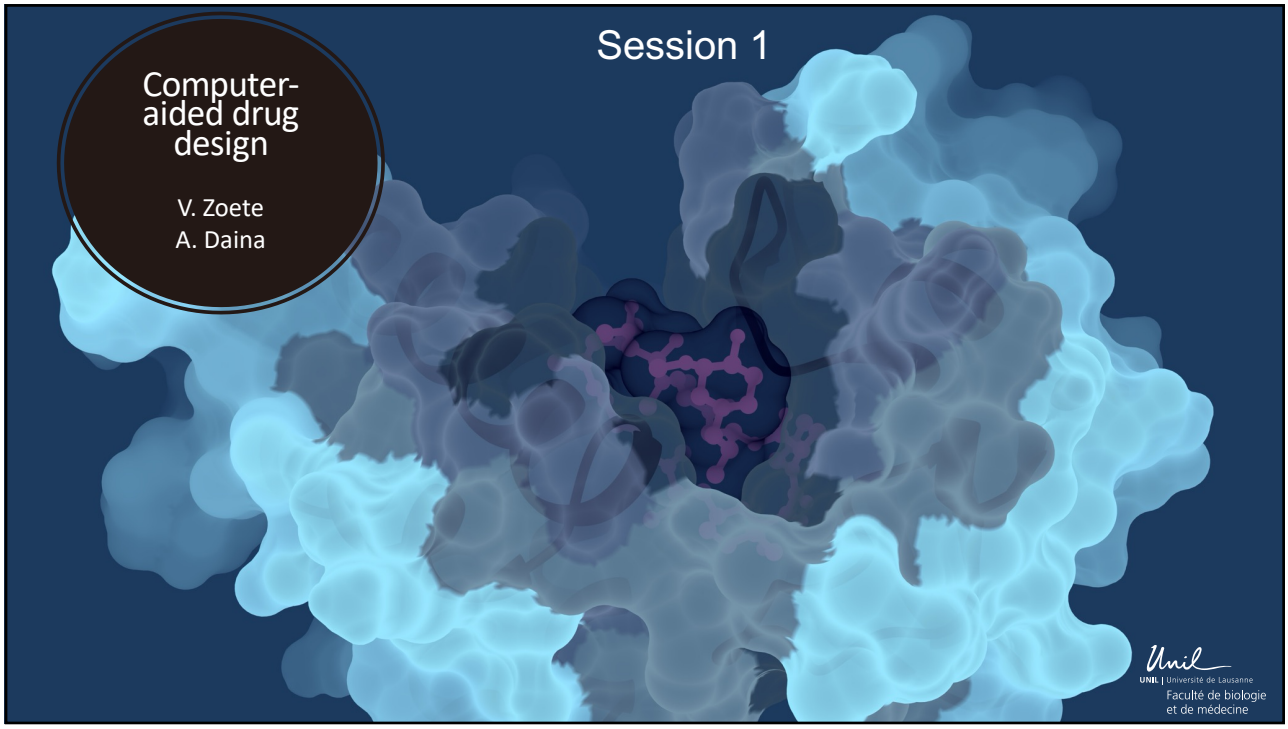



1

## Lecture and Practice Proceedings & Objectives

- Have a flavor of the broadness of the drug design applications,
- Acquire the basic theoretical background,
- Practice the molecular graphics techniques,
- Know the free web-based tools developed at SIB,
- Use them for structure-based and ligand-based design

➔ You should be able to perform simple tasks of computer-aided drug design on whatever computer connected to the internet



2

# Lectures & Practices Agenda

| Session | Lecture | Practice |
|---|---|---|
| 1 | Prologue: molecular representation | |
| | Introduction to (computer-aided) drug design | |
| | Origin of 3D structures | |
| | Molecular recognition | Use of UCSF chimera to analyze protein-ligand complexes |
| 2 | Binding free energy estimation | |
| | Introduction to molecular docking | Ligand-protein docking with AutoDock Vina |
| 3 | Introduction to molecular (virtual) screening | Ligand-based virtual screening with **SwissSimilarity** |
| 4 | Short introduction on target prediction of small molecules | Use of **SwissTargetPrediction** to perform reverse screening. |
| 5 | Introduction to ADME, pharmacokinetics, druglikeness | Estimate physicochemical, pharmacokinetic, druglike and related properties with **SwissADME** |
| 6 | Short introduction to bioisosterism | Use of **SwissBioisostere** to perform bioisosteric design |



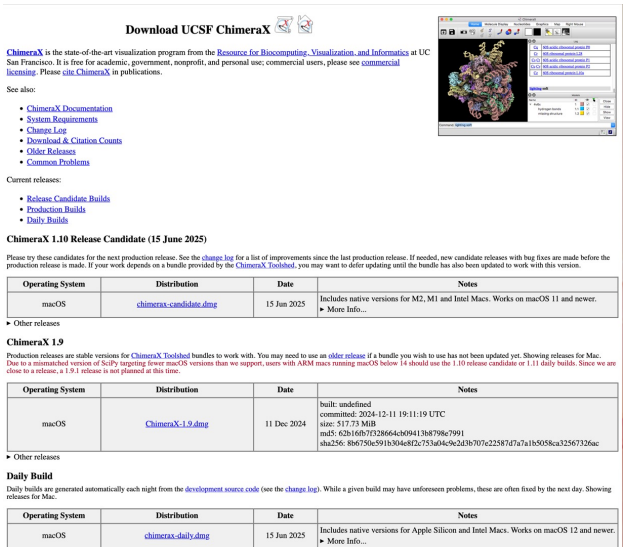
3

# Installing UCSF ChimeraX

In this lecture/practical you will use the software UCSF ChimeraX for 3D structure visualisation and analysis.

This software is:
- free for teaching or academic research
- available for the most current platforms (Windows, Mac, Linux)
- open source (you can modify it for your needs if you know how to code in python. This is out of the scope of this lecture).

You can download the latest production release here: https://www.cgl.ucsf.edu/chimerax/download.html



**Download UCSF ChimeraX**

**ChimeraX** is the state-of-the-art visualization program from the Resource for Biocomputing, Visualization, and Informatics at UC San Francisco. It is free for academic, government, nonprofit, and personal use; commercial users, please see commercial licensing. Please cite ChimeraX in publications.

See also:

- ChimeraX Documentation
- System Requirements
- Change Log
- Download & Citation Counts
- Older Releases
- Common Problems

Current releases:

- Release Candidate Builds
- Production Builds
- Daily Builds

**ChimeraX 1.10 Release Candidate (15 June 2025)**

Please try these candidates for the next production release. See the change log for a list of improvements since the last production release. If needed, new candidate releases with bug fixes are made before the production release is made. If your work depends on a bundle provided by the ChimeraX Toolshed, you may want to defer updating until the bundle has also been updated to work with this version.

| Operating System | Distribution | Date | Notes |
|---|---|---|---|
| macOS | chimerax-candidate.dmg | 15 Jun 2025 | Includes native versions for M2, M1 and Intel Macs. Works on macOS 11 and newer. ▸ More Info... |

▸ Other releases

**ChimeraX 1.9**

Production releases are stable versions for ChimeraX Toolshed bundles to work with. You may need to use an older release if a bundle you wish to use has not been updated yet. Showing releases for Mac. Due to a mismatched version of SciPy targeting fewer macOS versions than we support, users with ARM macs running macOS below 14 should use the 1.10 release candidate or 1.11 daily builds. Since we are close to a release, a 1.9.1 release is not planned at this time.

| Operating System | Distribution | Date | Notes |
|---|---|---|---|
| macOS | ChimeraX-1.9.dmg | 11 Dec 2024 | built: undefined<br>committed: 2024-12-11 19:11:19 UTC<br>size: 517.73 MiB<br>md5: 62b16fb7f328664cb09413b8798e7991<br>sha256: 8b6750e591b304e8f2c753a04c9e2d3b707e22587d7a7a1b5058ca32567326ac |

▸ Other releases

**Daily Build**

Daily builds are generated automatically each night from the development source code (see the change log). While a given build may have unforeseen problems, these are often fixed by the next day. Showing releases for Mac.

| Operating System | Distribution | Date | Notes |
|---|---|---|---|
| macOS | chimerax-daily.dmg | 15 Jun 2025 | Includes native versions for Apple Silicon and Intel Macs. Works on macOS 12 and newer. ▸ More Info... |

Please, install this software on your machine.
It will be mandatory for the practicals, but also useful for the theoretical lectures



4

## The dedicated web site

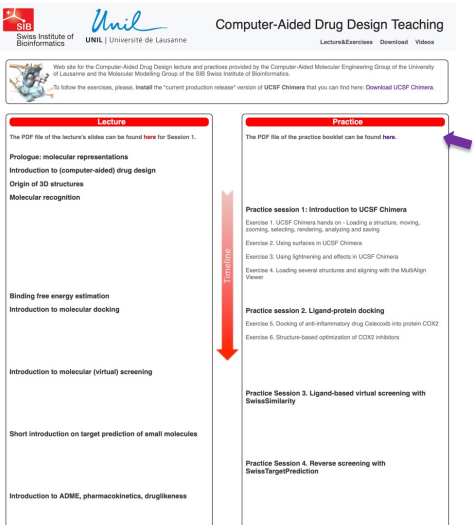
This teaching has been conceived to alternate theoretical lectures and practicals, so that you will:

- experiment yourself the visualisation and analysis of ligand-protein 3D structures
- get a flavor of different tools of computer-aided drug design

To facilitate the process, a web site has been especially conceived for this teaching. You can find it here:

http://www.drug-design-teaching.ch

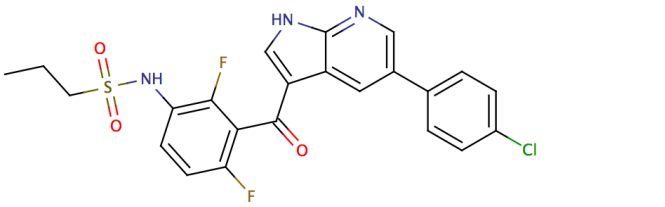
1. This web site will indicate you when to switch between lecture and practicals. For instance, you will be able to make Session 1 exercices just after the lecture on molecular recognition
2. The booklet of the practicals and the PDF of the lecture can be downloaded from the web site too

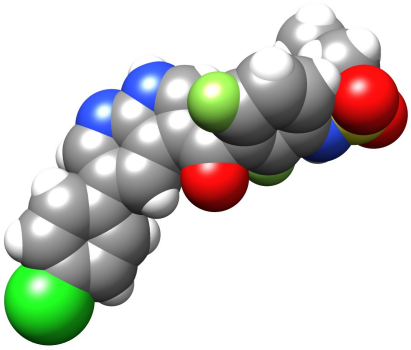


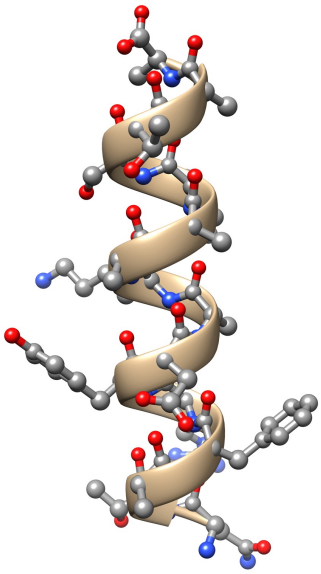



5

## Prologue: molecular representations







6

## 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.)

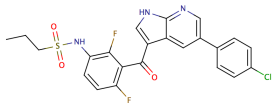
Small molecules can be represented in 1D, 2D or 3D:

Example of Vemurafenib (BRAF V600E inhibitor)

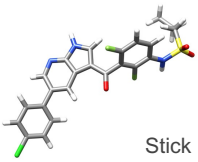
1D

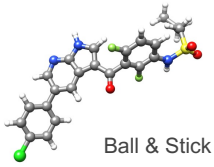
**SMILES**: CCCS(=O)(=O)NC1=C(F)C(C(=O)C2=CNC3=NC=C(C=C23)C2=CC=C(Cl)C=C2)=C(F)C=C1
**InChI**: 1S/C23H18ClF2N3O3S/c1-2-9-33(31,32)29-19-8-7-18(25)20(21(19)26)22(30)17-12-28-23-16(17)10-14(11-27-23)13-3-5-15(24)6-4-13/h3-8,10-12,29H,2,9H2,1H3,(H,27,28)

