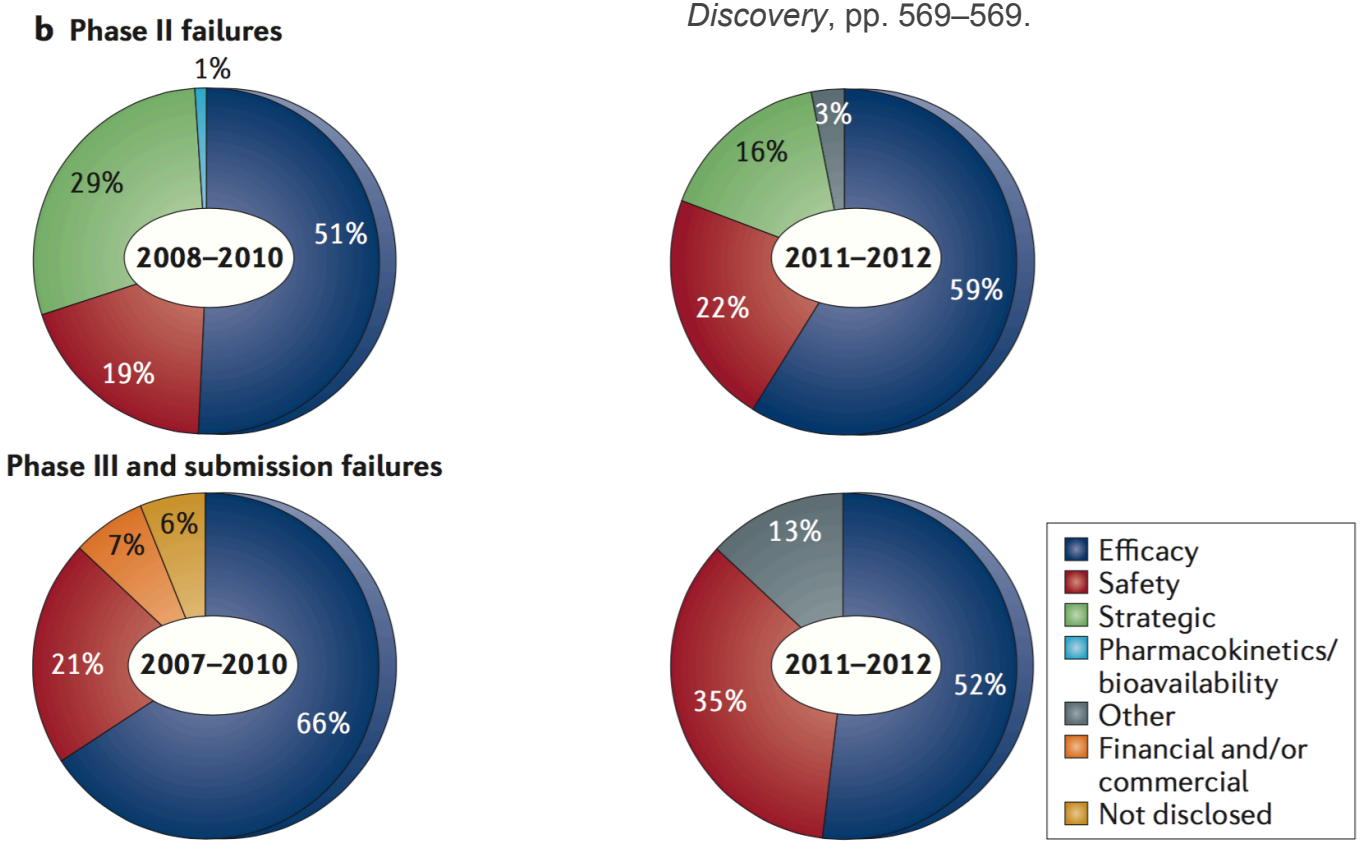


SIB Swiss Institute of Bioinformatics

Drug design: some figures

Reasons for failure in clinical trials:

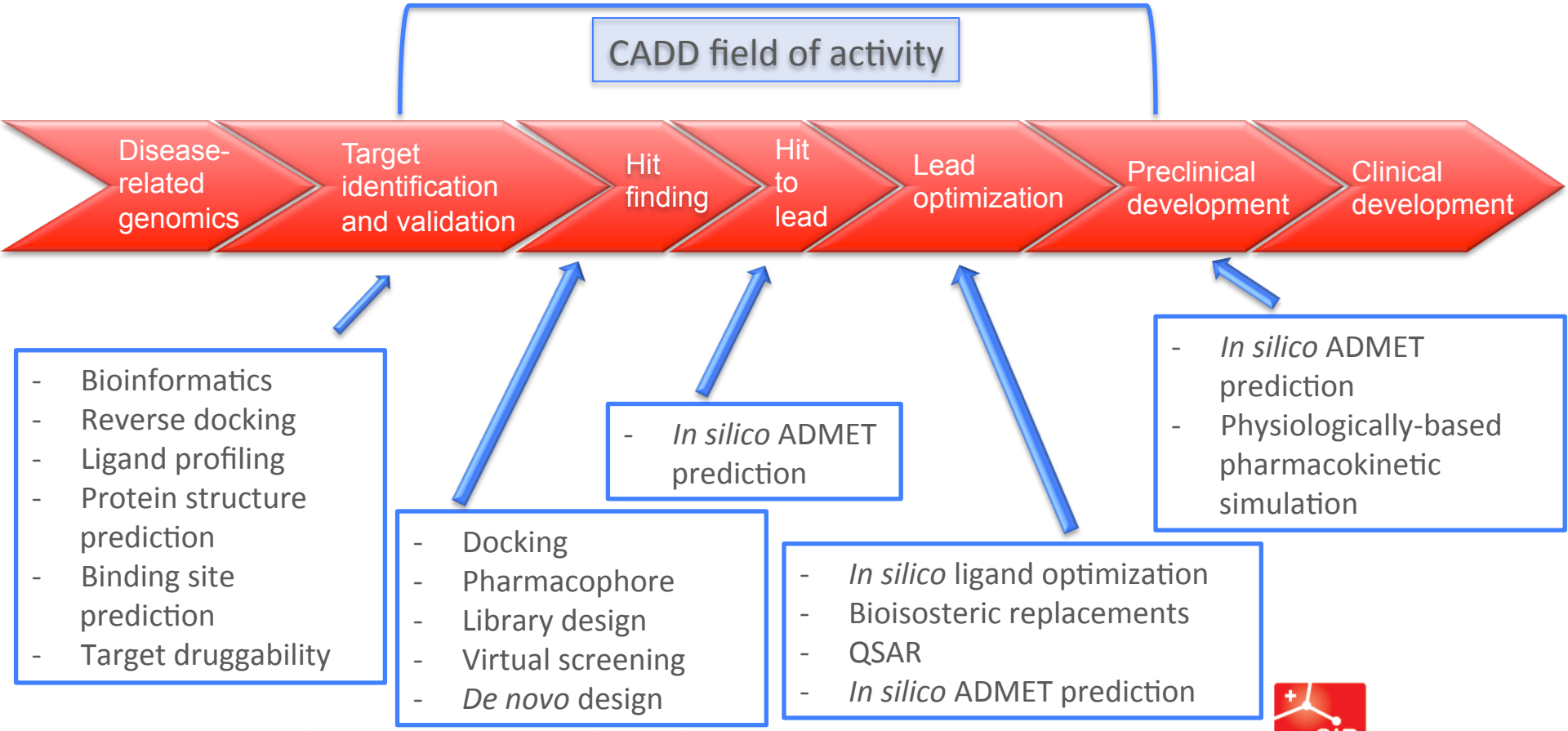
Arrowsmith, J., & Miller, P. (2013, August). Trial watch: phase II and phase III attrition rates 2011-2012. *Nature Reviews. Drug Discovery*, pp. 569–569.



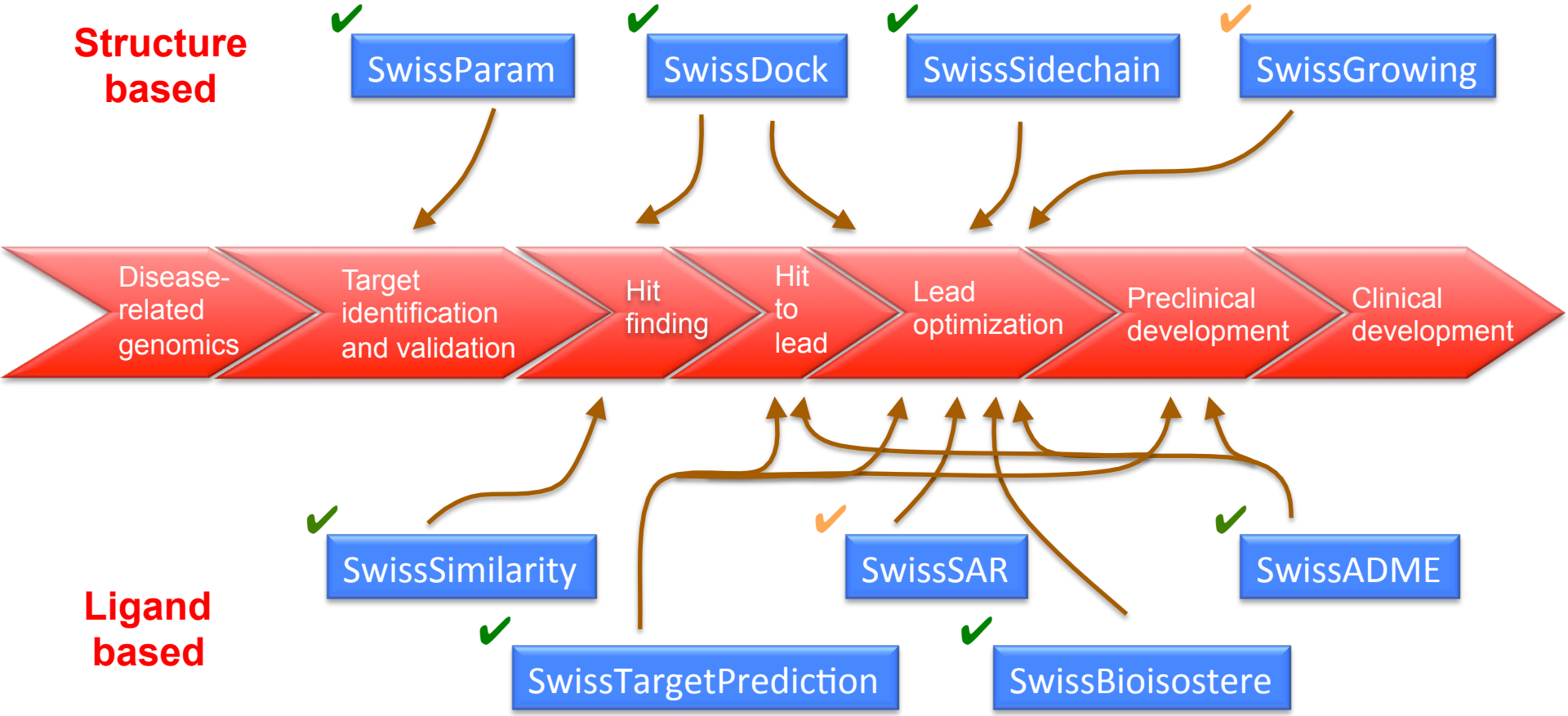
Computer-Aided Drug design (CADD)

Goal: use of **computing resources**, algorithms and 3D visualization (software, web-services, databases) to:

- help obtaining **rational ideas** about how to **create** or **modify** molecules,
- and to **take decisions** in the execution of the drug design process



Computer-Aided Drug design (CADD) at SIB

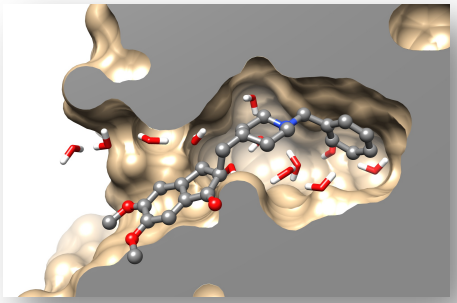


✔ : online ✔ : in development

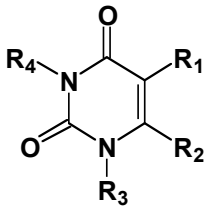
Computer-Aided Drug design (CADD)

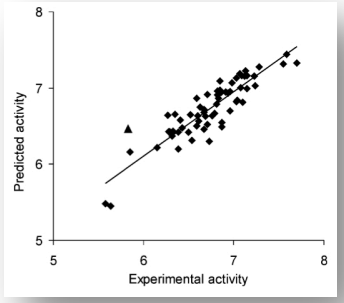
Two main categories of approaches to discover, create, optimize and evaluate active molecules:

- Structure-based approaches. Use the 3D structure of the targeted macromolecule. Ex: Molecular docking.

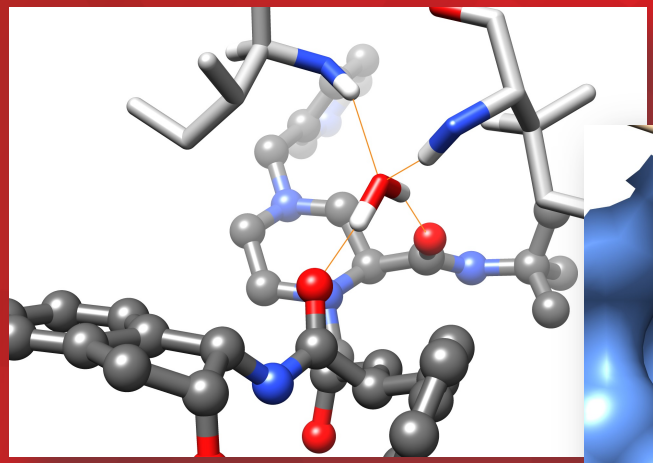


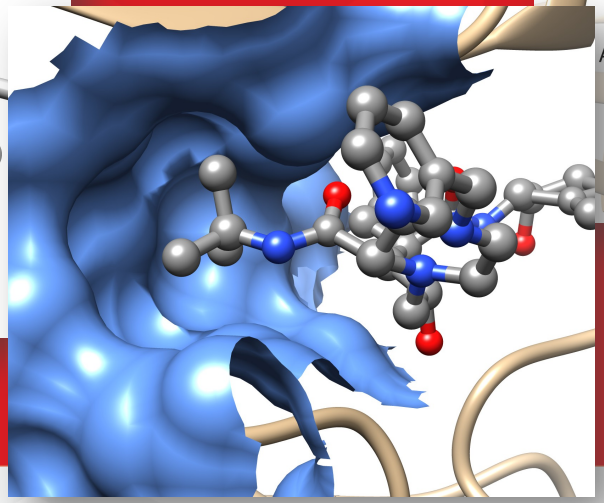
- Ligand-based approaches. Use the information derived from known ligands. Ex: Quantitative Structure-Activity Relationships (QSAR), bioisosteric replacements.

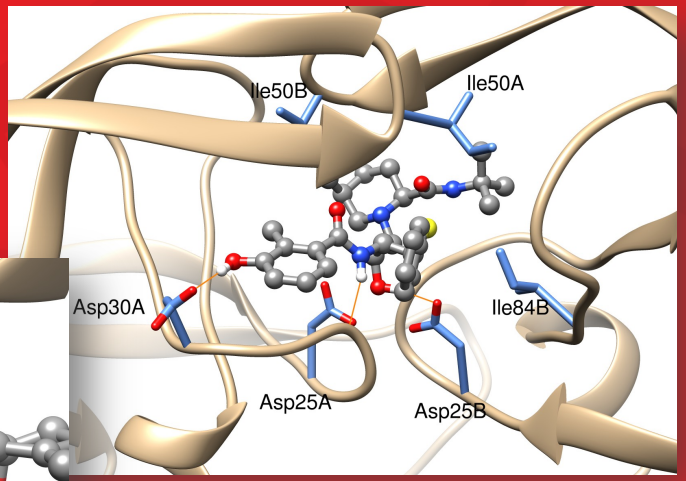




Intermolecular Recognition



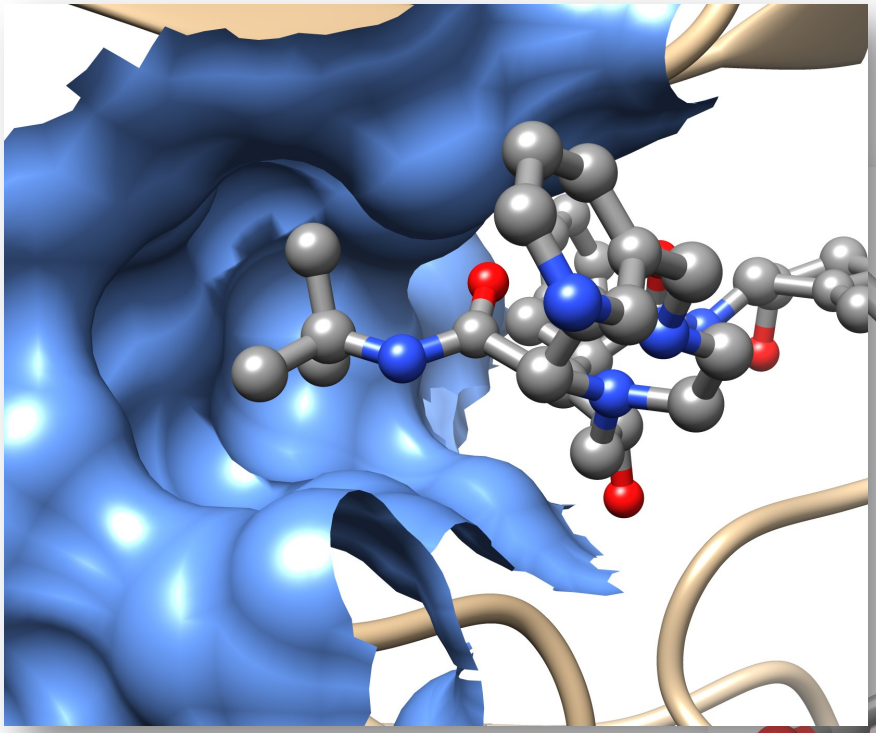


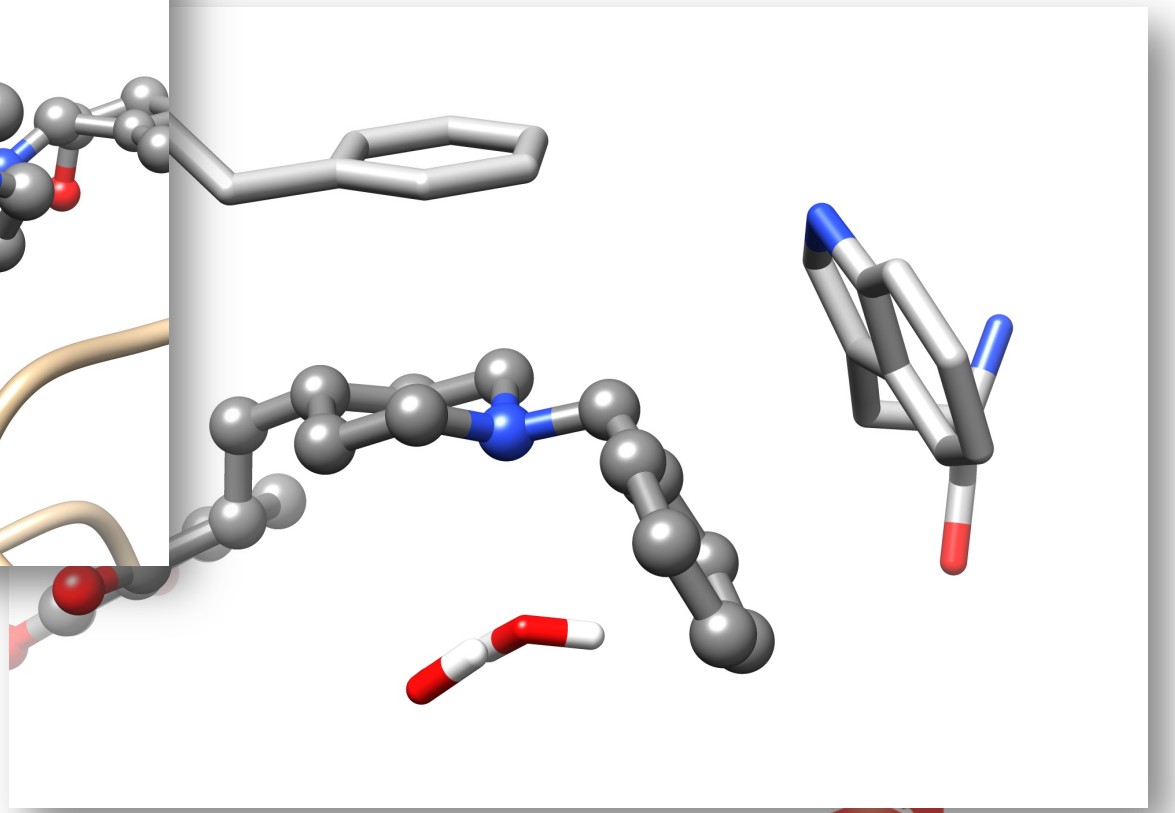


SIB
Swiss Institute of Bioinformatics

Molecular recognition - ligand and target

Designing a bioactive molecule is directly (structure-based design) or indirectly (ligand-based design) related to **mutual molecular recognition** between both partners: the **ligand** (small molecule) and its **biotarget** (protein).





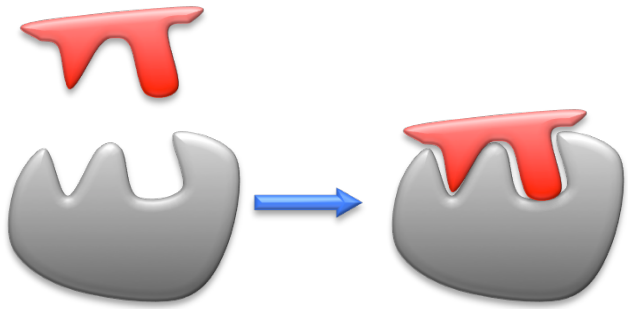
Crystal structure of a **ligand-protein complex** resolved by **X-ray** diffraction and displayed in Molecular Graphics environment (**UCSF Chimera**)

Molecular recognition - historical models

"Lock and key" model.

Emil Fischer in the 1890s.
The protein has a particular shape into which the ligand fits exactly.

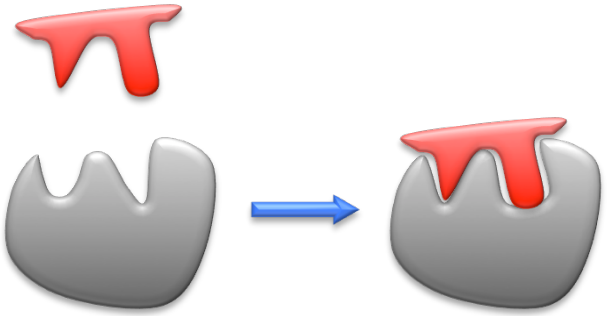
LIGAND



PROTEIN

Induced fit model

Daniel Koshland 1958.
The binding site of the macromolecule is flexible and its shape can be modified as the ligand interacts with it.



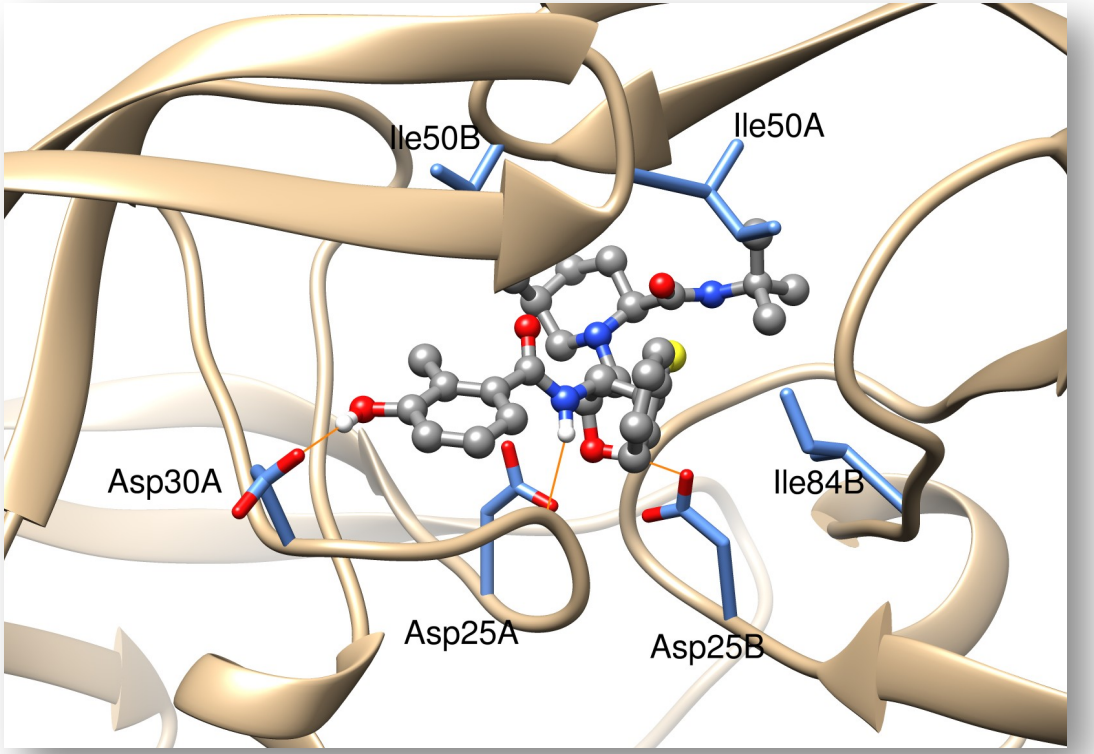
Molecular recognition:
Collection of **interactions** between molecules that govern their **binding**.

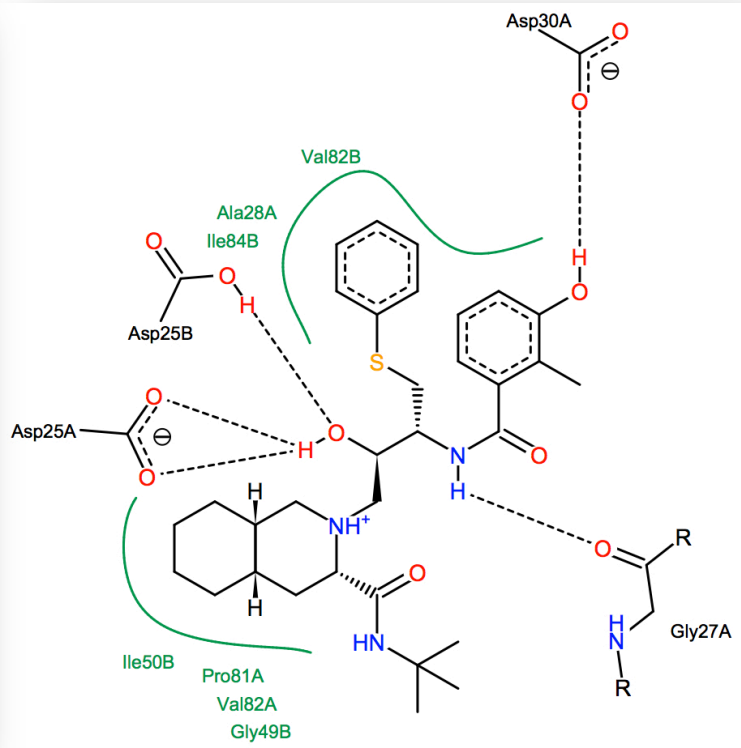
Qualitative **nature** of the interactions?
Quantitative **intensity** of the molecular recognition?