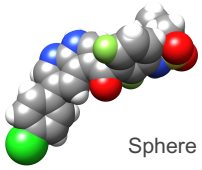
2D

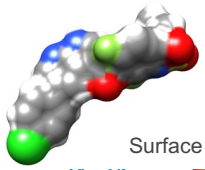


3D


Stick


Ball & Stick


Sphere


Surface



7

## Molecular representations – biological macromolecules

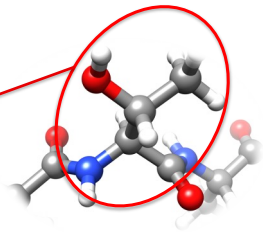
Biological macromolecules can also be represented in 1D, 2D or 3D:

Example of proteins

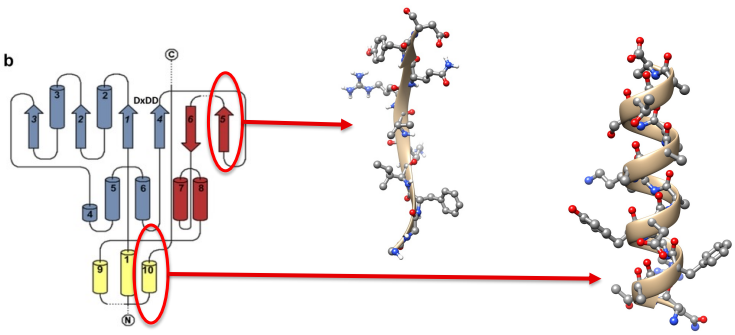
1D

Sequence if amino-acids:

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



2D





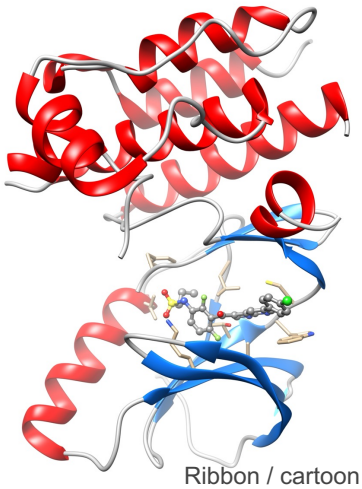

8

## Molecular representations – biological macromolecules

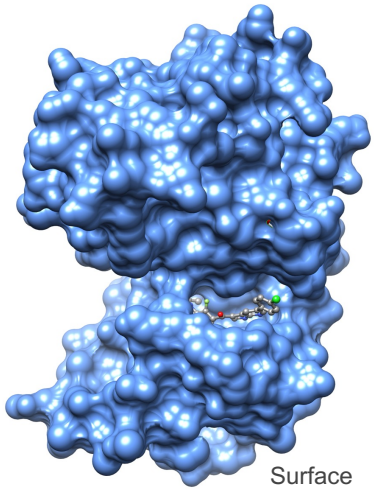
Biological macromolecules can also be represented in 1D, 2D or 3D:

Example of proteins

3D



Ribbon / cartoon



Surface



9

## Molecular representations – biological macromolecules

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

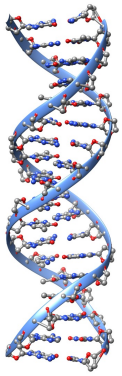
Example of DNA

1D

Sequence if nucleotides:

GAGTAGCACGTGCTACTC

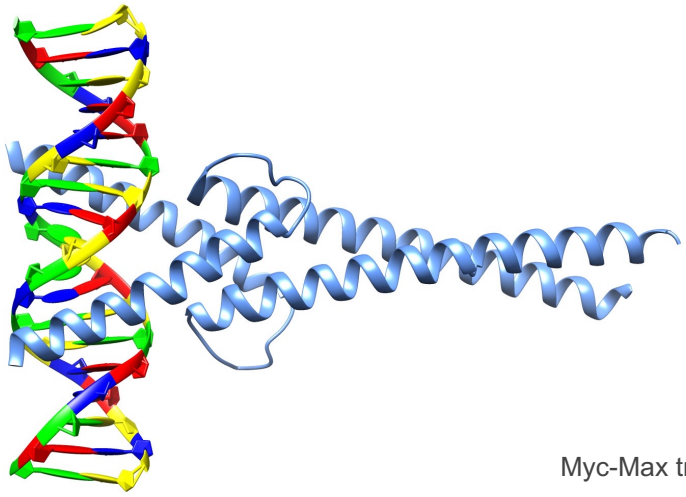
3D





10

## Molecular representations – biological macromolecules

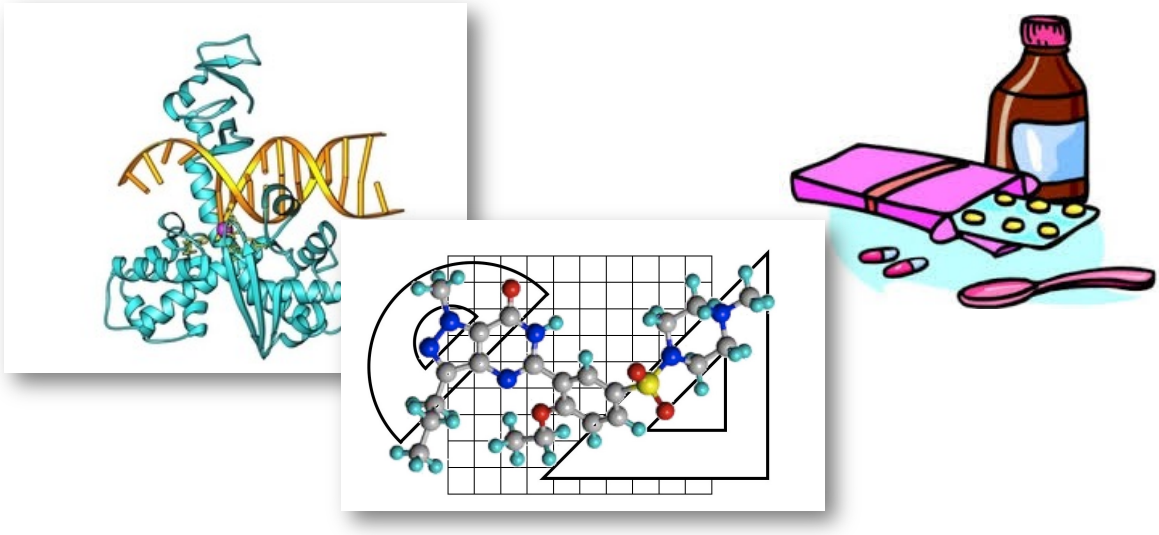


Myc-Max transcription factor



11

## Overview of the Drug Design Pipeline





12

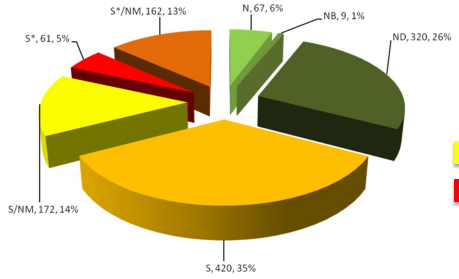
# Drugs: Definition and Origin

**Drug** (here = active ingredient):

- A **substance** administered to a patient with possibly various objectives:
  - 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 ?



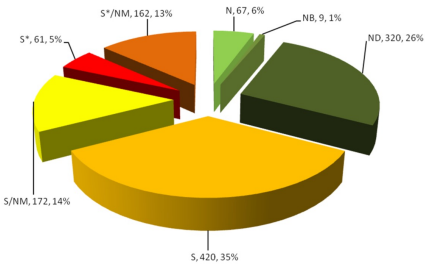

| | |
|---|---|
| N | Unaltered natural product |
| NB | Botanical drug (defined mixture) |
| ND | Natural product derivative |
| S | Synthetic drug |
| S* | Synthetic drug (NP pharmacophore) |
| /NM | Mimic of natural product |

Newman, D. J. & Cragg, G. M. (**2016**). *Journal of Natural Products*, *79*(3), 629.



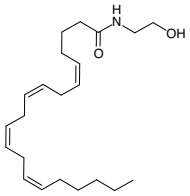
13

# Drugs: Definition and Origin

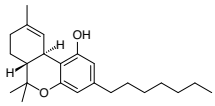



| | |
|---|---|
| N | Unaltered natural product |
| NB | Botanical drug (defined mixture) |
| ND | Natural product derivative |
| S | Synthetic drug |
| S* | Synthetic drug (NP pharmacophore) |
| /NM | Mimic of natural product |

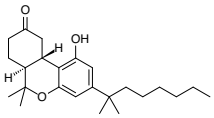
**Anandamide.** Natural, endogenous, ligand of cannabinoid receptors



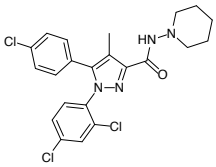
**Tetrahydrocannabinol (THC).** Natural ligand of cannabinoid receptors, from plant. Analgesic, antiemetic



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



**Rimonabant.** Synthetic ligand. Anorectic anti-obesity.

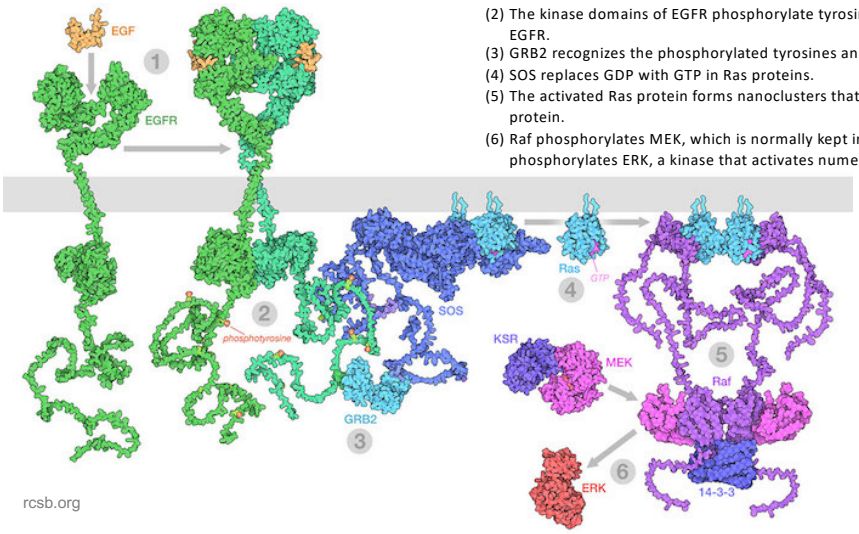




14

## Drug Design: Aim

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



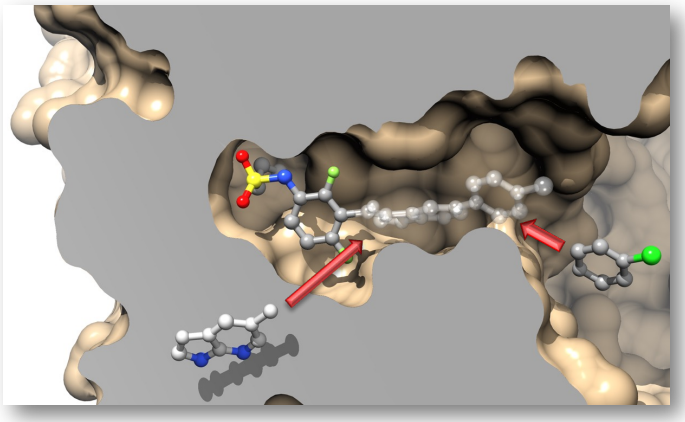

(1) EGF binds to its receptor (EGFR), leading to its dimerization.
(2) The kinase domains of EGFR phosphorylate tyrosines on the unstructured C-terminal portion of EGFR.
(3) GRB2 recognizes the phosphorylated tyrosines and recruits SOS.
(4) SOS replaces GDP with GTP in Ras proteins.
(5) The activated Ras protein forms nanoclusters that activate Raf, with the help of the 14-3-3 protein.
(6) Raf phosphorylates MEK, which is normally kept inactive by complexation with KSR. MEK then phosphorylates ERK, a kinase that activates numerous processes in the cell.