Molecular recognition - type of interactions

Non covalent interactions between atoms :

- non-polar interactions (shape recognition)
- electrostatic interactions (salt bridge and hydrogen bond)
- π interactions
- metal/ion interactions

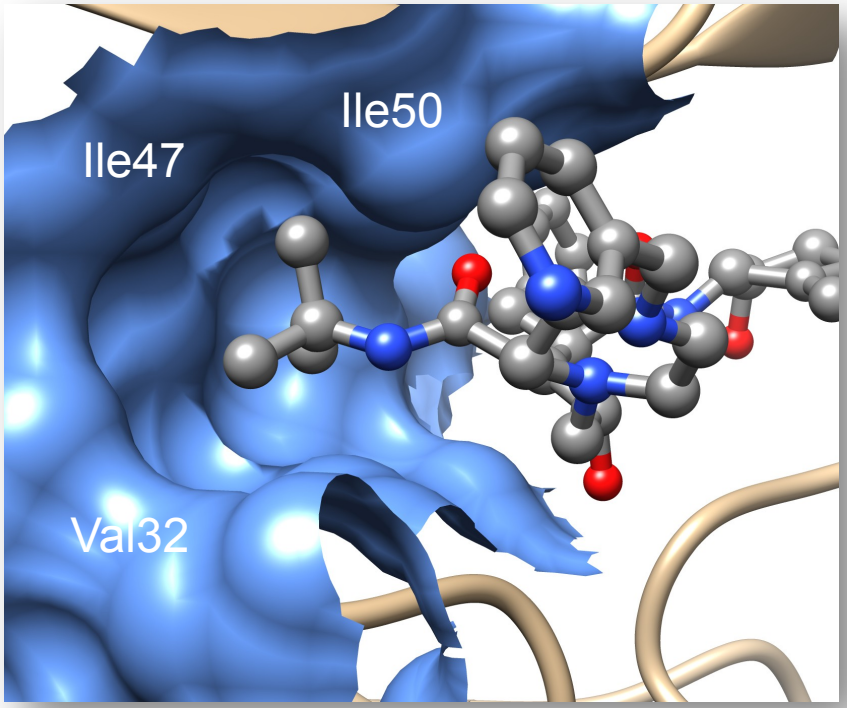




Crystallized complex Nelfinavir / HIV-1 protease (1OHR in PDB)

Molecular recognition – Non-polar interactions

Mainly involves **contacts between hydrophobic** parts of the molecular partners



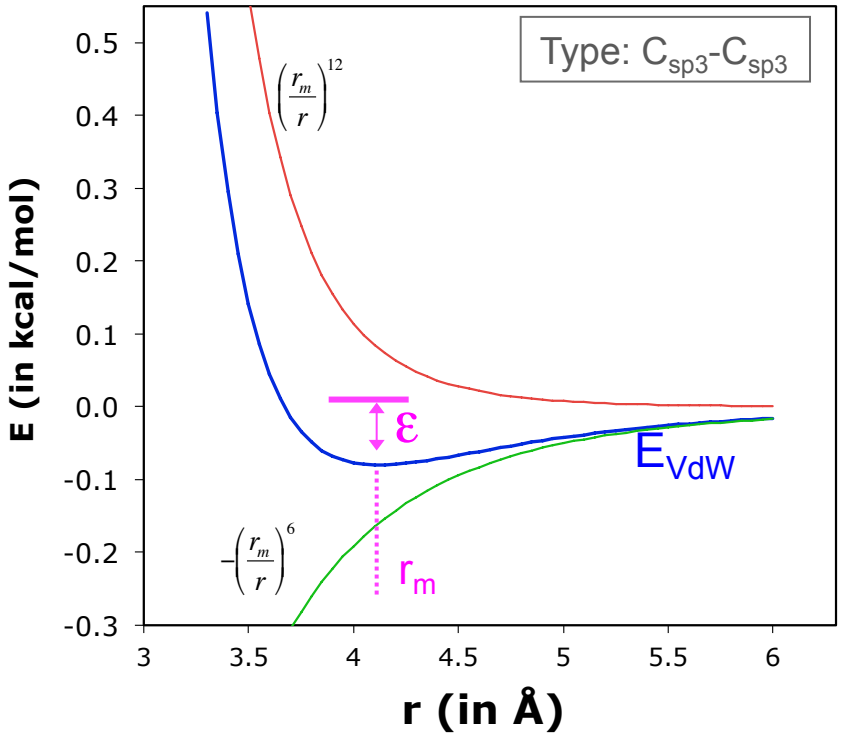
Ex: non-polar contacts between t-butyl function of indinavir and Ile50, Ile47 and Val32 of HIV-1 protease (1HSG in PDB)

Optimal contact distance between atoms

→ **Shape complementarity**

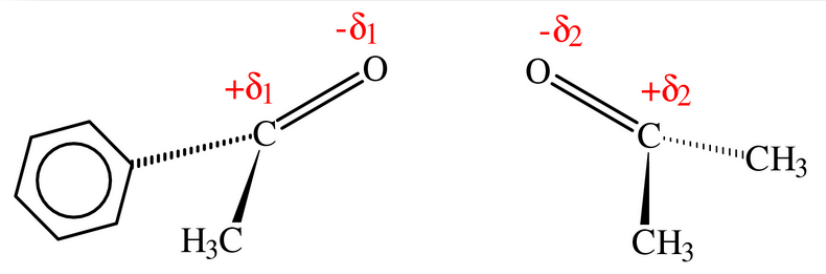
$$E = \varepsilon\left[\left(\frac{r_m}{r}\right)^{12} - 2\left(\frac{r_m}{r}\right)^{6}\right]$$

Lennard - Jones potential



Repulsive : Pauli exclusion principle
Attractive: induced dipole / induced dipole

Molecular recognition – Electrostatic interactions (salt bridge)

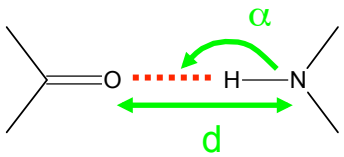


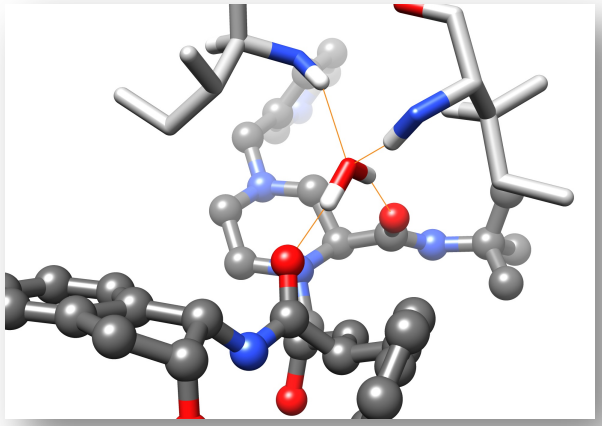
$$E_{elec} = \frac{q_i \, q_j}{4\pi\varepsilon_0 \varepsilon \, r_{ij}}$$

Coulomb law

where ε is the dielectric constant, 1 for vacuum, 4-20 for protein core, and 80 for water

Hydrogen bonds (H-bonds)



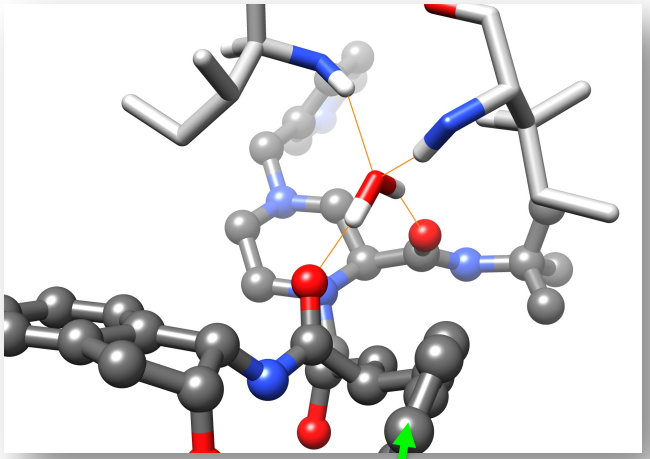


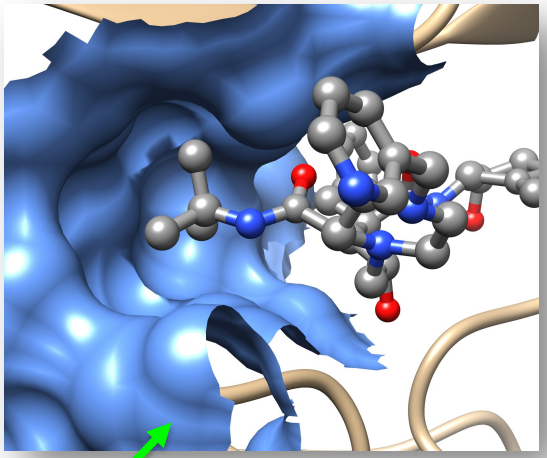
Electrostatic interactions are **local and directional** (H-bonds even more than salt bridges)

→ **Directionality / locality of interactions**
Specificity of molecular recognition

Ex: H-bond network between Indinavir, active site water molecule and Ile50 backbone NH of HIV-1 protease (1HSG in PDB)

Molecular recognition – Potency and specificity

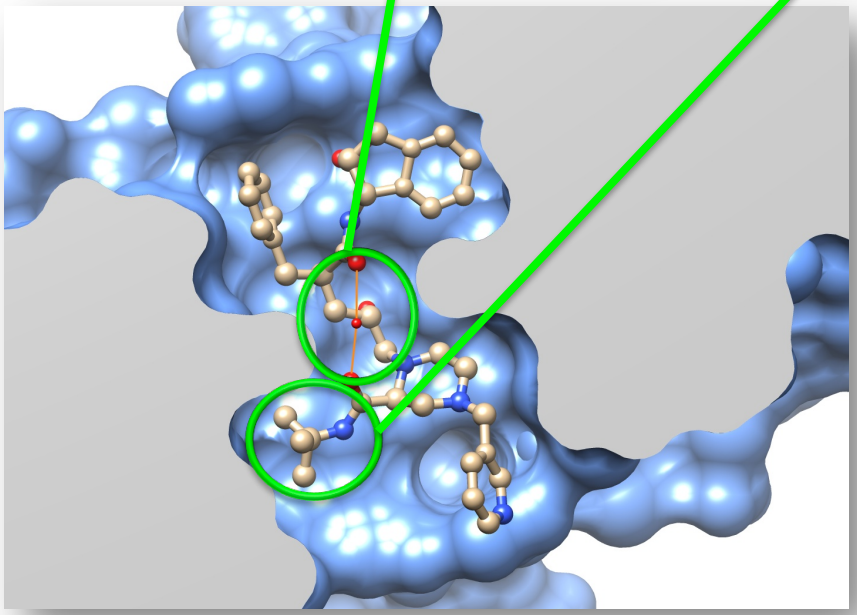




Various and numerous ligand-protein interactions:

- local and directional interactions

- shape complementarity

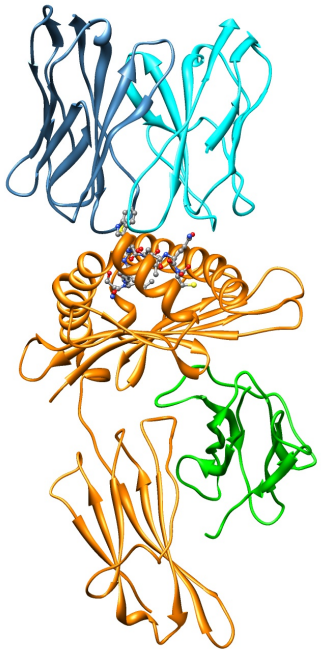


Specificity
(Limits side effects of drugs)

Affinity/potency
(Increased efficacy of drugs)

Closely related to Pharmacodynamic (PD) properties

Molecular recognition - Molecular Mechanics - Molecular Dynamics



- Adding explicit droplet of water:

System solvated with explicit water molecules (TIP3P model):

- **~ 29,500** water molecules
- **~ 100,000** atoms in total

- Molecular Dynamics (**MD**)

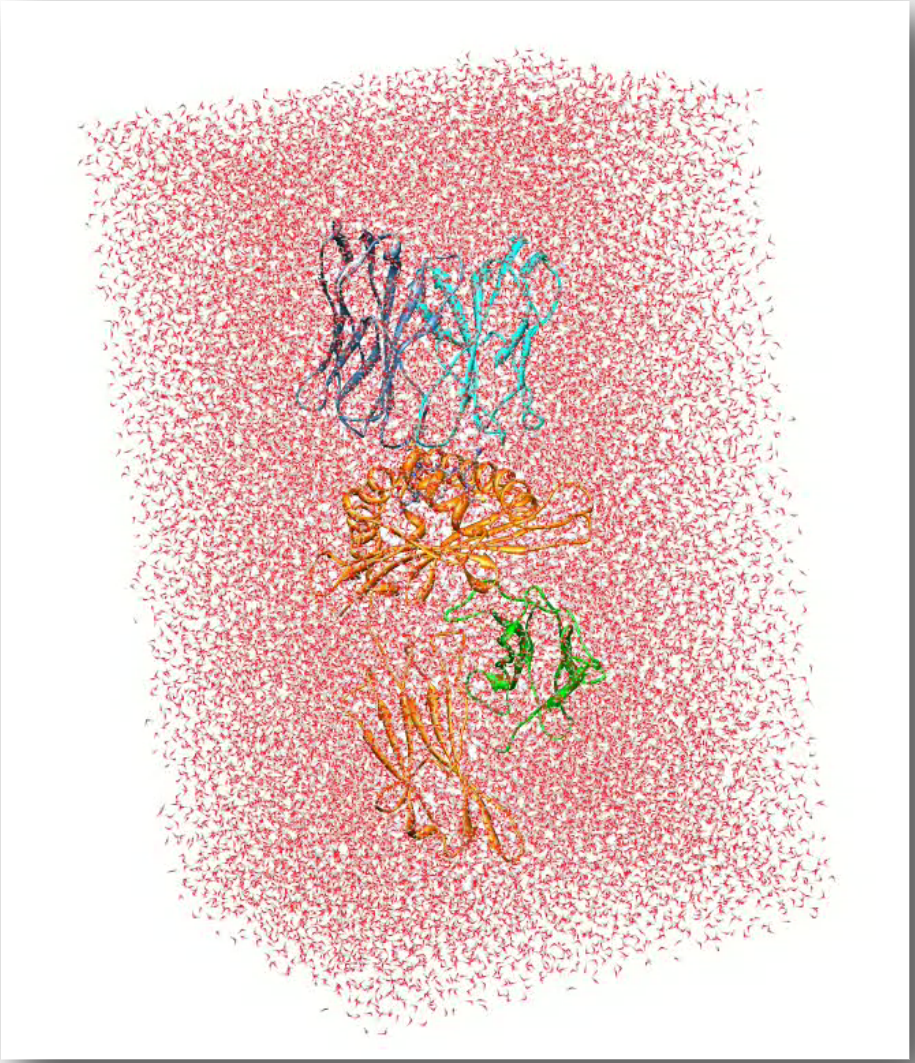
Atom motions are calculated to follow Newton's equation of motion, at **300 K** and **1 atm**.

Typical simulation times: from **0.5 ns to ~ 100 ns** (10^{-9} s).

Energy terms averaged over 200 to 500 **"snapshots"** extracted from the MD simulation **trajectory**.

➔ **Simulation closer to physiologic reality, but much more computationally intensive**

Molecular recognition - Molecular Dynamics



- Adding explicit droplet of water:

System solvated with explicit water molecules (TIP3P model):

- **~ 29,500** water molecules
- **~ 100,000** atoms in total

- Molecular Dynamics (**MD**)

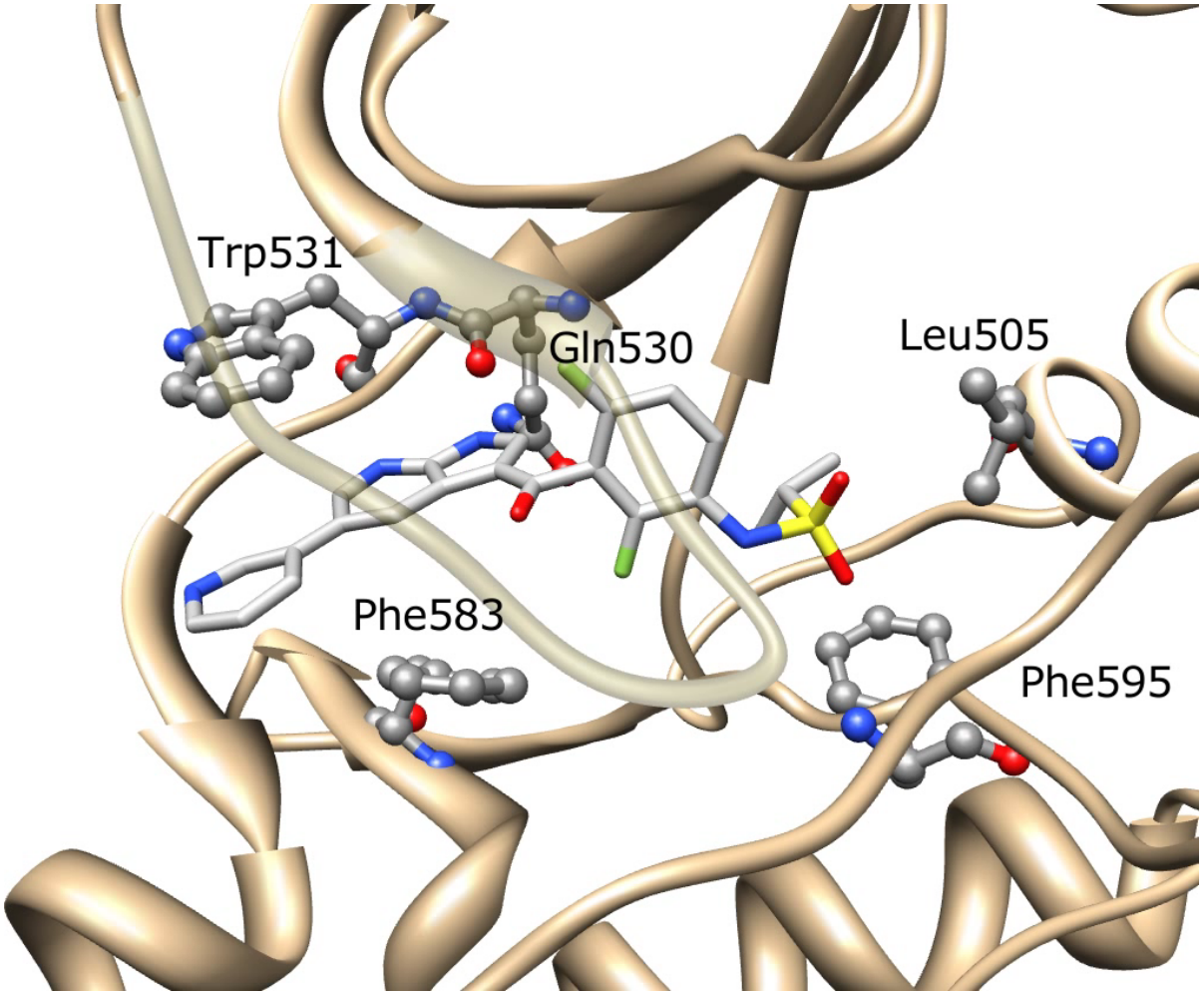
Atom motions are calculated to follow Newton's equation of motion, at **300 K** and **1 atm**.

Typical simulation times: from **0.5 ns to ~ 100 ns** (10^{-9} s).

➔ **Simulation much closer to physiologic reality, but much more computational intensive**

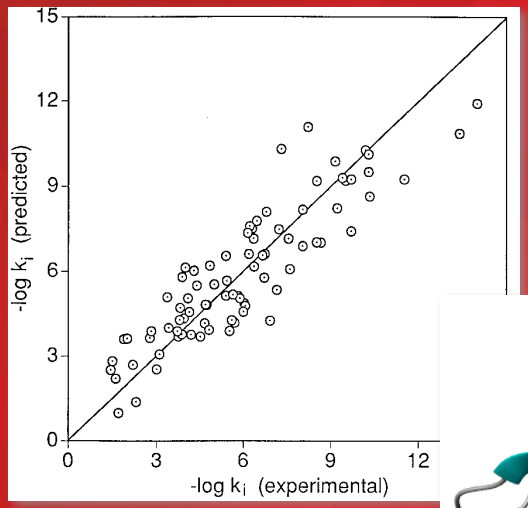
Molecular recognition - Molecular Dynamics

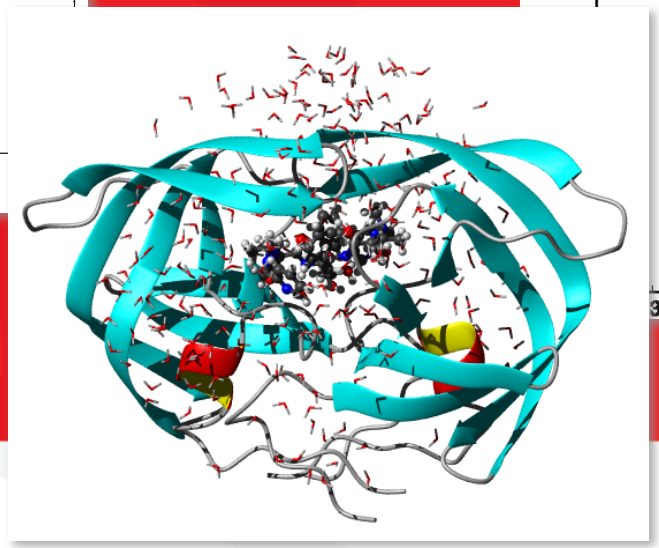
Typical motions in a ligand/protein complex at room temperature:

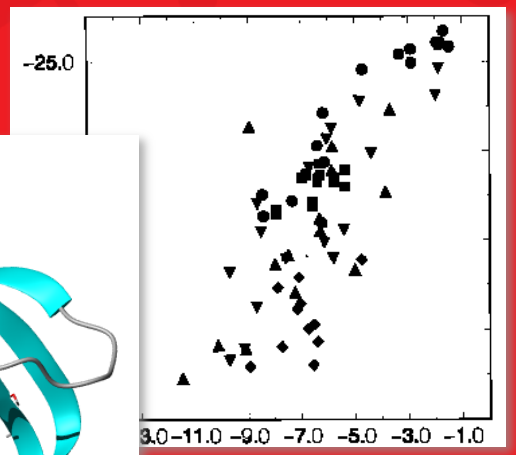


PLX3203 in BRAF

Binding Affinity Estimation







SIB
Swiss Institute of Bioinformatics

Affinity estimation - Binding, dissociation and inhibition constants. - Binding free energy

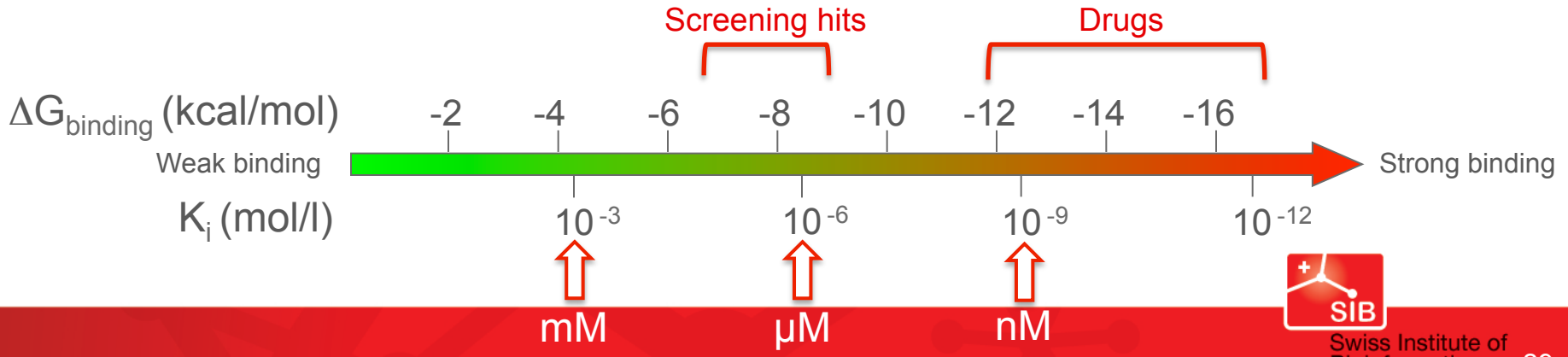
$$A + B \leftrightarrows A'B' \qquad K_b = \frac{[A'B']}{[A][B]}$$

$$A'B' \leftrightarrows A + B \qquad K_d = K_i = \frac{[A][B]}{[A'B']}$$

K_b : binding constant, K_d : dissociation constant, K_i : inhibition constant

Calculated — Determined Exp.

$$\Delta G_{binding} = -RT \ln K_b = RT \ln K_i = \Delta H - T\Delta S$$



Affinity estimation - The computational methods

Many available methods
All give estimations

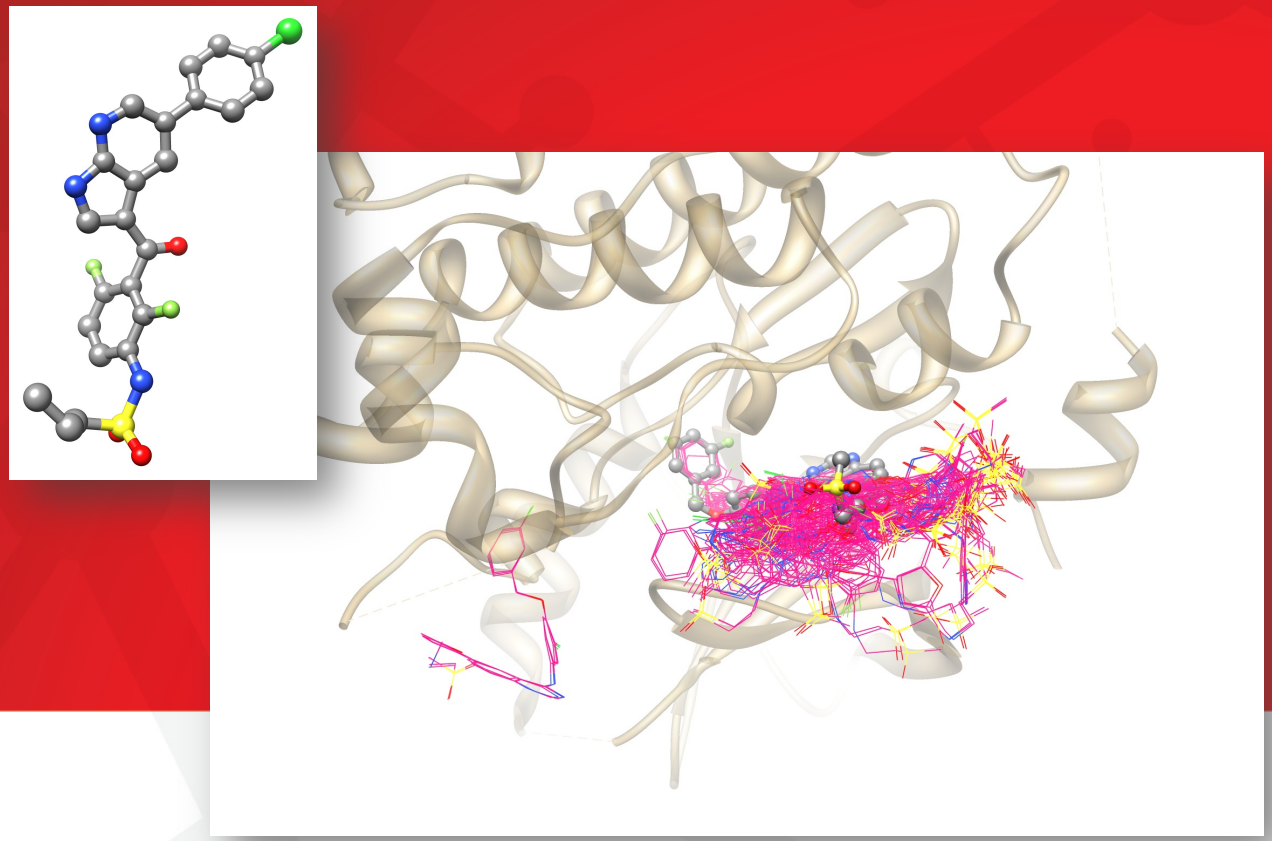
Ligand-based

- 2D QSAR. Ex: Hansch equations
- 3D QSAR. Ex: CoMFA

Structure-based

- **Force field** methods:
  - Free energy simulation (FEP, TI)
  - MM-PBSA, MM-GBSA
  - Linear interaction energy (LIE)
- **Empirical** scoring functions (regression based approaches). Ex: LUDI score
- **Knowledge**-based approaches (Potential of Mean Force). Ex: PMF score