15

## Drug Design: Aim

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



To address:

**Molecular recognition**; i.e. "Lock and key" (E. Fischer)

➡ Potency, Selectivity

But also **ADMET**,

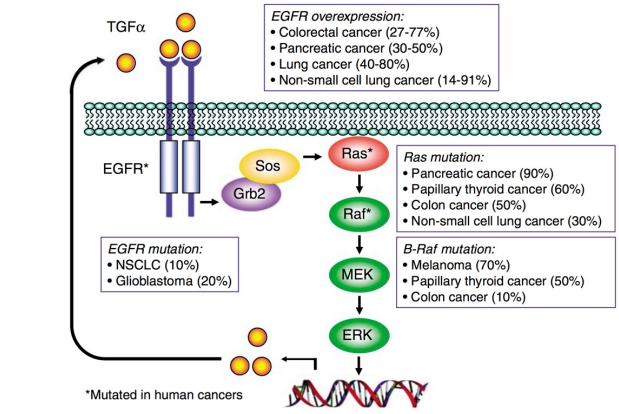
- Absorption
- Distribution
- Metabolism
- Excretion
- Toxicity



16

# Drug Design: Aim

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




Roberts P.J., et al.,Oncogene (2007) 26, 3291–3310

## Possible drugs:

**EGFR:**

Afatinib, Almonertinib, Brigatinib, Dacomitinib, Erlotinib, Gefitinib, Icotinib, Lapatinib, Neratinib, Olmutinib, Osimertinib, Pyrotinib, Simotinib, Sorafenib, Vandetanib

**RAS:**

Adagrasib, Sotorasib

**RAF:**

Dabrafenib, Encoratinib, Vemurafenib

**MEK:**

Binimetinib, Cobimetinib, Selumetinib, Trametinib

**ERK:**

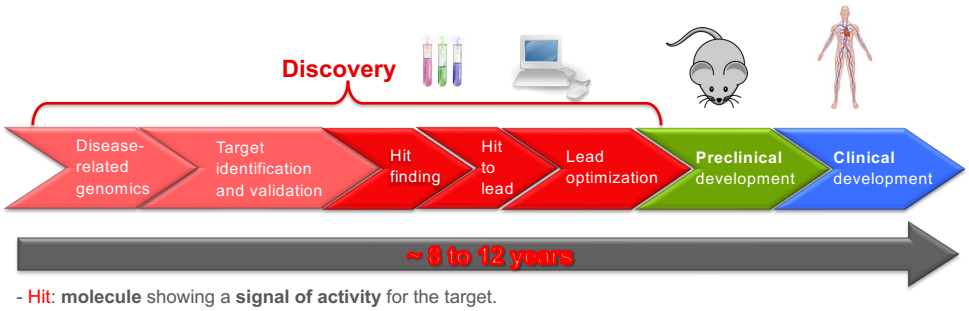
Ulixertinib





# Drug Design: Pipeline

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




- Hit: **molecule** showing a **signal of activity** for the target.
- Hit finding: process to discover hits, generally **using Molecular Screening** (HTS).
- Hit-to-lead: Selection of select hits. **Activity confirmation, re-testing** for dose-response. Filters (toxicity, ...).
- Lead: molecule showing **promising and confirmed properties.**
- Lead optimization: Modest and targeted **chemical modifications** of the lead **to refine** the properties of lead.
- Preclinical development: **animal** pharmacology/toxicology **testing:** reasonably **safe to proceed with human?**
- Clinical development: **safety**, **dosage**, **efficacy side-effects** in human

## Drug design : some figures

**Globally:**

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

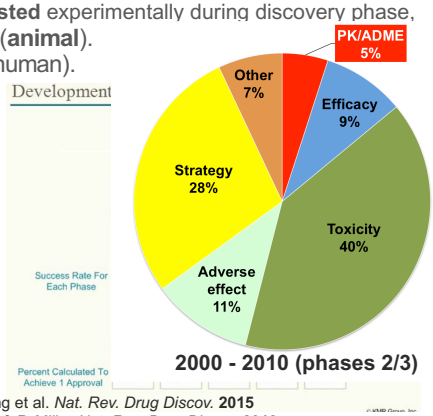
**Typical project:**

- **Millions of chemical structures** ("virtual molecules") created and/or evaluated **in computer**,
- **Thousands of molecules synthesized and tested** experimentally during discovery phase,
- **3 to 10 molecules** tested in preclinical trials (**animal**).
- **1 to 3 molecules** to enter in **clinical trials** (human).

**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

➔ **Risky and expensive.**




M.J. Waring et al. *Nat. Rev. Drug Discov.* **2015**
J. Arrowsmith & P. Miller *Nat. Rev. Drug Discov.* **2013**



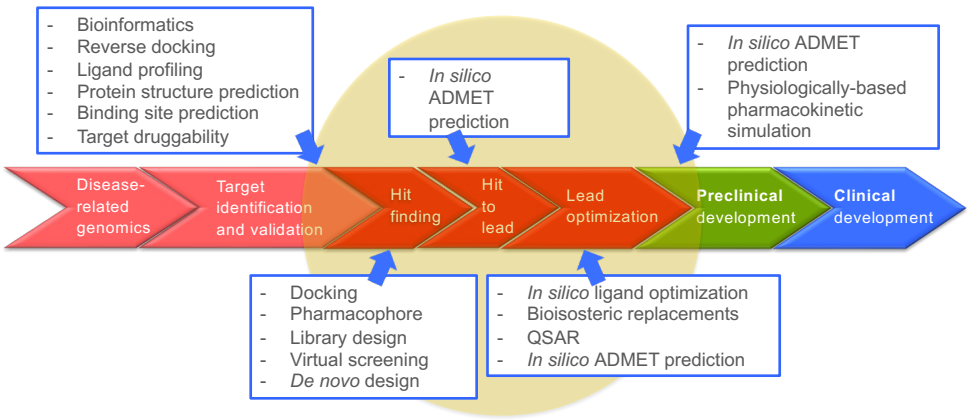


## Computer-Aided Drug design (CADD)

Objective: use of **computing resources**, algorithms and 3D visualization (programs, web-services, databases) to **support**:

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

CADD is including a lot of different approaches, methods, techniques and tools:



# Origin of the 3D structures



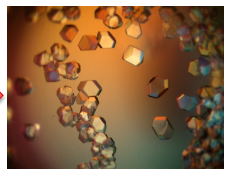


# Experimental methods – Xray crystallography

Purified protein

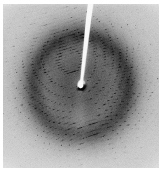
Crystallization

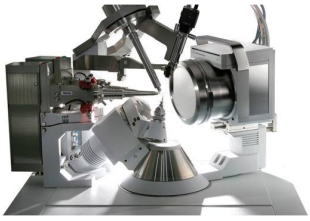
Protein crystals



X-ray

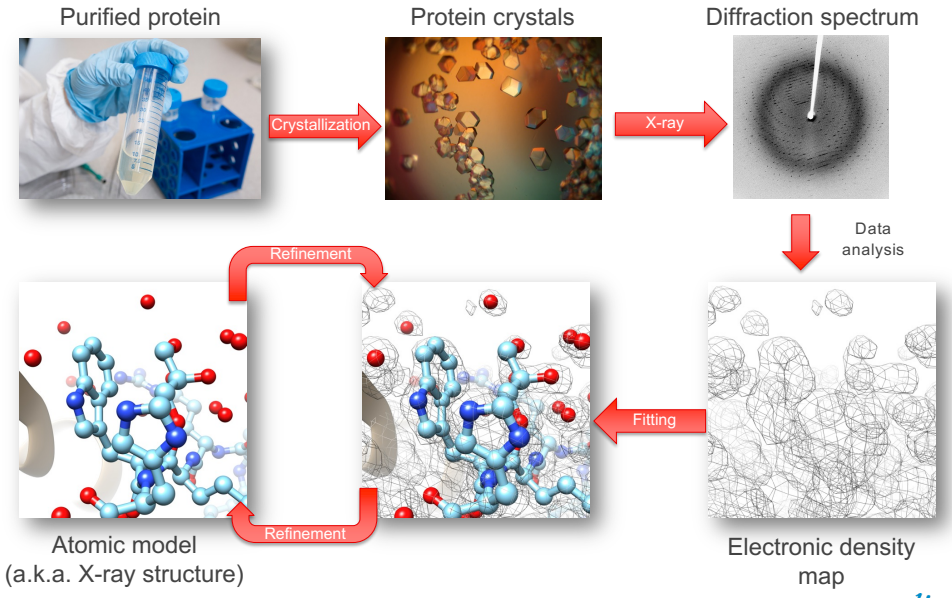
Diffraction spectrum





Xray diffraction

## Experimental methods – Xray crystallography



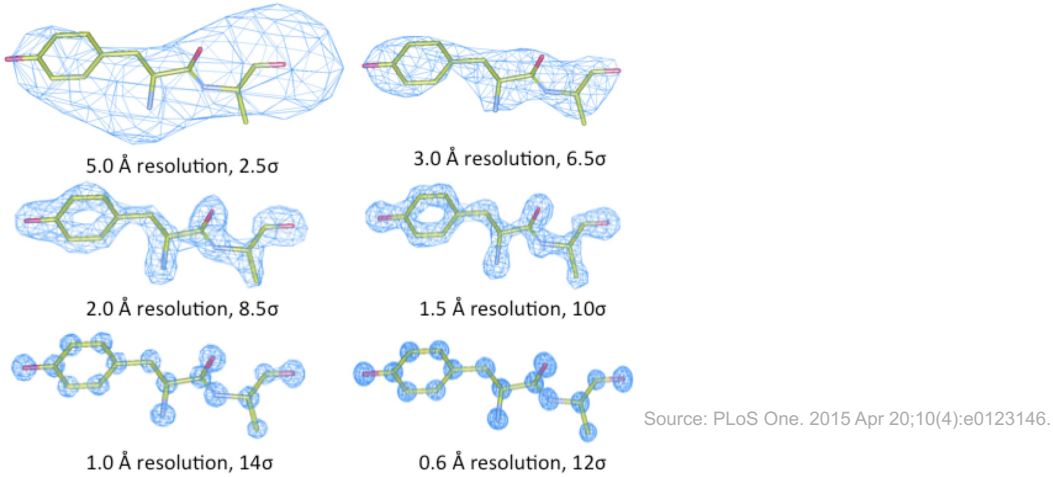



23

## Experimental methods – Xray crystallography

important measures of accuracy:

- **Resolution** (in Å): measures the amount of detail that may be seen in the experimental data. The lower the better (typically around 2 Å)




Source: PLoS One. 2015 Apr 20;10(4):e0123146.



24

## Experimental methods – Xray crystallography

3 important measures of accuracy:

- **Resolution** (in Å): measures the amount of detail that may be seen in the experimental data. The lower the better (typically around 2 Å)
- **R-value**: measures how well the atomic model is supported by the experimental data found in the structure factor file (Perfect fit R-value = 0.0; Random fit R-value = 0.63; Typical R-value ~ 0.20) The atomic model is used to simulate a diffraction spectrum, which is compared to the experimental one.
- **R-free value**: idem than R-value, but calculated for a set of experimental data that have not been used to create the model (~10% of the data are removed before refinement, in order to be used in this test). Generally, R-free value > R-value; Typically R-free value ~ 0.26 for a good quality structure.

Typical limitations:

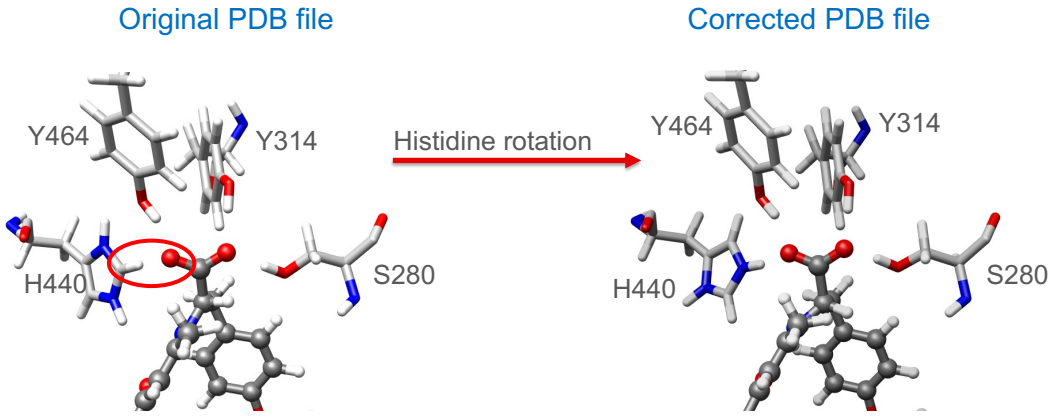
- Hydrogen atoms are generally not visible
- Some regions are not defined (e.g. flexible loops or flexible side chains)
- X-ray structures are models. They can be totally wrong!!



25

## Experimental methods – Xray crystallography

Xray structures are models. They can be wrong!




hPPARα in complex with GW409544 (1K7L in PDB)

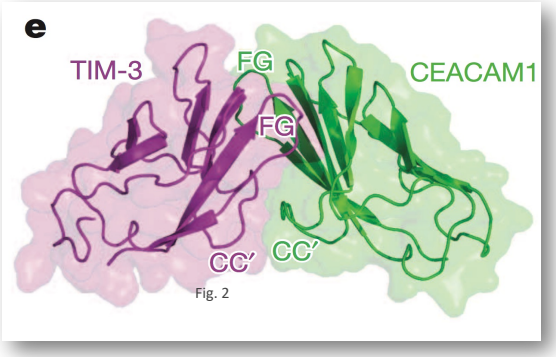


26

## Experimental methods – Xray crystallography

Xray structures are models. They can be totally wrong!