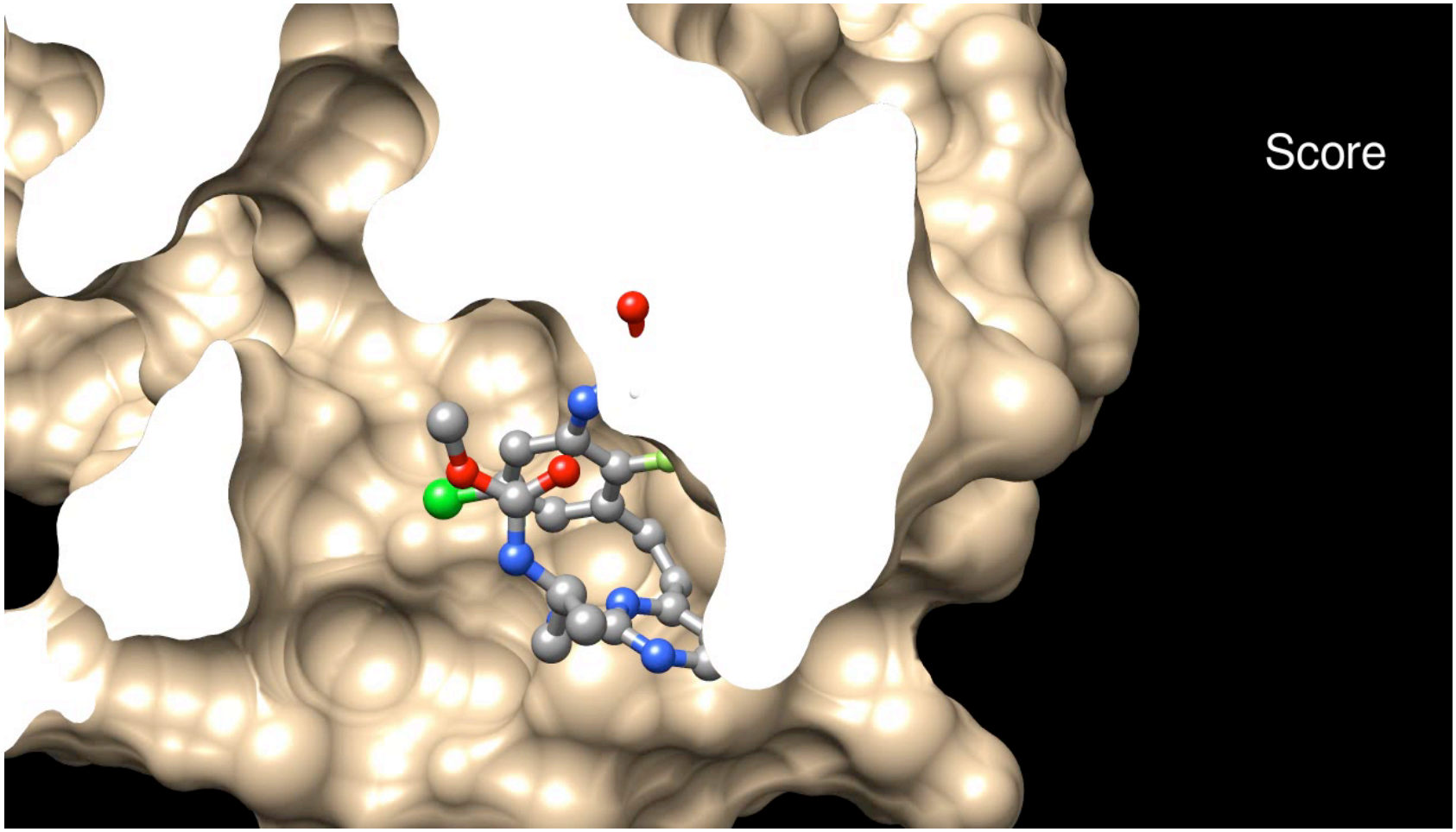
Small Molecule Docking



SIB
Swiss Institute of Bioinformatics

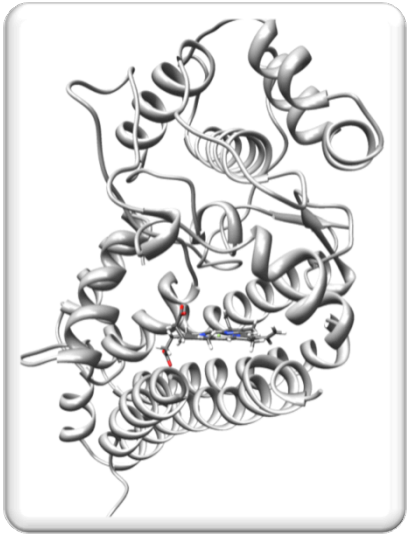
Docking small molecules into protein cavities

Find the **location**, the **orientation** and the **conformation** of the small molecule on the protein surface

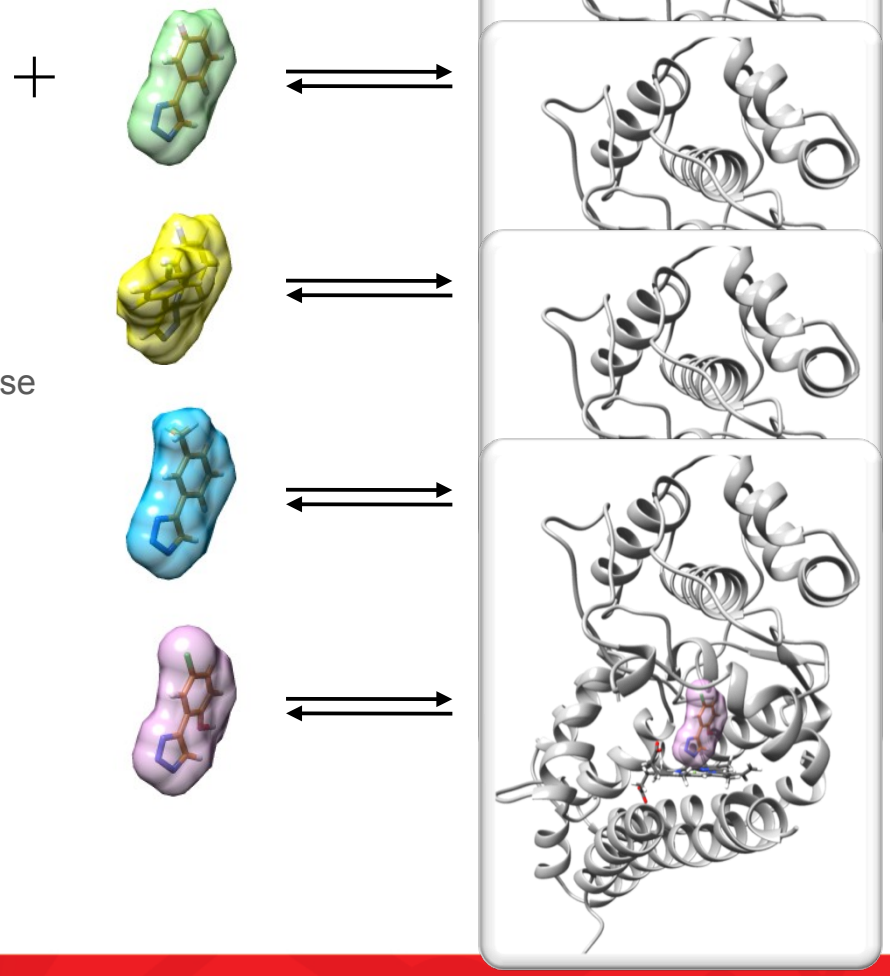


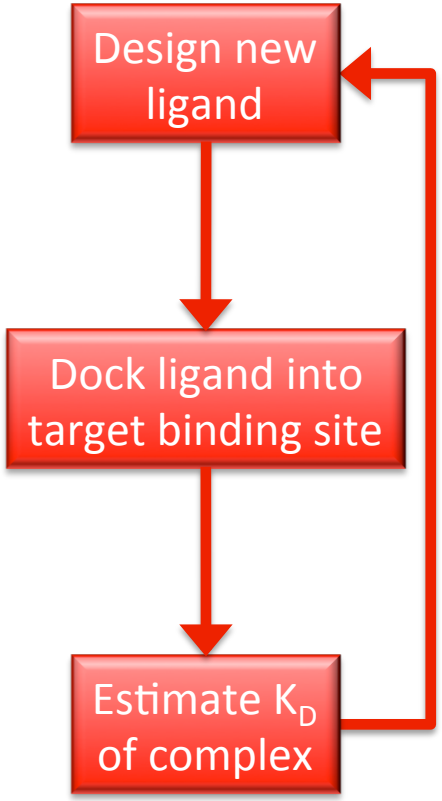
Docking small molecules into protein cavities

Towards structure-based ligand design….



Indoleamine 2,3-Dioxygenase (IDO)





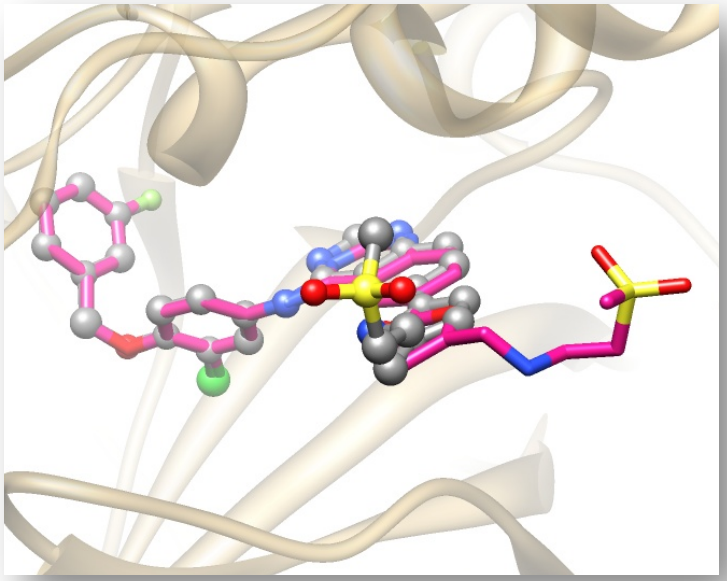
Docking - Definitions

Pose: location, orientation and conformation of a small molecule on a macromolecule surface (cavity, pocket of groove) ~ tentative binding modes.

Native binding mode: experimentally defined binding mode (X-ray, NMR). Expected to be the best binding mode in term of binding free energy.

Docking: predicting the (native) binding mode using molecular modeling approaches.

Success: ability to predict a binding mode close to the native binding mode (when known, *i.e.* exercise or benchmark of the approach).



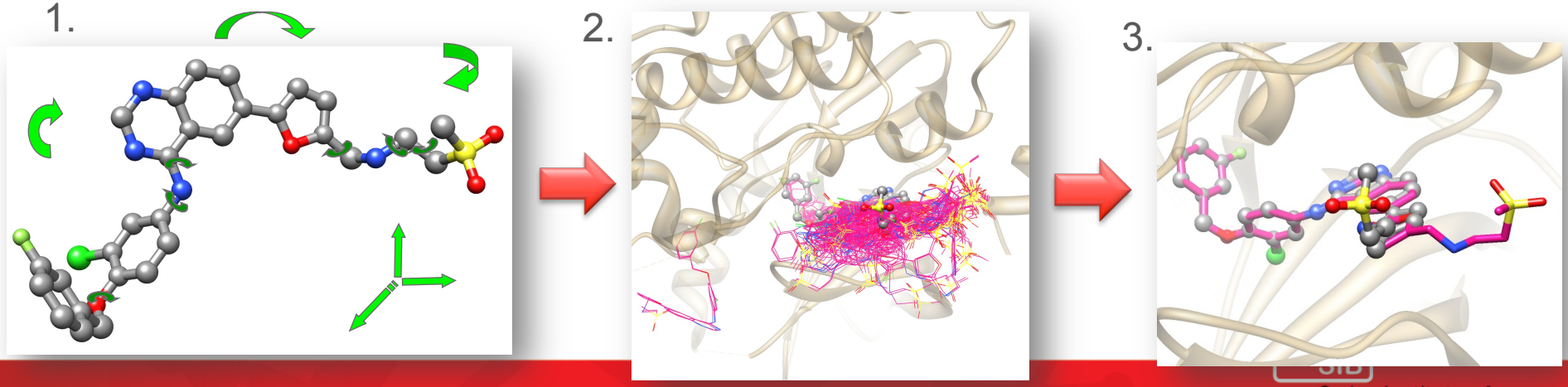
Docking - Basics

The “standard” methodology:

1. **Generate a large number of poses**. Sampling the posing space (orientational/translational/conformational) of the ligand into the protein binding site
2. **Assess the binding strength**. Scoring each possible ligand pose (~ fast evaluation of the ligand affinity)
3. **Selecting** the pose with the most favorable binding (**best score**) ➔ **predicted binding mode**

Different levels of approximation:

- protein and ligand are rigid (the past!)
- the protein is rigid, the ligand is flexible (today, except HTVS)
- protein and ligand are flexible (possible today at computational cost)



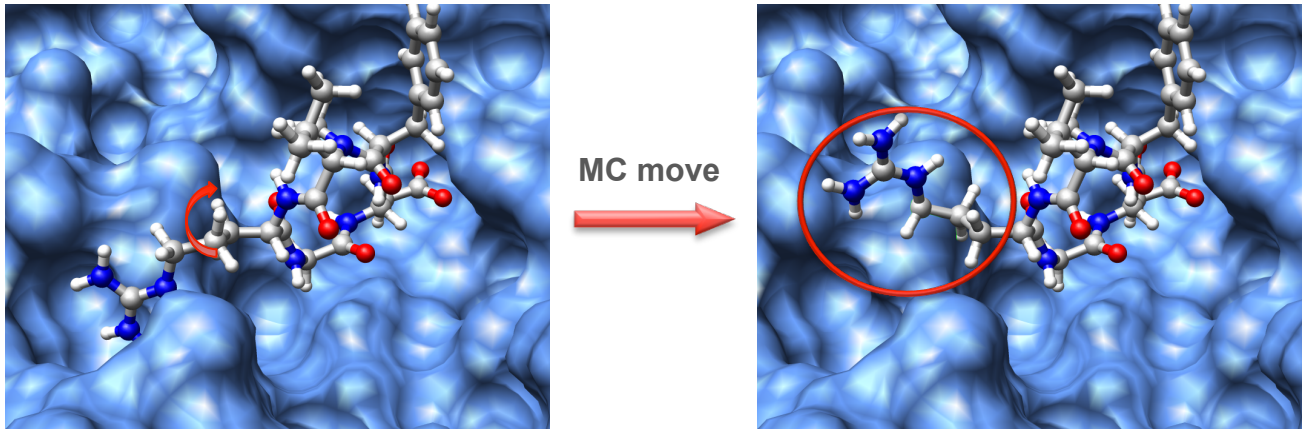
Docking - Posing algorithm

Three types of posing search:

Systematic search	• Combinatorial exploration of degrees of freedom • Incremental construction
Stochastic search	• Monte Carlo (MC) • Evolutionary algorithm (EA) • Particle swarm optimization
Deterministic search	• Minimization • Molecular dynamics simulation (MD)

Docking - Posing algorithm - Stochastic search.
- Monte Carlo (MC)

1. Generate an **initial pose** (ligand random conformation, translation and rotation) and **score it**
2. Generate an **new pose from the previous one** (through random conformational change, translation, rotation) and **score it**
3. Use **metropolis criterion**(*) to determine whether the new pose is retained
4. Repeat steps 2-3 until the number of desired poses is obtained (typically >100,000)



(*) Metropolis criterion:
If the difference in energy between the new and previous pose (ΔE) is negative (i.e. the new pose has better interactions with the protein), then the new pose is accepted.
If ΔE is positive, a random number between 0 and 1, $0<X<1$, is generated and the new pose is accepted only if $\exp(-\Delta E/RT)>X$.