Huang, Y.-H., et al. CEACAM1 regulates TIM-3-mediated tolerance and exhaustion. *Nature,* 2015, *517*(7534), 386–390.




Xray structure of the complex
CEACAM1/TIM3
PDB ID: 4QYC
Resolution: 3.4Å
R-value: 0.232

Correction

**5DZL**

**Crystal structure of the protein human CEACAM1**

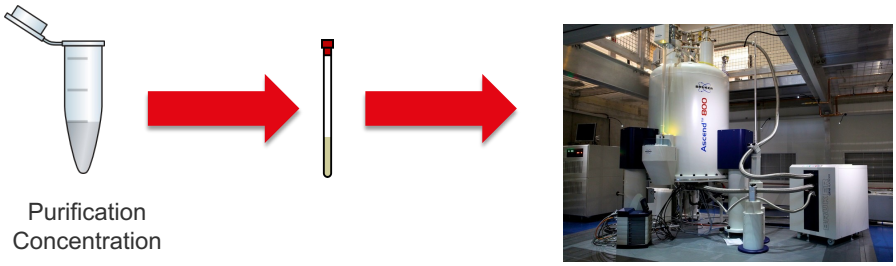
**DOI:** 10.2210/pdb5dzl/pdb Entry 5DZL **supersedes** 4QYC

It was a homodimer of CEACAM1…!



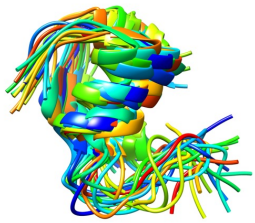


## Experimental methods – NMR spectroscopy



Purification
Concentration

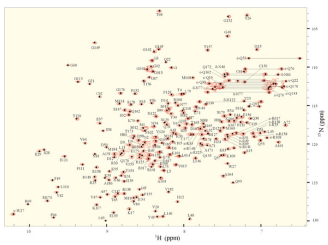




Ensemble of structures
Example: insulin

Distance constraints

Modeling

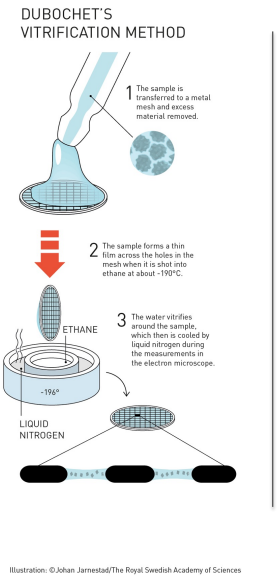


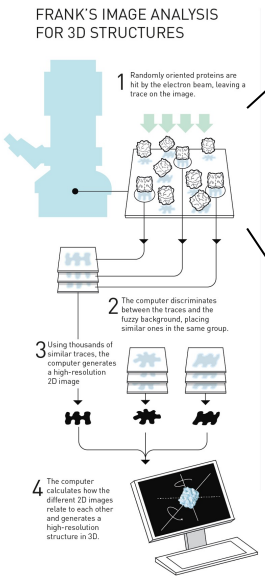
2D-Spectra

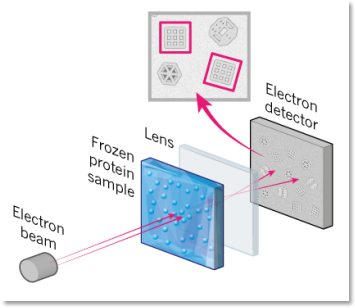
Pros : Structure in solution

Cons:
- Limited to small proteins
- Low resolution
- Highly flexible regions

## Experimental methods – CryoEM



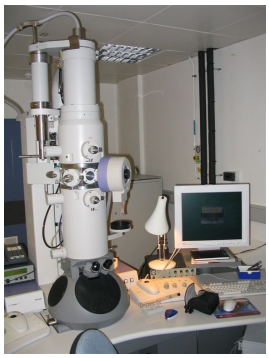



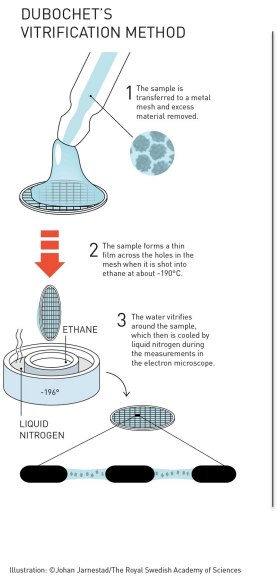
Illustration: ©Johan Jarnestad/The Royal Swedish Academy of Sciences

- Very power electronic beam
- Better resolution than light (smaller wave length)
- In vacuo in the microscope
- Frozen sample (77 K or 4 K)
- Vitrified water



29

## Experimental methods – CryoEM



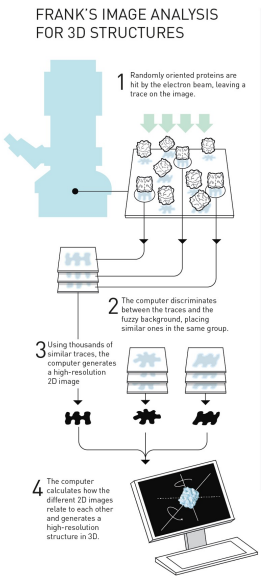




Illustration: ©Johan Jarnestad/The Royal Swedish Academy of Sciences

Until recently:

- Only low resolution structures. Need to be used together with Xray crystallography or NMR (for example, insertion of Xray structures into the Cryo-EM density map)
- Limited to large-size systems (which can actually be seen as a pros or a cons)

Nowadays:

- Resolution close to that of Xray crystallography
- Applicable to smaller systems
- More Cryo-EM structures produced every year than NMR structures
- Capture structures in relevant states (isolated molecules, in solution, at a given salt concentration and pH)



30

## Experimental methods - Summary

| Technique | Advantages | Disadvantages |
|---|---|---|
| Xray crystallography | High resolution (1 to 3 Å ) | Requires to crystallize the protein<br>Does not allow studying transmembrane or very flexible proteins |
| NMR | Does not require protein crystallization<br>~ High resolution | Generally limited to small proteins |
| Cryo-EM | Does not necessitate to crystallize the protein: possible to study transmembrane proteins, and more flexible proteins than Xray.<br>New techniques allow studying smaller proteins, and increasing resolution | Generally limited to large proteins<br>Low resolution, 4 to 20 Å (a lot of progresses have been done recently) |





## Where to find experimental 3D structures? The protein databank

Public experimental 3D structures are stored in the **Protein Data Bank (PDB)**

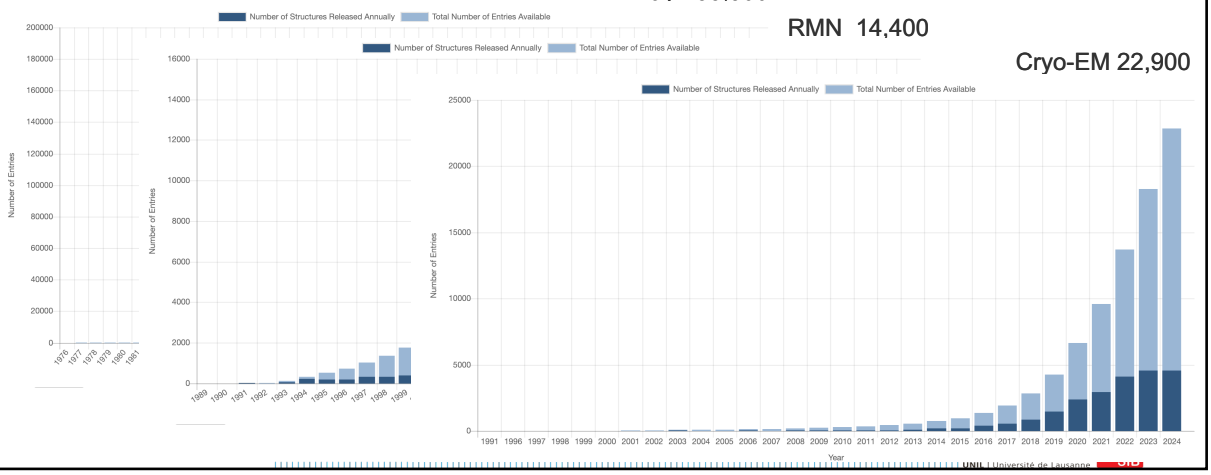
**Worldwide Protein Data Bank (wwPDB)** https://www.wwpdb.org
- RCSB Protein Data Bank (RCSB PDB) https://www.rcsb.org
- Protein Data Bank in Europe (PDBe) https://www.ebi.ac.uk/pdbe
- Protein Data Bank Japan (PDBj) https://pdbj.org

226'000 structures in Oct 2024

X-Rav 188.300

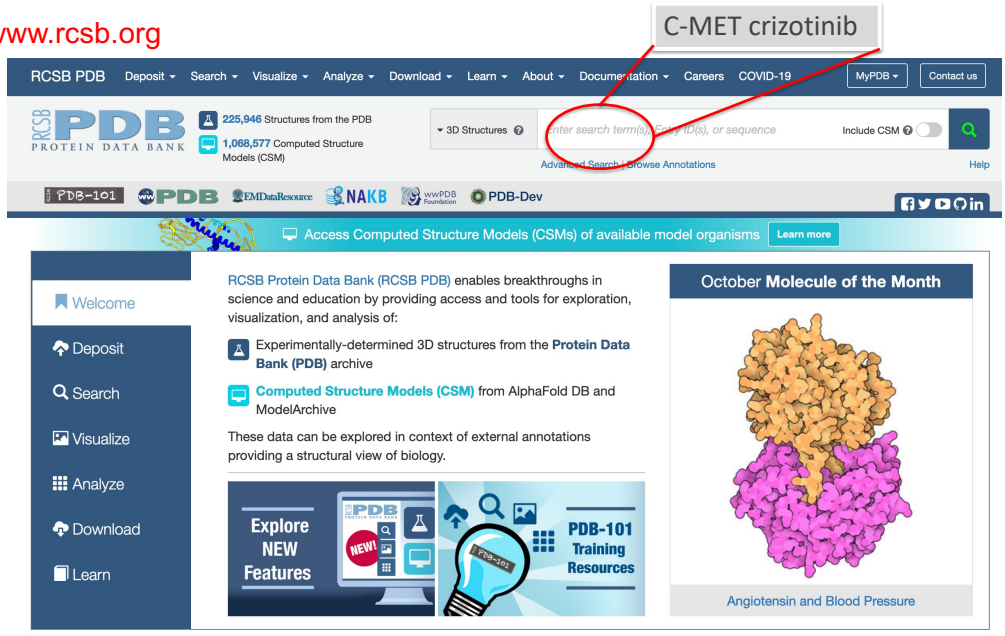
RMN 14,400

Cryo-EM 22,900



## Where to find experimental 3D structures? The protein databank

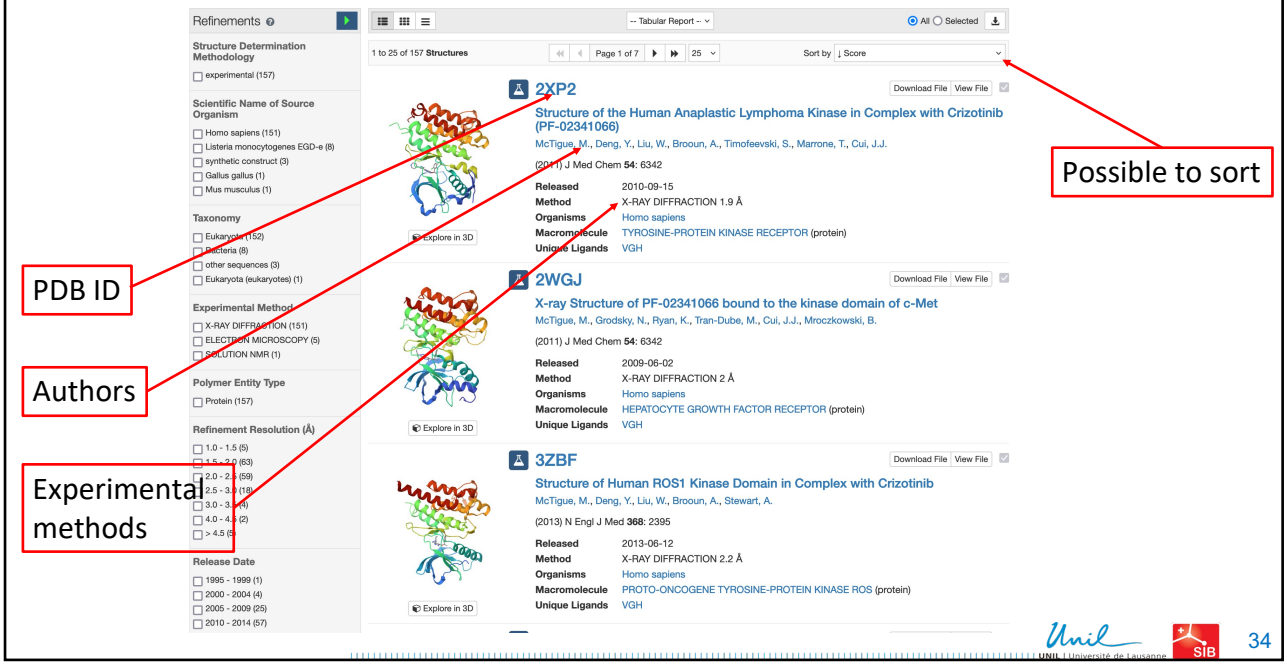
https://www.rcsb.org






33

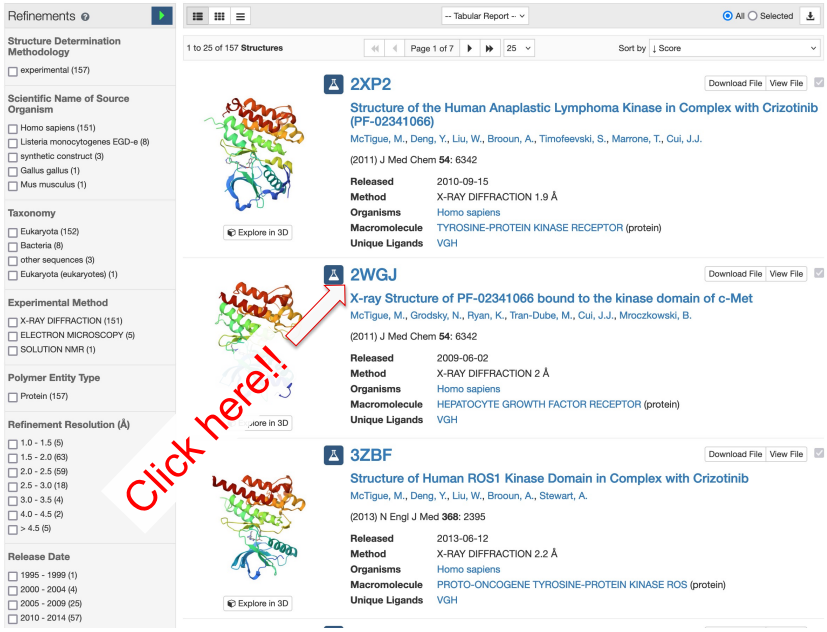
## Where to find experimental 3D structures? The protein databank






34

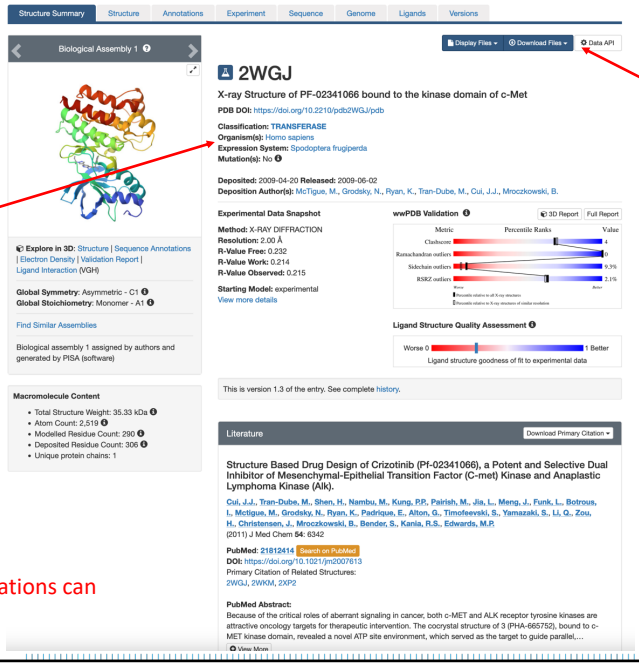
## Where to find experimental 3D structures? The protein databank




35

35

## Where to find experimental 3D structures? The protein databank




Organism (origin of the sequence) & expression system (synthesis et post-translational modifications)

Dowload or online visualization

Experimental method and quality

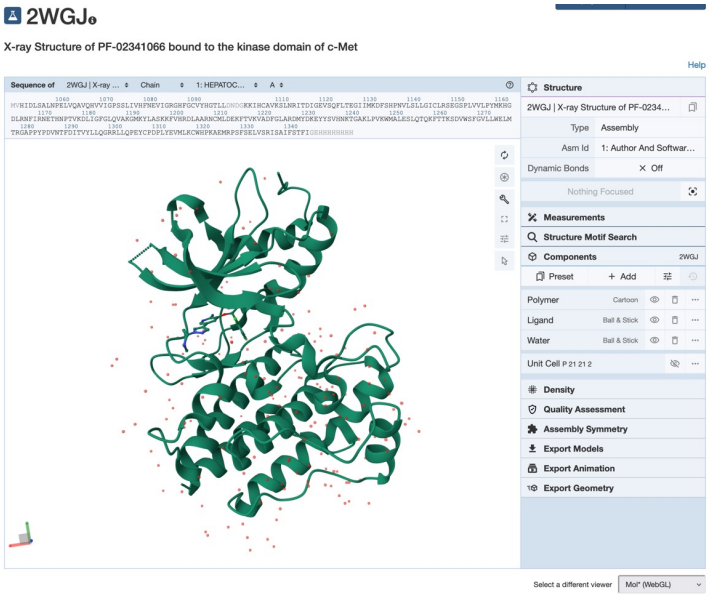
Note: post-translational modifications can differ between organisms

36

36

## Where to find experimental 3D structures? The protein databank

Online visualization

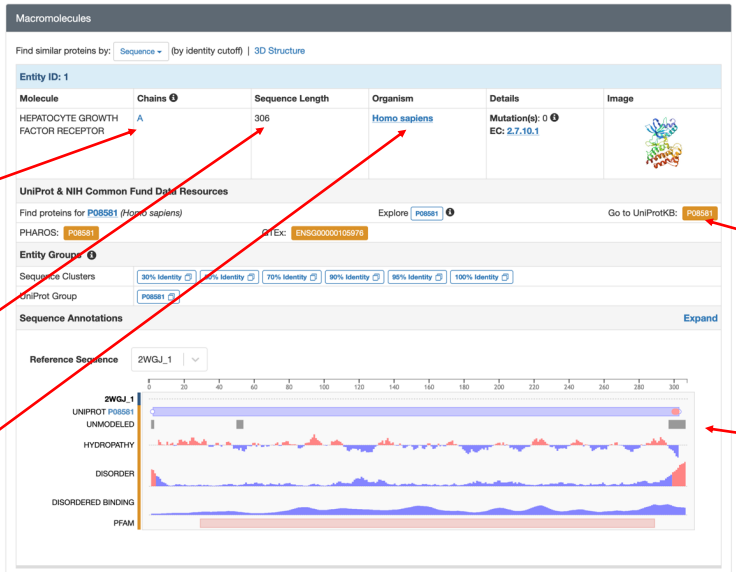





37

## Where to find experimental 3D structures? The protein databank

Information regarding the protein, and what is present in the experimental structure




ID of the protein chain

Number of residues in the protein chain

Source organism

Link to Uniprot for this protein

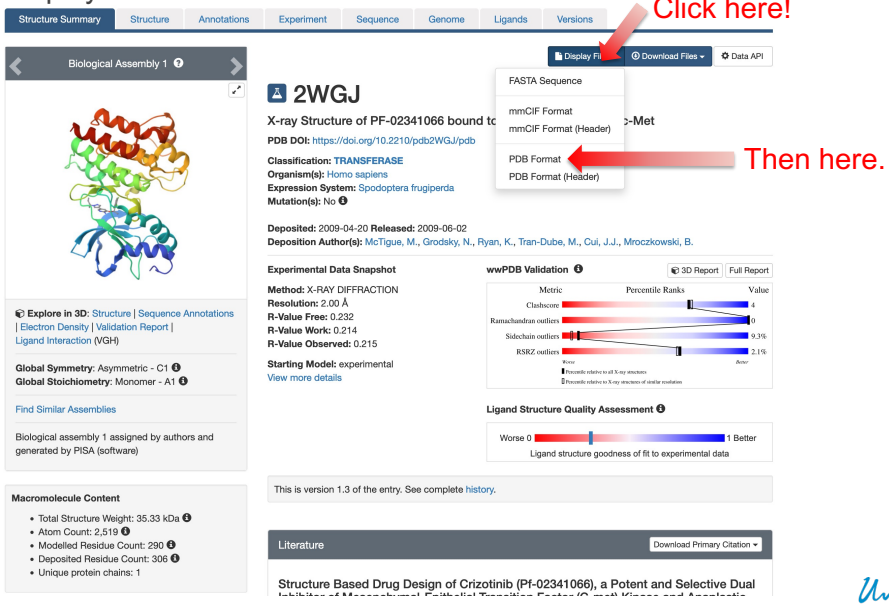
Information on sequence, mutations, and missing regions



38

## Where to find experimental 3D structures? The protein databank

Download / display








## Where to find experimental 3D structures? The protein databank

Header with information about the protein and experimental conditions

```
HEADER    TRANSFERASE                             20-APR-09   2WGJ
TITLE     X-RAY STRUCTURE OF PF-02341066 BOUND TO THE KINASE DOMAIN OF C-MET
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: HEPATOCYTE GROWTH FACTOR RECEPTOR;
COMPND   3 CHAIN: A;
COMPND   4 FRAGMENT: TYROSINE KINASE DOMAIN, RESIDUES 1051-1348;
COMPND   5 SYNONYM: HGF RECEPTOR, SCATTER FACTOR RECEPTOR, SF RECEPTOR, HGF/SF
COMPND   6 RECEPTOR, MET PROTO-ONCOGENE TYROSINE KINASE, C-MET;
COMPND   7 EC: 2.7.10.1;
COMPND   8 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   3 ORGANISM_COMMON: HUMAN;
SOURCE   4 ORGANISM_TAXID: 9606;
SOURCE   5 EXPRESSION_SYSTEM: SPODOPTERA FRUGIPERDA;
SOURCE   6 EXPRESSION_SYSTEM_TAXID: 7108;
SOURCE   7 EXPRESSION_SYSTEM_CELL_LINE: SF9;
SOURCE   8 EXPRESSION_SYSTEM_VECTOR_TYPE: BACULOVIRUS;
SOURCE   9 EXPRESSION_SYSTEM_PLASMID: PFASTBAC1
KEYWDS    C-MET, KINASE, INHIBITOR, TRANSFERASE, ATP-BINDING, NUCLEOTIDE-
KEYWDS   2 BINDING, TYROSINE-PROTEIN KINASE
EXPDTA    X-RAY DIFFRACTION
AUTHOR    M.MCTIGUE,N.GRODSKY,K.RYAN,M.TRAN-DUBE,J.J.CUI,B.MROCZKOWSKI
REVDAT   5  13-DEC-23 2WGJ    1       REMARK
REVDAT   4  08-MAY-19 2WGJ    1       REMARK
REVDAT   3  28-SEP-11 2WGJ    1       AUTHOR JRNL   REMARK FORMUL
REVDAT   3 2                  1       VERSN
REVDAT   2  01-SEP-10 2WGJ    1       COMPND KEYWDS JRNL   REMARK
REVDAT   2 2                  1       HETSYN MASTER
REVDAT   1  02-JUN-09 2WGJ    0
JRNL      AUTH  J.J.CUI,M.TRAN-DUBE,H.SHEN,M.NAMBU,P.P.KUNG,M.PAIRISH,L.JIA,
JRNL      AUTH 2 J.MENG,L.FUNK,I.BOTROUS,M.MCTIGUE,N.GRODSKY,K.RYAN,
JRNL      AUTH 3 E.PADRIQUE,G.ALTON,S.TIMOFEEVSKI,S.YAMAZAKI,Q.LI,H.ZOU,
JRNL      AUTH 4 J.CHRISTENSEN,B.MROCZKOWSKI,S.BENDER,R.S.KANIA,M.P.EDWARDS
JRNL      TITL  STRUCTURE BASED DRUG DESIGN OF CRIZOTINIB (PF-02341066), A
JRNL      TITL 2 POTENT AND SELECTIVE DUAL INHIBITOR OF
JRNL      TITL 3 MESENCHYMAL-EPITHELIAL TRANSITION FACTOR (C-MET) KINASE AND
JRNL      TITL 4 ANAPLASTIC LYMPHOMA KINASE (ALK).
JRNL      REF   J.MED.CHEM              V.  54  6342 2011
JRNL      REFN                  ISSN 0022-2623
JRNL      PMID  21812414
JRNL      DOI   10.1021/JM2007613
REMARK   2
REMARK   2 RESOLUTION.    2.00 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : REFMAC 5.1.24
REMARK   3   AUTHORS     : MURSHUDOV,SKUBAK,LEBEDEV,PANNU,STEINER,
REMARK   3                : NICHOLLS,WINN,LONG,VAGIN
REMARK   3
REMARK   3    REFINEMENT TARGET : MAXIMUM LIKELIHOOD
REMARK   3
REMARK   3 DATA USED IN REFINEMENT.
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) : 2.00
REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) : 19.96
REMARK   3   DATA CUTOFF            (SIGMA(F)) : NULL
REMARK   3   COMPLETENESS FOR RANGE        (%) : 78.1
REMARK   3   NUMBER OF REFLECTIONS             : 17195
```