Docking - Scoring - Evaluation of Binding ~Affinity estimation

The two roles of scoring functions:

1. Rank poses for one ligand in a given target (docking ➔ predict binding mode)
2. Rank the binding modes of different ligands for a given target (compounds **selection** in **lead optimization** or **virtual screening** in hit-finding in **large databases**).

➔ must be a quick estimate of the binding.

Structure-based

- **Force field** methods:
  - Simplified force fields
  - MM-PBSA, MM-GBSA
- **Empirical** scoring functions (regression based approaches). Ex: LUDI score
- **Knowledge**-based approaches (Potential of Mean Force). Ex: PMF score, DrugScore, GoldScore

Docking - Success

Success: ability to predict a binding mode close to the native binding mode (redocking, i.e. exercise or test of the approach).

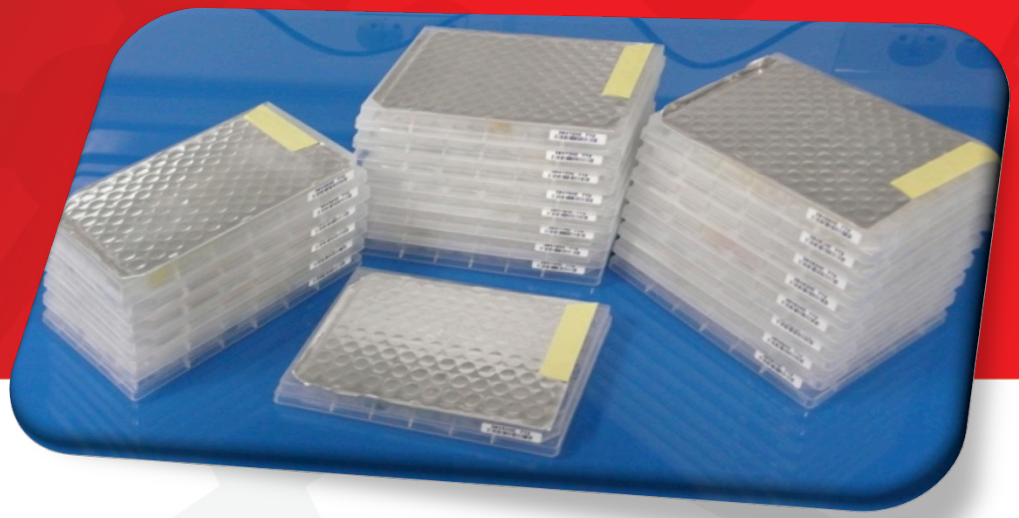
Success rate : ~ 50 to 90 % in benchmarks (re-docking)

in “real” application (cross docking) the success rate decreases by at least 30%

➔ room for improvement.

➔ need for high precision docking programs, handling protein flexibility (but also water, ion, ...).

Experimental Screening



SIB
Swiss Institute of
Bioinformatics

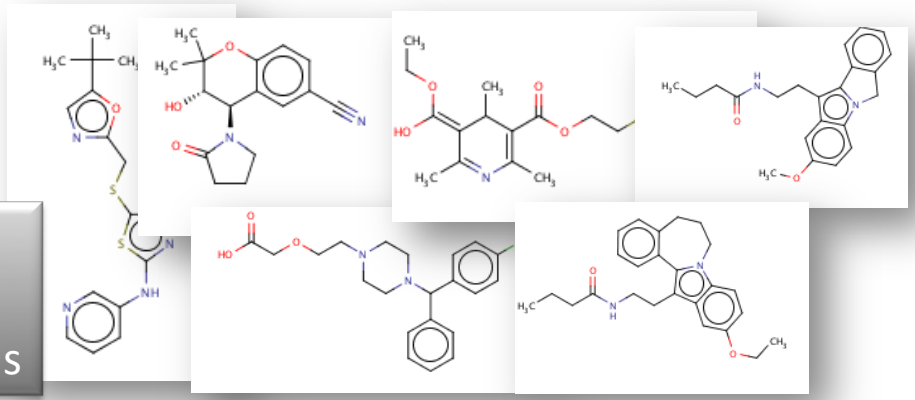
Molecular screening – experimental HTS

Experimental High Throughput Screening (HTS) consists in testing, using an automatized assay, a large collection of small molecules against a given target

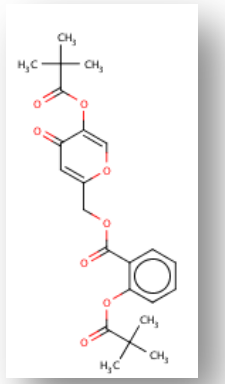
Target

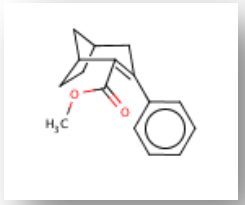
+

Collection Of small compounds



Automatized assay





Hits

Molecular screening – experimental HTS

Example of commercially available "general" collections:

Vendor	Num. of compounds (Sep. 2014)
Asinex	323,891
Chembridge	1,037,068
ChemDiv	1,493,991
Enamine	1,896,486
LifeChemicals	379,184
Otava	264,024
Princeton	970,438
SPECS	212,782
TimTec	749,038
Vitas	877,583

For more, see http://zinc.docking.org

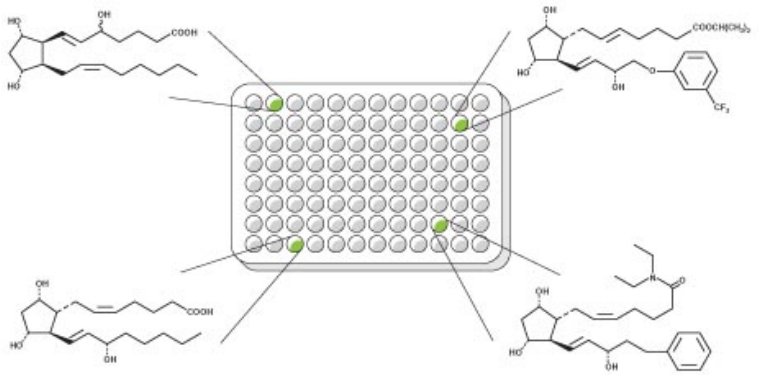
Cost:

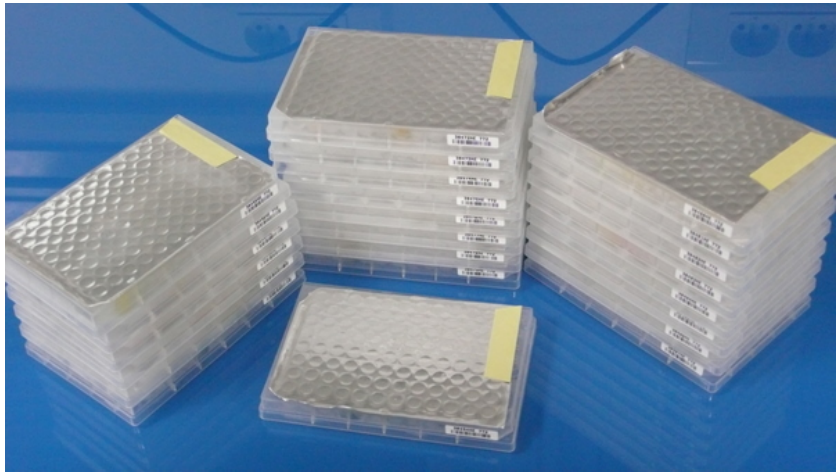
- ~ 1 to 15 $ per molecule when buying entire large collection
- ~ 100 $ per molecule for cherry picking a few compounds among different collections

Pharmaceutical companies have proprietary collections of up to 10 millions compounds

Molecular screening – experimental HTS

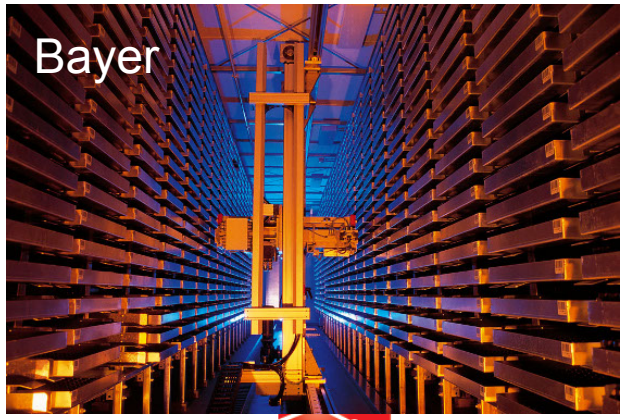
Format of the experimental collections:





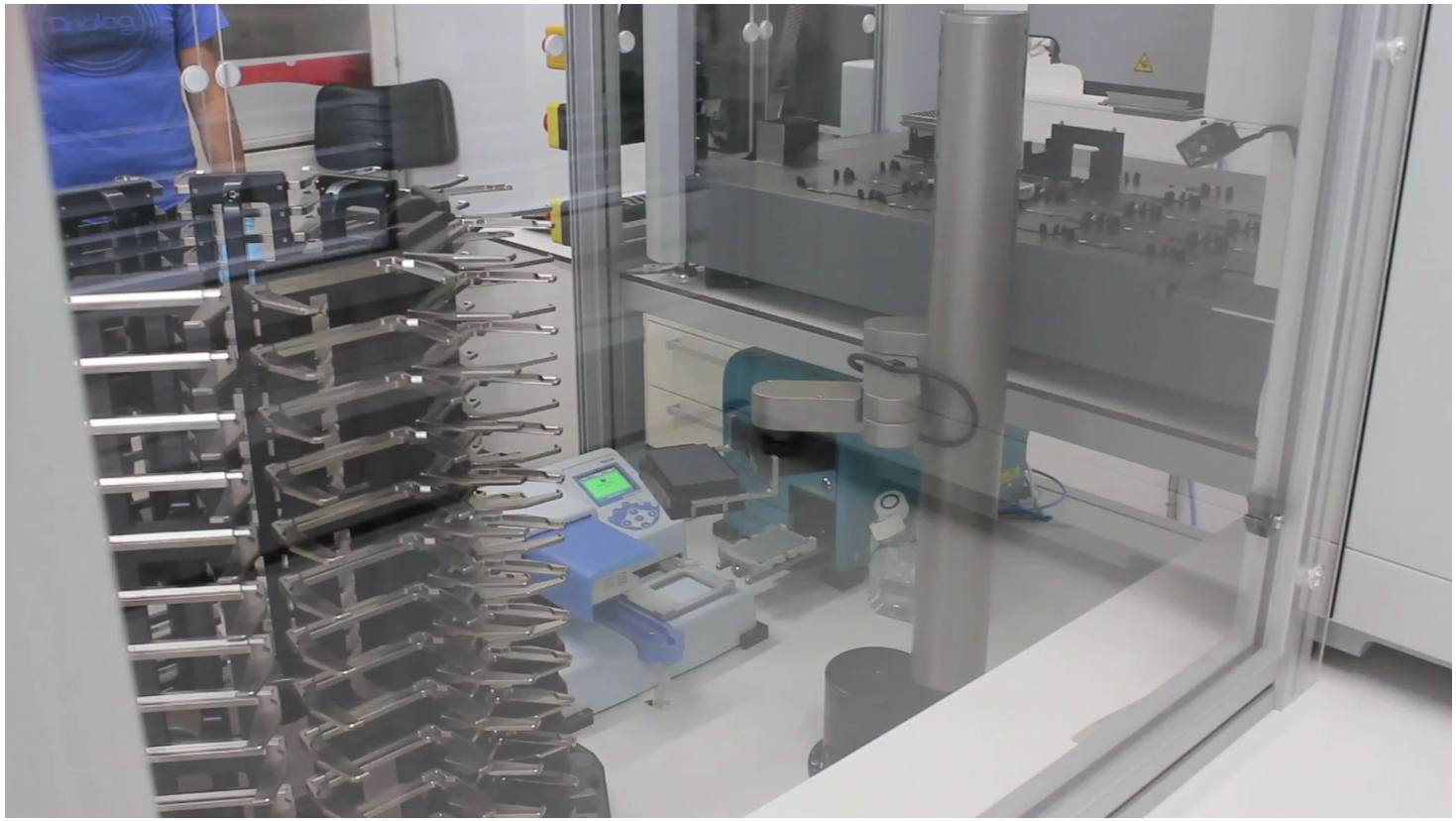
University collection…

… pharmaceutical company collection.