Cartesian coordinates for each visible atom

Atom number

Atom

Residue

Pept. Chain code

Residue number

```
ATOM      1  N   HIS A1052       8.906 104.713   7.487  1.00104.39           N
ATOM      2  CA  HIS A1052       8.577 105.689   8.566  1.00104.36           C
ATOM      3  C   HIS A1052       7.812 105.015   9.708  1.00104.17           C
ATOM      4  O   HIS A1052       6.636 104.667   9.557  1.00104.23           O
ATOM      5  CB  HIS A1052       7.775 106.857   7.991  1.00104.48           C
ATOM      6  CG  HIS A1052       8.434 108.189   8.169  1.00104.79           C
ATOM      7  ND1 HIS A1052       8.502 108.828   9.389  1.00104.91           N
ATOM      8  CD2 HIS A1052       9.048 109.006   7.281  1.00105.09           C
ATOM      9  CE1 HIS A1052       9.135 109.978   9.246  1.00105.27           C
ATOM     10  NE2 HIS A1052       9.476 110.111   7.976  1.00105.36           N
ATOM     11  N   ILE A1053       8.492 104.829  10.842  1.00103.87           N
ATOM     12  CA  ILE A1053       7.946 104.104  11.996  1.00103.48           C
ATOM     13  C   ILE A1053       8.315 104.787  13.324  1.00103.11           C
ATOM     14  O   ILE A1053       9.473 105.159  13.539  1.00103.17           O
ATOM     15  CB  ILE A1053       8.402 102.606  11.956  1.00103.56           C
ATOM     16  CG1 ILE A1053       7.415 101.766  11.136  1.00103.74           C
ATOM     17  CG2 ILE A1053       8.557 102.007  13.355  1.00103.48           C
ATOM     18  CD1 ILE A1053       7.832 101.545   9.688  1.00104.18           C
ATOM     19  N   ASP A1054       7.320 104.941  14.201  1.00102.54           N
ATOM     20  CA  ASP A1054       7.461 105.666  15.473  1.00101.93           C
ATOM     21  C   ASP A1054       7.917 104.789  16.644  1.00101.29           C
ATOM     22  O   ASP A1054       7.646 103.580  16.671  1.00101.35           O
ATOM     23  CB  ASP A1054       6.151 106.379  15.846  1.00102.11           C
ATOM     24  CG  ASP A1054       5.023 106.096  14.864  1.00102.49           C
ATOM     25  OD1 ASP A1054       4.487 104.964  14.873  1.00102.82           O
ATOM     26  OD2 ASP A1054       4.609 106.948  14.046  1.00102.86           O
ATOM     27  N   LEU A1055       8.596 105.414  17.612  1.00100.34           N
ATOM     28  CA  LEU A1055       9.150 104.707  18.771  1.00 99.31           C
ATOM     29  C   LEU A1055       8.188 104.654  19.956  1.00 98.44           C
ATOM     30  O   LEU A1055       8.051 103.615  20.603  1.00 98.42           O
ATOM     31  CB  LEU A1055      10.476 105.346  19.212  1.00 99.43           C
ATOM     32  CG  LEU A1055      11.784 104.536  19.295  1.00 99.51           C
ATOM     33  CD1 LEU A1055      11.580 103.125  19.847  1.00 99.54           C
ATOM     34  CD2 LEU A1055      12.508 104.491  17.946  1.00 99.74           C
ATOM     35  N   SER A1056       7.528 105.777  20.233  1.00 97.33           N
ATOM     36  CA  SER A1056       6.629 105.901  21.383  1.00 96.17           C
ATOM     37  C   SER A1056       5.287 105.184  21.194  1.00 95.28           C
ATOM     38  O   SER A1056       4.480 105.111  22.127  1.00 95.16           O
ATOM     39  CB  SER A1056       6.398 107.381  21.725  1.00 96.31           C
ATOM     40  OG  SER A1056       5.676 108.049  20.703  1.00 96.12           O
ATOM     41  N   ALA A1057       5.062 104.653  19.993  1.00 94.08           N
```

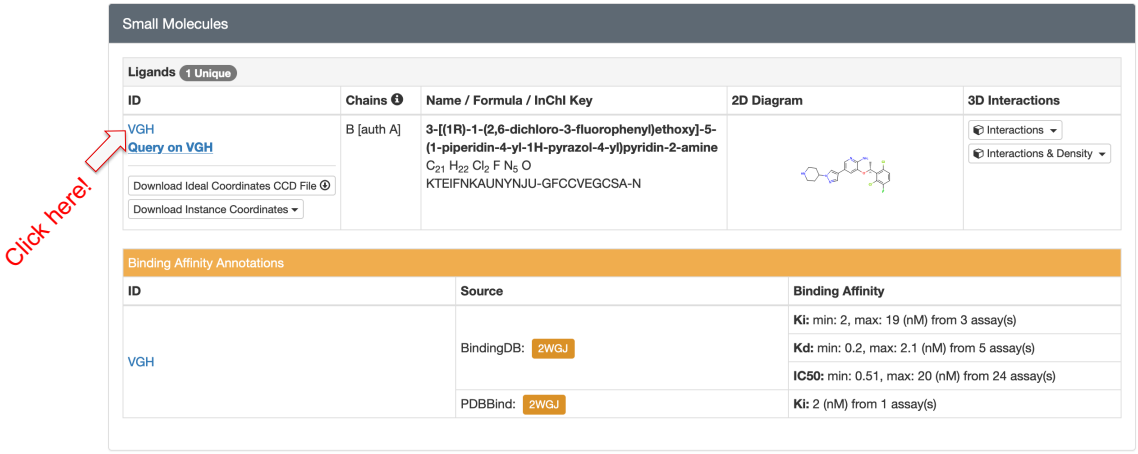
B-factor. Measures the variability of the atom's position. The higher the value, the more fluctuating the position of this atom

Need visualization software...

Ex.: Swiss PDB Viewer, UCSF ChimeraX, Pymol

# Where to find experimental 3D structures? The protein databank

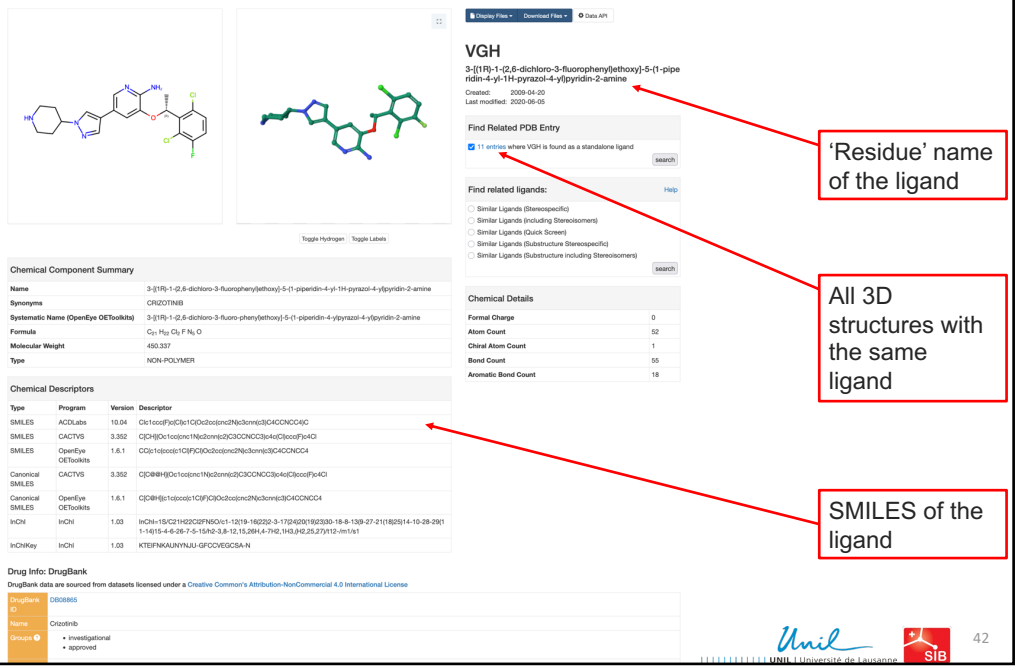
And for small molecules?








# Where to find experimental 3D structures? The protein databank



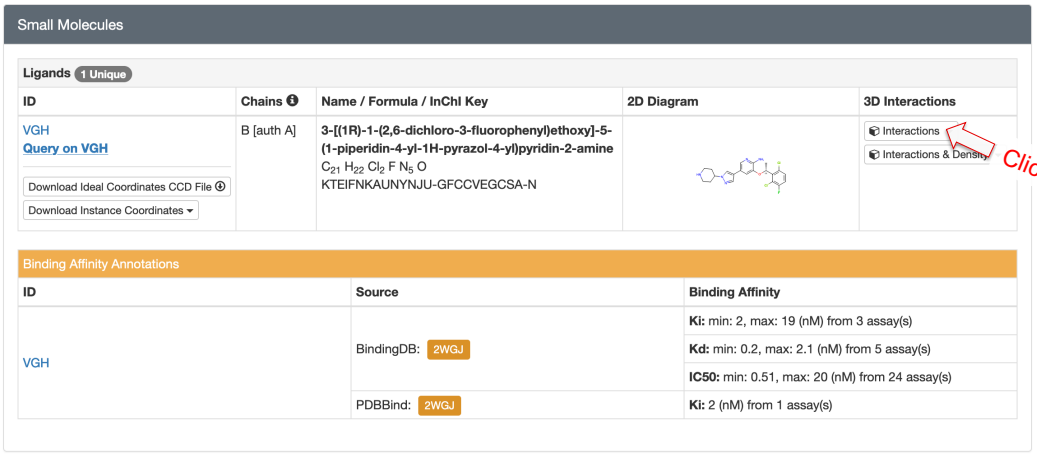
## Where to find experimental 3D structures? The protein databank

List of 3D structures, present in the PDB, and containing the ligand crizotinib




43

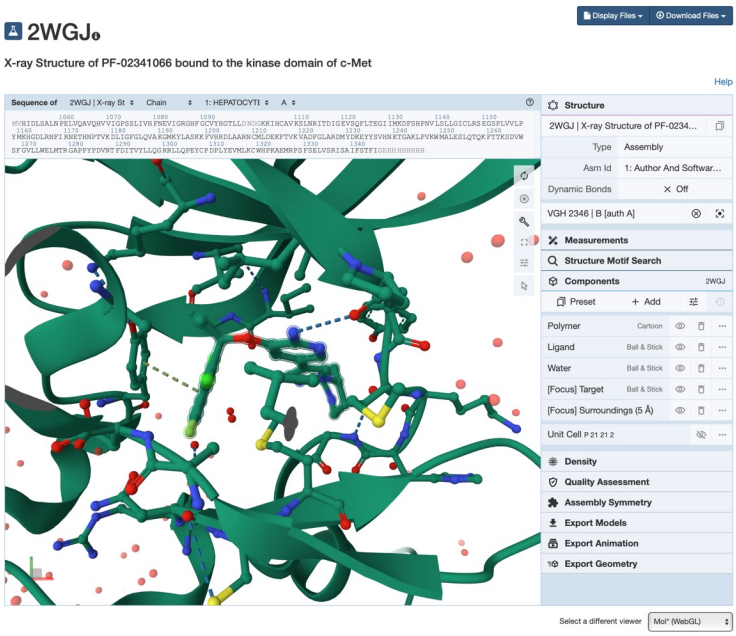
## Where to find experimental 3D structures? The protein databank






44

## Where to find experimental 3D structures? The protein databank

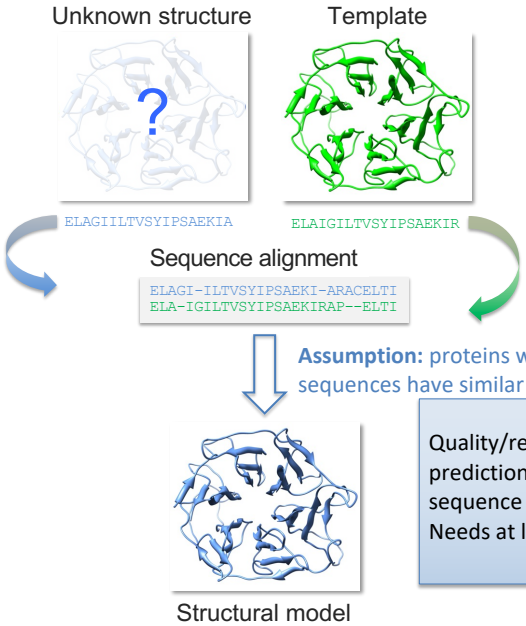



Analysis of ligand/protein interactions in 3D, in the web interface



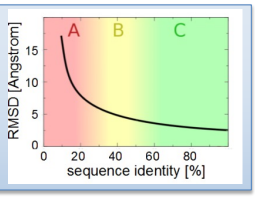


## And when there is no experimental structure? Homology modeling

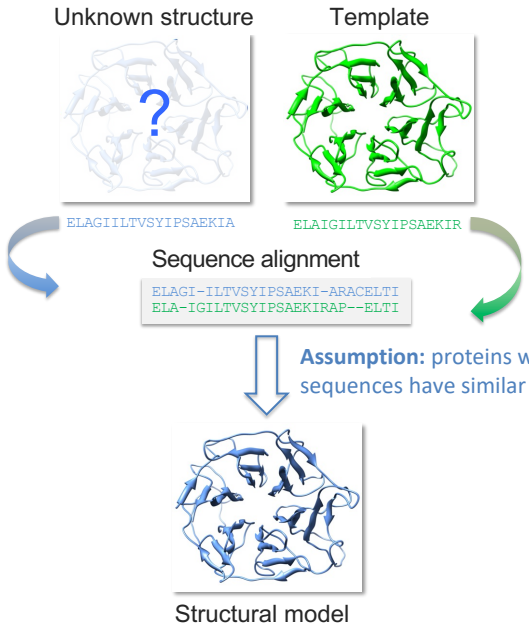



Homology modeling:
Building a theoretical model of the 3D structure of a protein based on experimentally known 3D structure of related proteins.

Quality/relevance of the prediction depends on the sequence identity.
Needs at least 40% seq. id.



## And when there is no experimental structure? Homology modeling




Homology modeling:
Building a theoretical model of the 3D structure of a protein based on experimentally known 3D structure of related proteins.

Programs et web servers:
- Modeller
- I-Tasser
- Robetta
- HHPred
- …

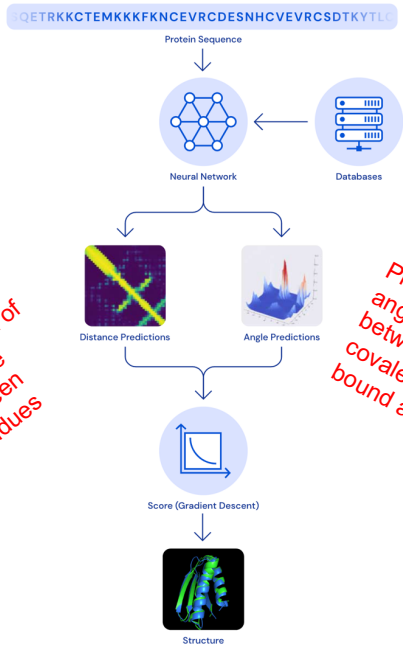
Databases of structural models:
- Swiss-model
- Modbase
- …





## And when there is no experimental structure? Homology modeling

AlphaFold:




Prediction of distance between residues

Prediction of angles between covalently bound atoms

Predicting distance between amino acids:

Sequence of target protein

Sequence alignment with all related proteins

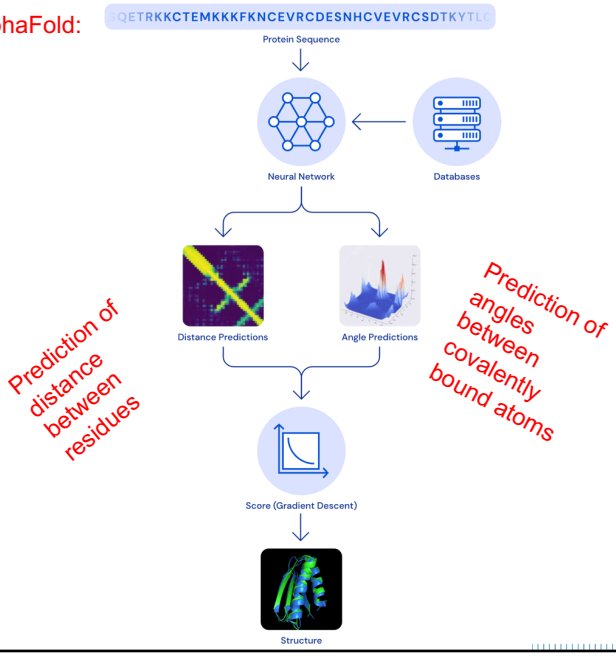
```
C T S Y P I K L M D F E R T S W Q A P R I M T G H K
C S S Y P I K L M D W E R T S W Q A P R I C T G Y K
C Q S Y P L K L M D F E R T S W Q V P R I P T G H K
C N S Y P L K L M D C E R T S W Q V P R I D T G C K
C S S Y P I K L M D F E R T S W Q A P R I F T G H K
C D S Y P V K L M D F E R T S W Q L P R I G T G H K
C C S Y P I K L M D K E R T S W Q A P R I M T G E K
C S S Y P A K L M D F E R T S W Q L P R I K T G H K
C T S Y P I K L M D D E R T S W Q A P R I L T G R K
```

Correlated mutations

Correlated mutations

## And when there is no experimental structure? Homology modeling

AlphaFold:




Predicting distance between amino acids:

Sequence of target protein

↓

Sequence alignment with all related proteins

↓

Correlations between residues in the sequence alignments indicate that these residues are close in 3D, even if distant in the sequence alignment

↓

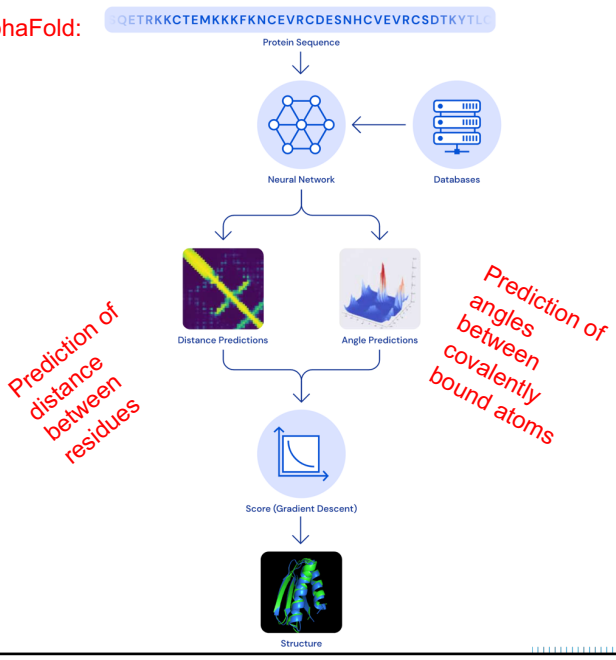
Prediction of the distance between residues





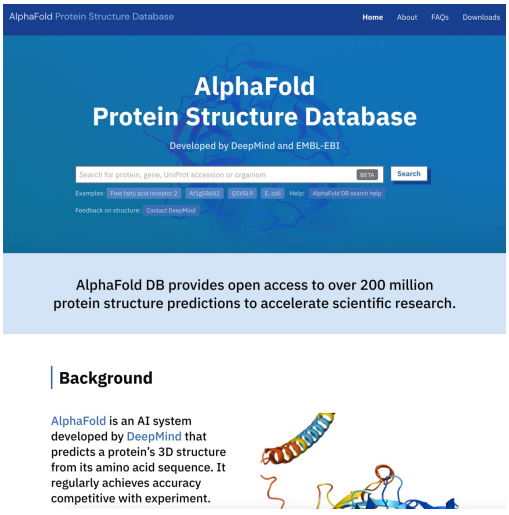
## And when there is no experimental structure? Homology modeling

AlphaFold:

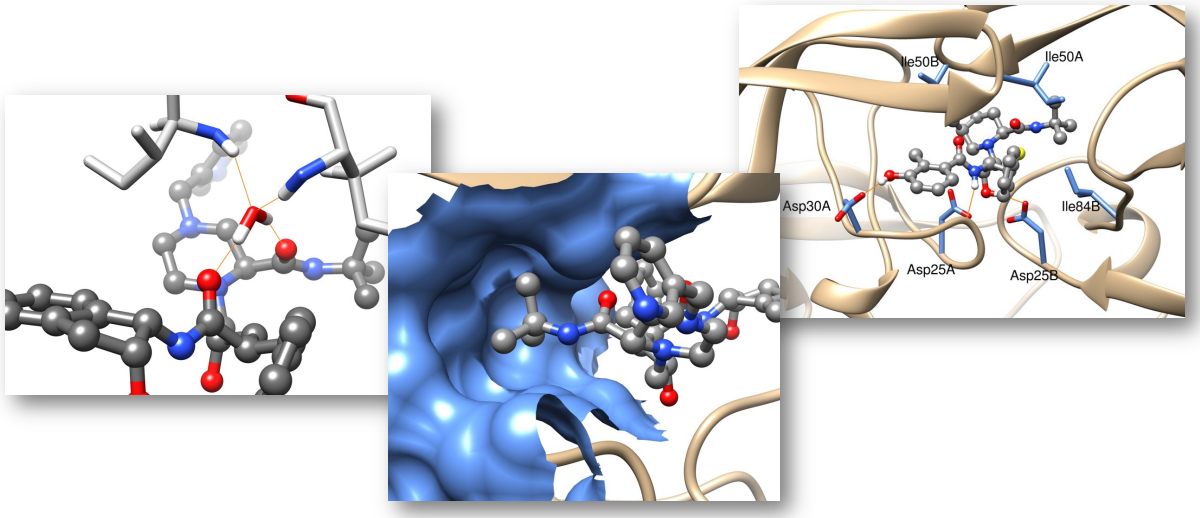



Database of models made with AlphaFold:

https://alphafold.ebi.ac.uk/



# Molecular Recognition

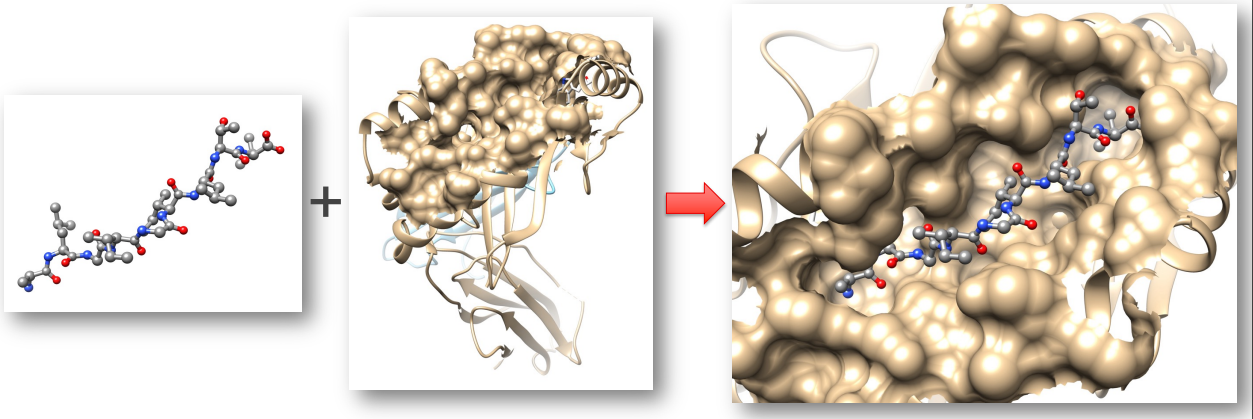





51

# Molecular recognition

Molecular interactions → Molecular recognition ⇒ Biological response





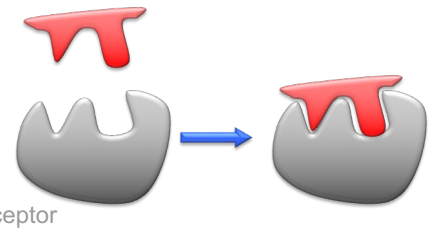
52

## Molecular recognition – Historical models

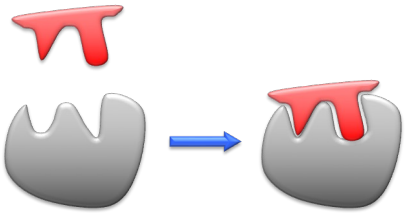
**"Lock and key" model.**
Emil Fischer in the 1890s.
The protein has a particular shape into which the ligand fits exactly.