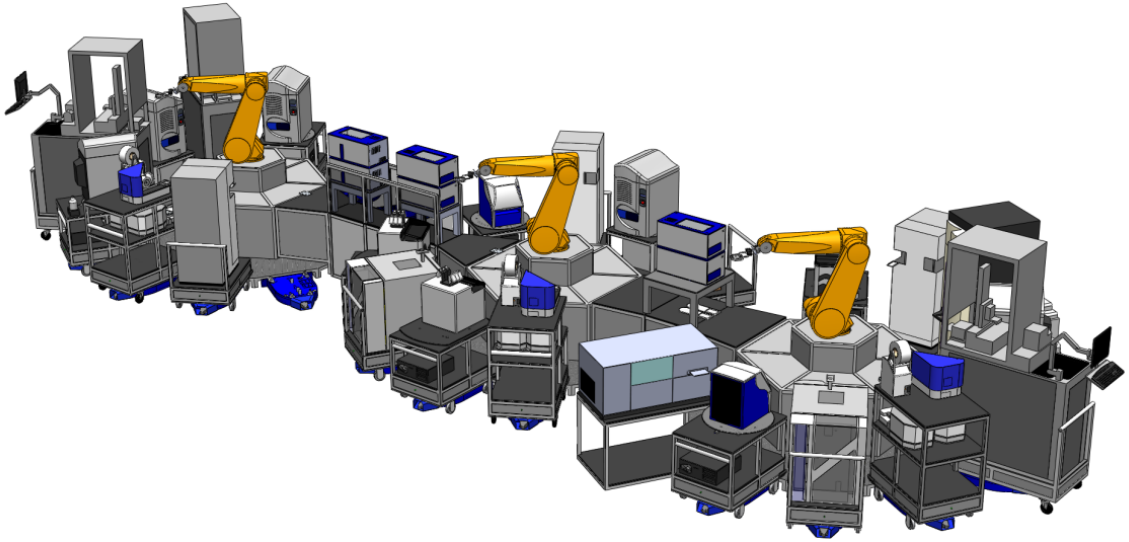
Molecular screening – experimental HTS

Assays are simplified, automated and performed using robots:



Molecular screening – experimental HTS

The industrial version :



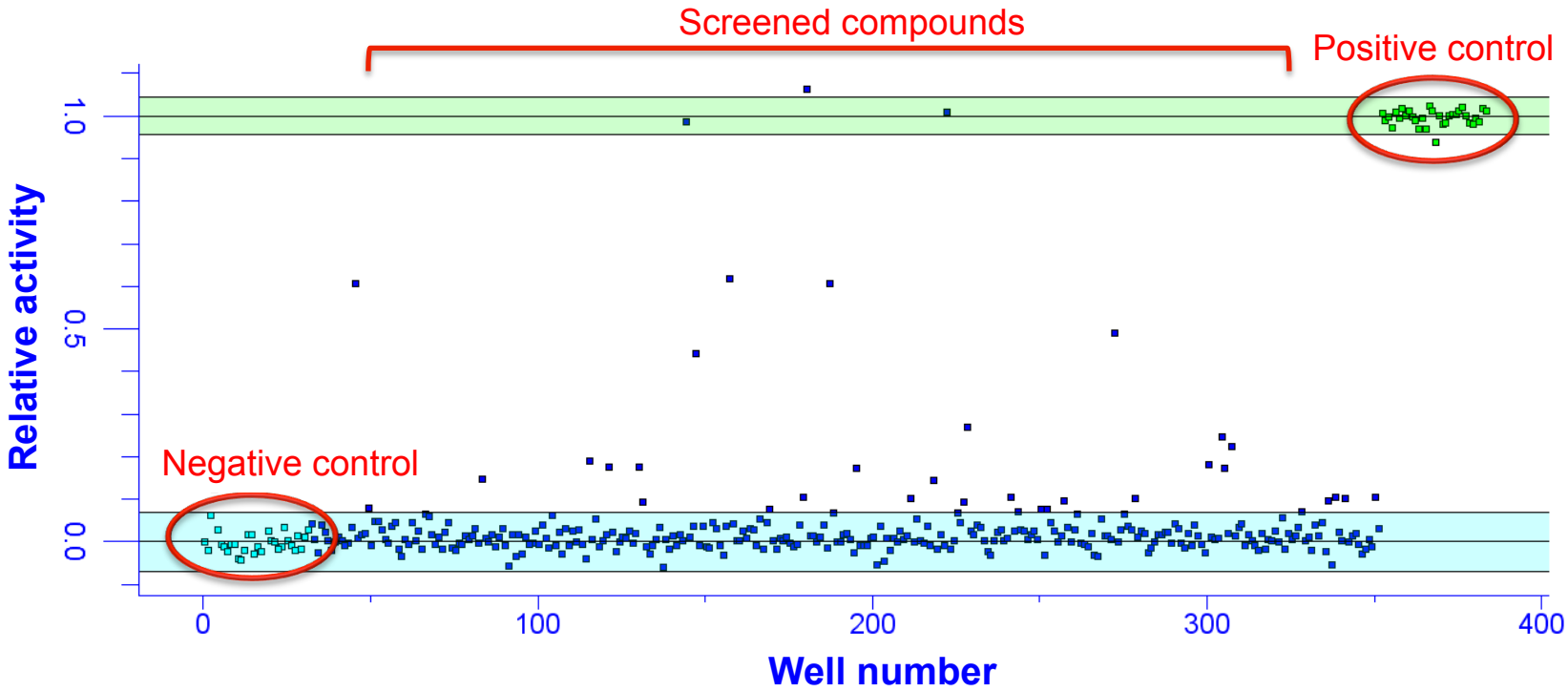
Typical capacity: more than **100,000 compounds per day** in high-density 1536-well formats.

Molecular screening – experimental HTS

Type of results

Ex.: screening Maybridge Hit Finder collection against IDO1

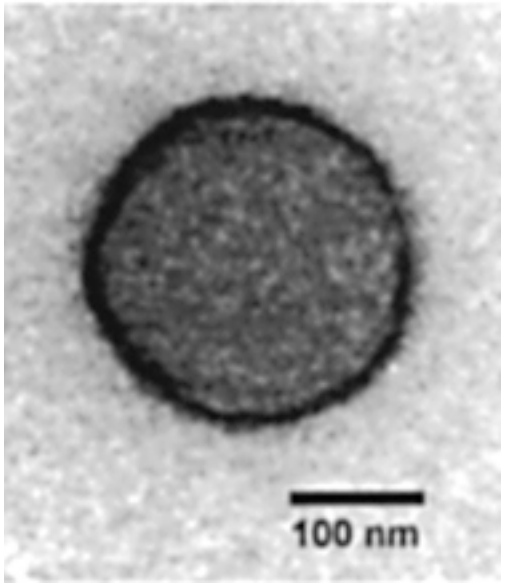
Röhrig, U. F.; Majjigapu, S. R.; Chambon, M.; Bron, S.; Pilotte, L.; Colau, D.; Van den Eynde, B. J.; Turcatti, G.; Vogel, P.; Zoete, V.; Michielin, O. Detailed Analysis and Follow-Up Studies of a High-Throughput Screening for Indoleamine 2,3-Dioxygenase 1 (IDO1) Inhibitors. Eur J Med Chem 2014, 84C, 284–301.

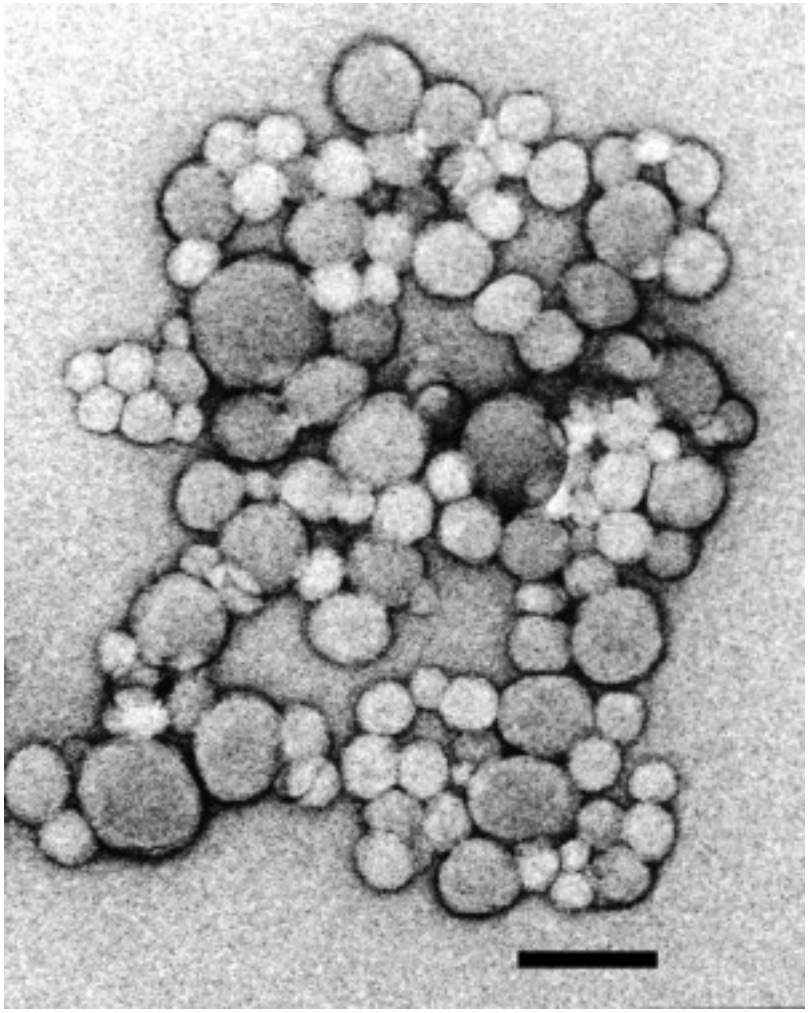


Molecular screening – experimental HTS. Example of exp. issue

Aggregators:
Promiscuous inhibitors form aggregates of 50-400 nm diameter that can absorb proteins and block them

McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A Common Mechanism Underlying Promiscuous Inhibitors From Virtual and High-Throughput Screening. *J. Med. Chem.* **2002**, *45*, 1712–1722.





In silico methods can be used to filter these compounds out

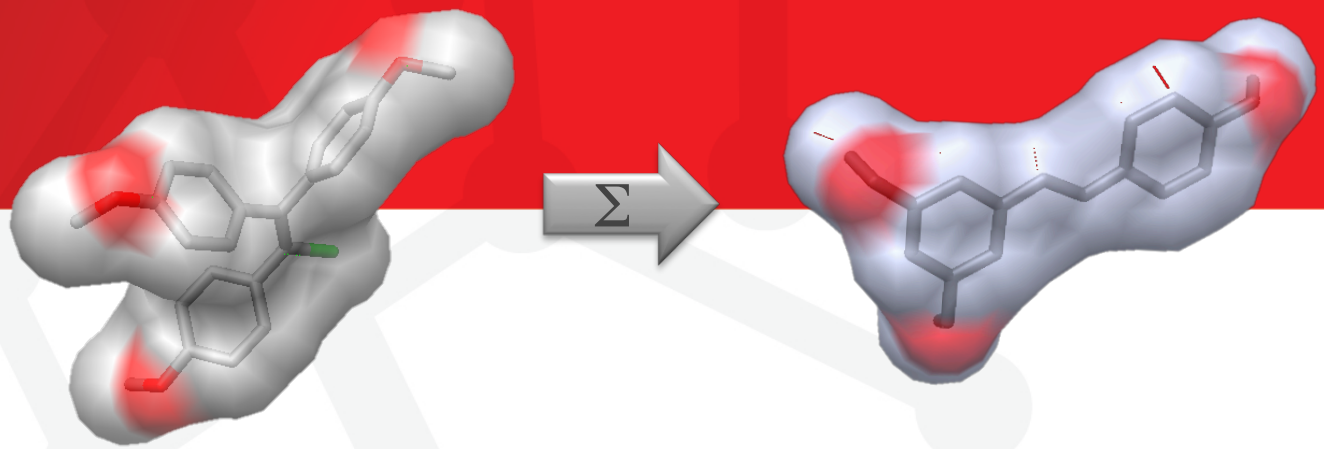
Molecular screening – experimental HTS

Success rate depends on the nature of the:

- **target**. Compound collections are full of molecules developed for intensively studied targets (e.g. GPCR). As a consequence, those targets lead to high success rate in HTS compared to new targets
- **interaction**. E.g. protein-protein interaction (PPI) inhibitors are hard to find
- **library**. E.g. kinase-focused libraries will provide higher success rate against kinases… but potentially less interesting discoveries

- Hits found in **half of the HTS campaigns**
- Number of hits corresponds to **1 to 3%** of the content of standard libraries against usual targets (0.1 to 1% for PPI).

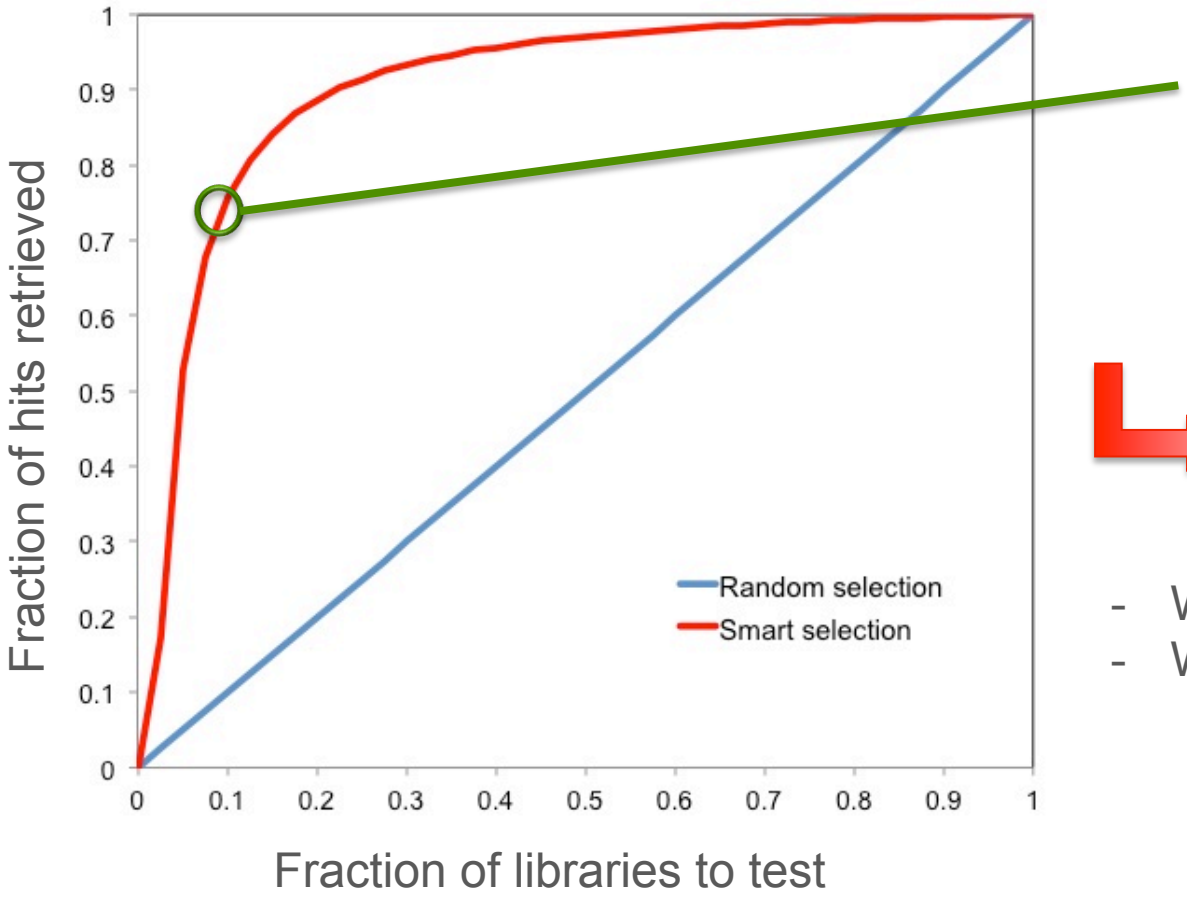
Ligand-based Virtual Screening



SIB
Swiss Institute of
Bioinformatics

Molecular screening – Virtual screening

Virtual screening: Use of CADD algorithms to decrease the number of molecules to test experimentally. i.e. Create a short list of molecules to test in priority



Only need to screen 10% of the library to retrieve 75% of the hits

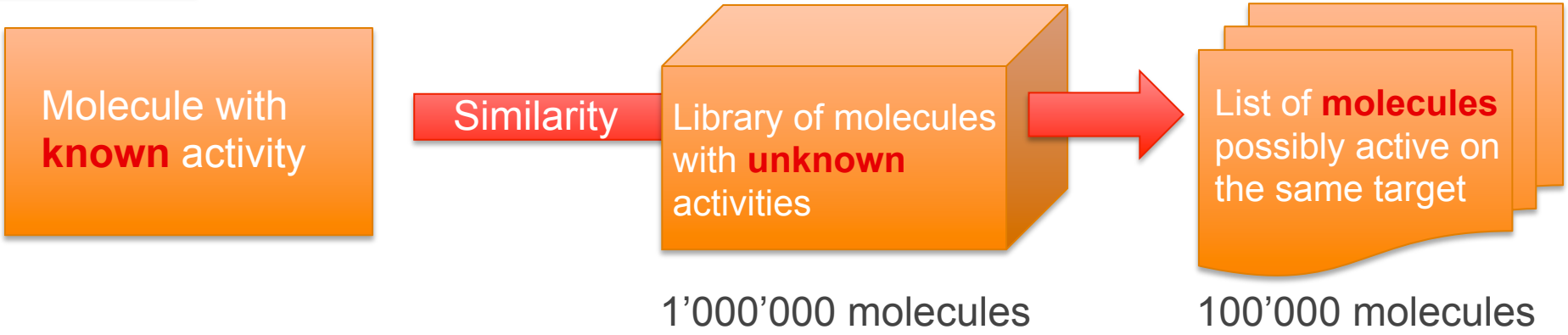
Enrichment procedure

- Will not find all actives
- Will not find only actives

Molecular screening – Ligand-based virtual screening

Assumption: **if two molecules are very similar, they are prone to be active on the same target**

Screening:



We need at least one molecule active on the target of interest

This molecule does not need to be a drug. It can be:
- a substrate
- a natural ligand
- a molecule from the competitors

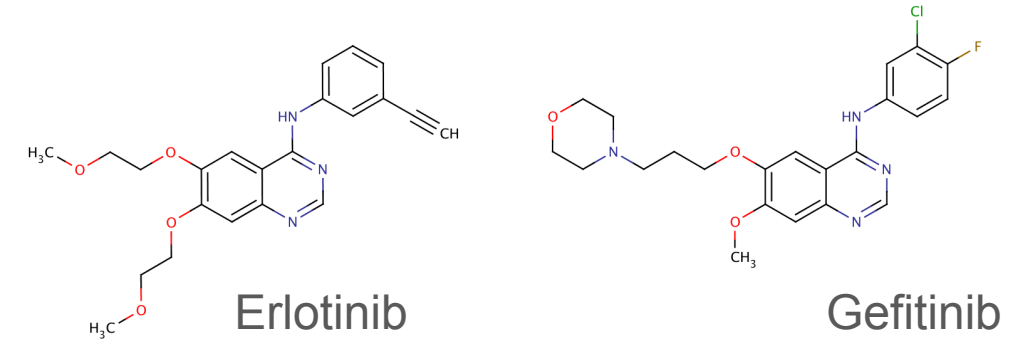
If several molecules are available, several screenings can be done.

→ Consensus screening

Molecular screening – Ligand-based virtual screening

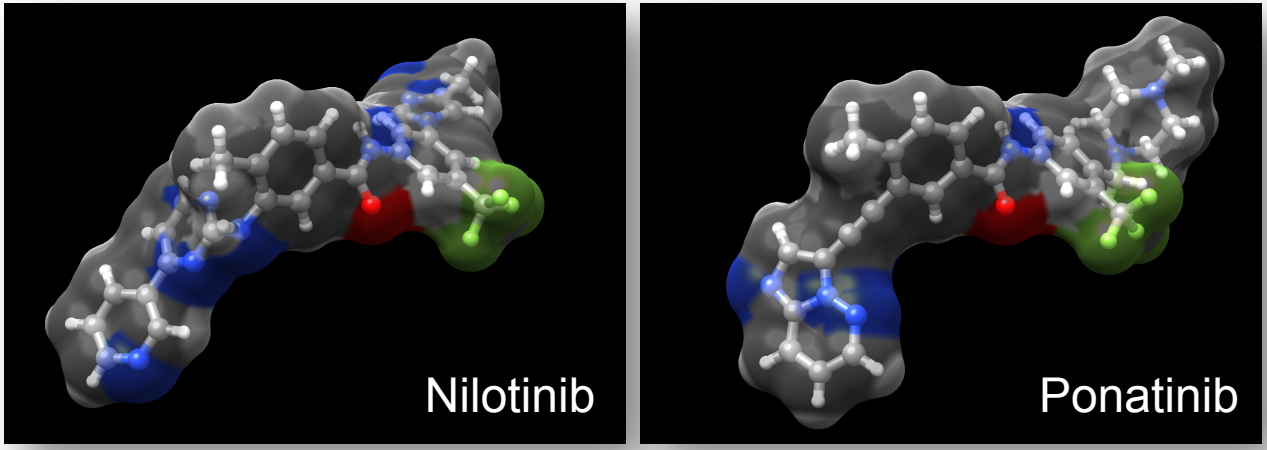
Assumption: **if two molecules are very similar, they are prone to be active on the same target**

- 2D: Similar by chemical structure (Fingerprints)



Very fast, but quite limited to similar scaffolds

- 3D: Similar by shape (electrostatics and lipophilicity)

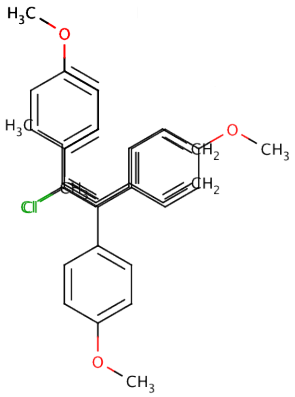


Slower, but allows various scaffolds

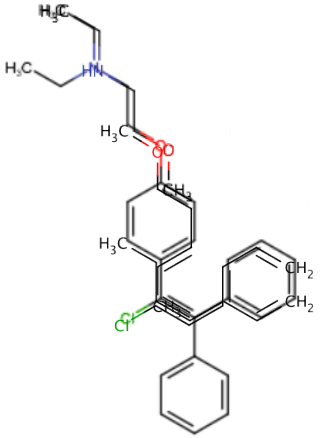
Molecular screening – 2D Ligand-based virtual screening

Chemical similarity (2D fingerprints)

Identify molecular features

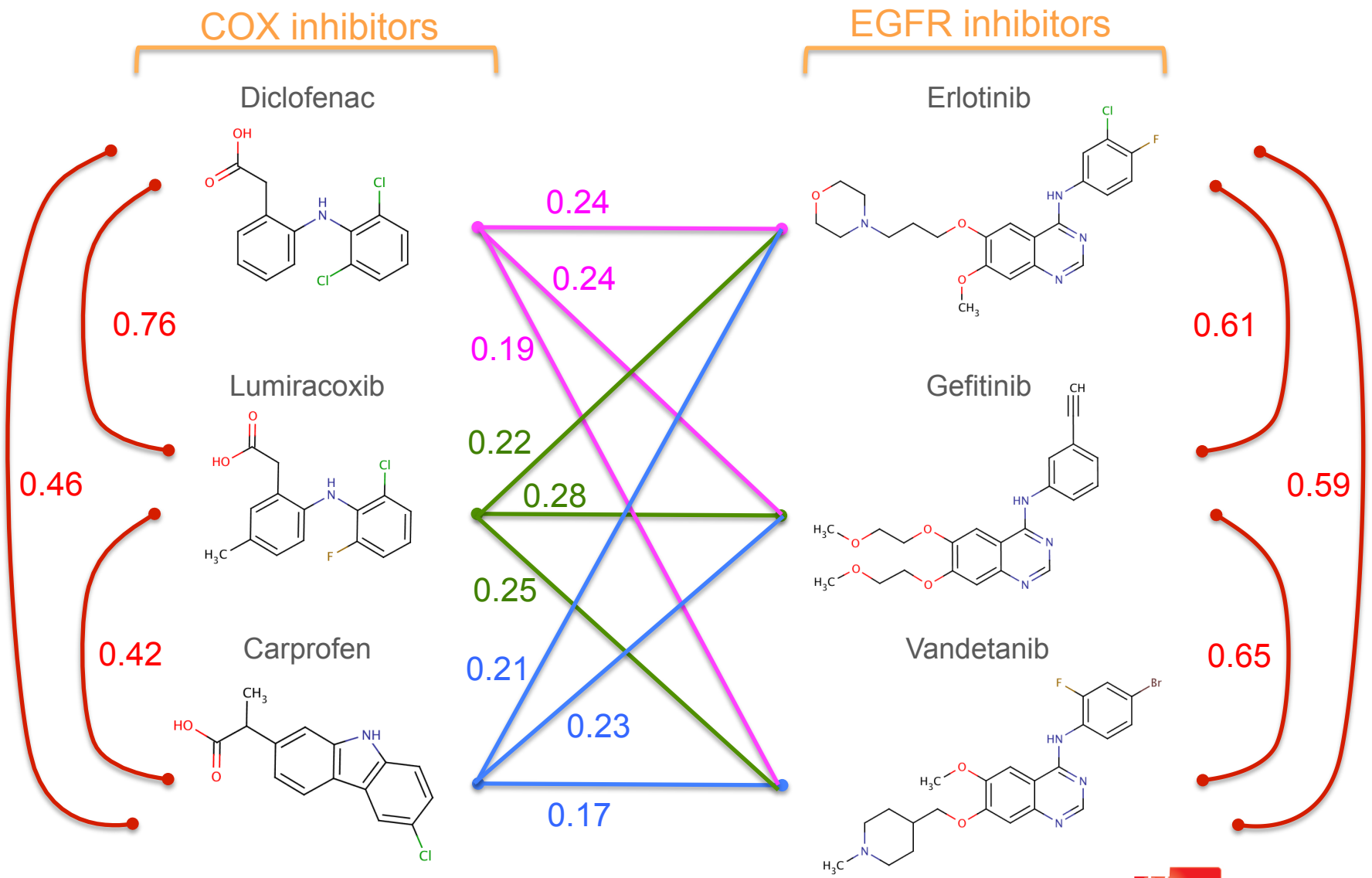


A=(0, 1, 0, 1, 0, 0, 1, 0, 0, …)



B=(0, 1, 0, 1, 0, 0, 1, 0, **1**, …)

Molecular screening – 2D Ligand-based virtual screening



Molecular screening – 2D Ligand-based virtual screening

Chemical similarity (2D fingerprints)

A=(0, 1, 0, 1, 0, 0, 1, 0, 0, …)

B=(0, 1, 0, 1, 0, 0, 1, 0, **1**, …)

The similarity value between molecules A and B is given by the **Tanimoto coefficient** T:

$$T = \frac{c}{a+b+c}$$, where

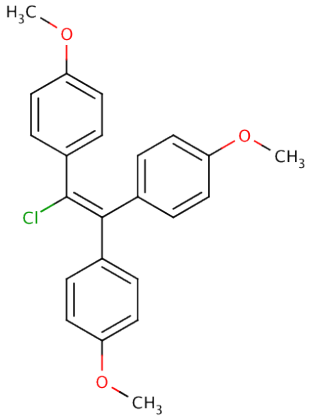
a is the count of bits at 1 in molecule A **but not** in molecule B
b is the count of bits at 1 in molecule B **but not** in molecule A
c is the count of bits at 1 in both molecules A **and** B

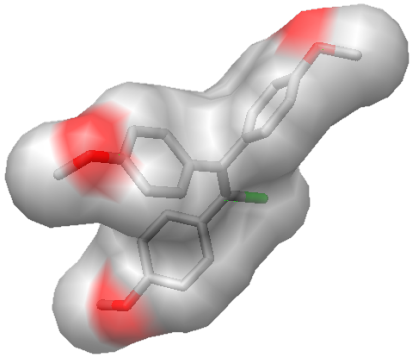
T ranges from 0 for totally different molecules to 1 for identical molecules

Molecular screening – 3D Ligand-based virtual screening

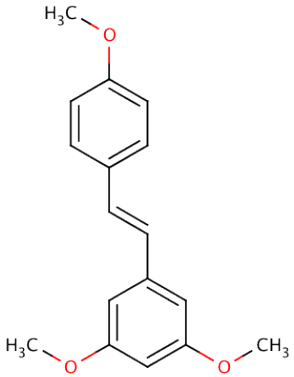
Ex: ROCS

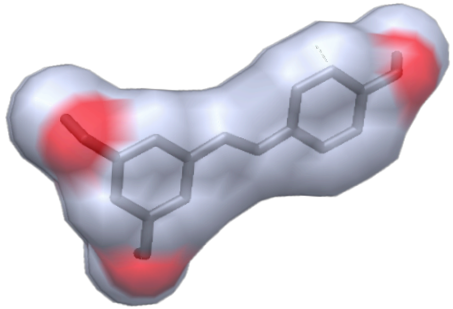
Molecules have similar shape if their volumes overlay well and any volume mismatch is a measure of dissimilarity.
ROCS uses a smooth Gaussian function to represent the molecular volume, so it is possible to routinely minimize to the best global match.





20/40 overlays per sec





Similar 3D shape

Grant, J.A., Gallardo, M.A., Pickup, B., *J. Comp. Chem.*, **1996**, *17*, 1653.

Molecular screening

Hands-on exercise:

Compare molecules using fingerprints, like a computer …

Molecular Modeling Group

Olivier Michielin

Vincent Zoete

Christophe Bovigny
Prasad Chaskar
Michel Cuendet
Antoine Daina
David Gfeller
Dennis Haake
Melita Irving
Justyna Iwaszkiewicz
Maria Johansson
Nahzli Dilek
Somi Reddy Majjigapu
Ute Röhrig
Thierry Schuepbach

All information : **http://www.molecular-modelling.ch**
Complete list of CADD tools : **http://www.click2drug.org**

SIB Swiss Institute of Bioinformatics

Questions, comments, feedbacks : **antoine.daina@isb-sib.ch**
vincent.zoete@isb-sib.ch