**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.

Ligand



Receptor



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

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

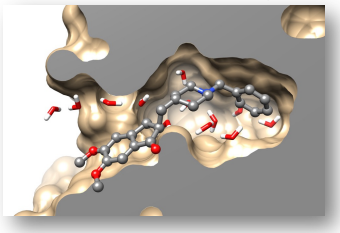




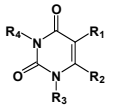
## Molecular recognition and CADD

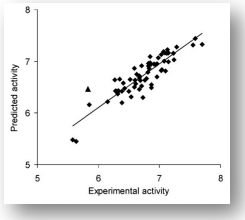
Two main categories of CADD approaches to discover, create, optimize and evaluate active molecules:

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



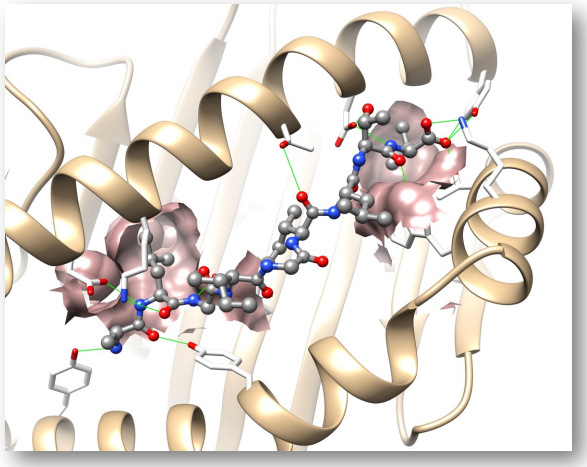
- Ligand-based approaches. Use the information derived from known ligands. Ex: Quantitative Structure-Activity Relationships (QSAR), bioisosteric replacements.



## 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

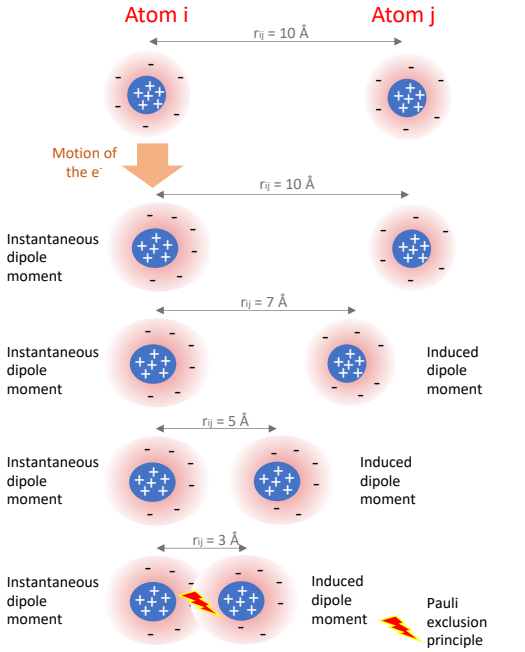


Crystal structure of HLA-A2*0201 in complex with MART-1/Melan-A



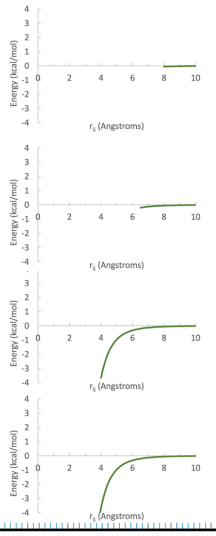
55

## Molecular recognition – Van der Waals interactions



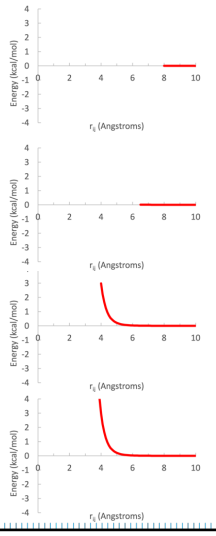

London dispertion energy

$$E = -\frac{B_{ij}}{r_{ij}^6}$$



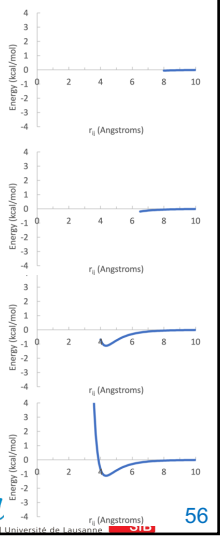
Pauli exclusion principle

$$E = \frac{A_{ij}}{r_{ij}^{12}}$$



van der Waals energy

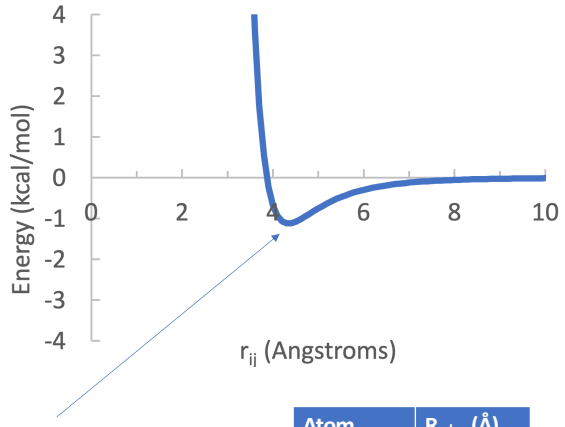
$$E = \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6}$$





56

## **Molecular recognition** – Van der Waals interactions




Optimum at $r_{ij} = R_{i,vdw} + R_{j,vdw}$

$R_{vdw}$: van der Waals radius

**Optimal interaction when atoms are « touching » each other**

| Atom | $R_{vdw}$ (Å) |
|---|---|
| Hydrogen | 1.2 |
| Carbon | 1.7 |
| Nitrogen | 1.55 |
| Oxygen | 1.52 |
| Sulfur | 1.8 |

Described by the Lennard-Jones potential

$$E = \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}$$

Interaction energy follows $1/r^6$ and $1/r^{12}$

↳ **Short range interaction**
Typically 3.5 Å

The optimal energy is weak between a given pair of atoms (Typically 0.5 kcal/mol)

However it is **cumulative** over all atoms involved in molecular recognition





## **Molecular recognition** – Van der Waals interactions

Do not require charges or partial charges on atoms

↳ van der Waals interactions are considered as **non-polar interactions**
… even though they are electrostatic by nature

Interactions particularly important for non-polar residues:
- Alanine, Valine, Leucine, Isoleucine, Proline
- Cysteine, Methionine
- Phenylalanine, Tyrosine, Tryptophan

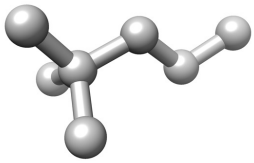
## **Molecular recognition** – Van der Waals interactions

Each atom tries to be positioned at optimal distance from its neighbors

2 atoms

3 atoms

4 atoms



However, in molecules, atoms are also linked via covalent bonds, which force a geometry…

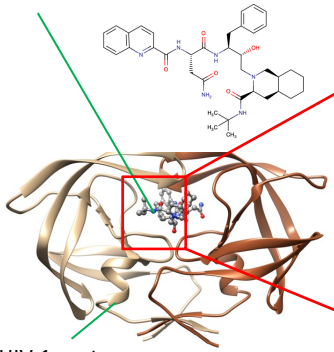




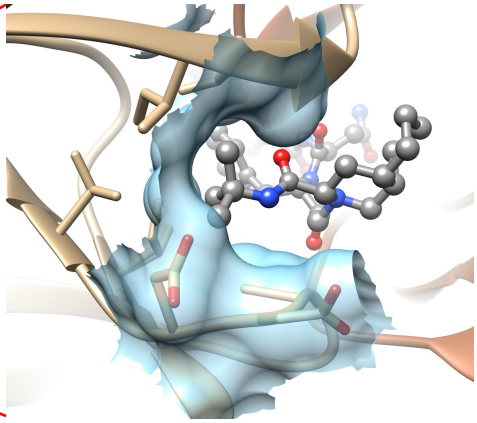
## **Molecular recognition** – Van der Waals interactions

Each atom tries to be positioned at optimal distance from its neighbors

**Saquinavir.** HIV-1 protease inhibitor
(Used in tri-therapy against HIV)

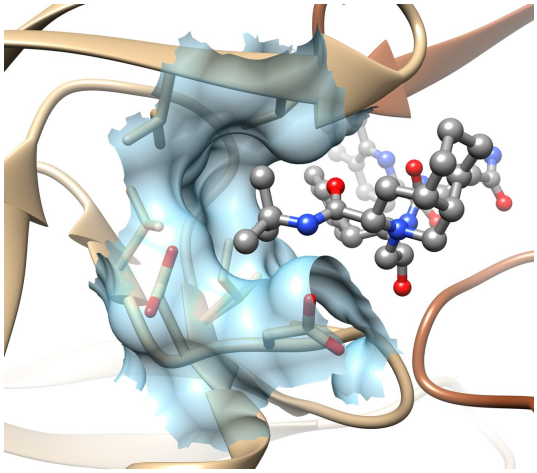


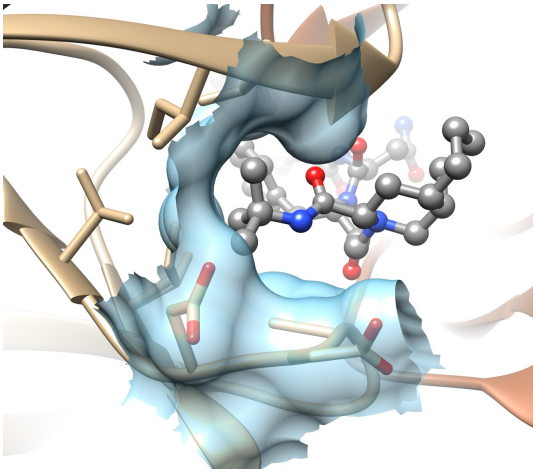
HIV-1 protease



## **Molecular recognition** – Van der Waals interactions

Each atom tries to be positioned at optimal distance from its neighbors



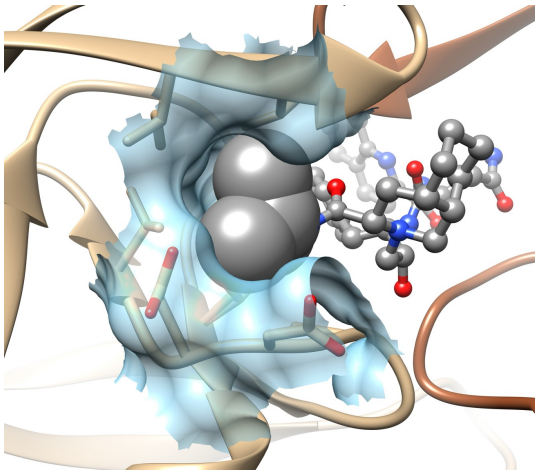


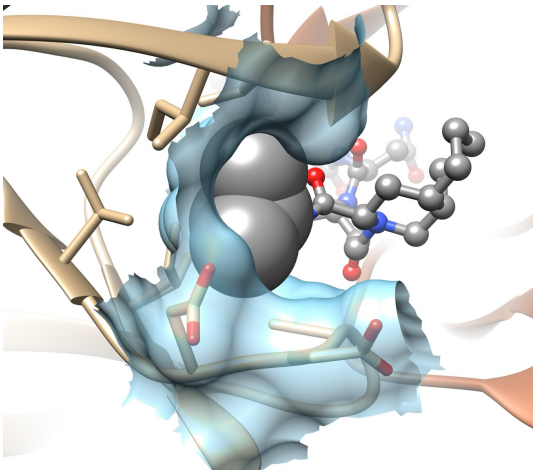


61

## **Molecular recognition** – Van der Waals interactions

Each atom tries to be positioned at optimal distance from its neighbors

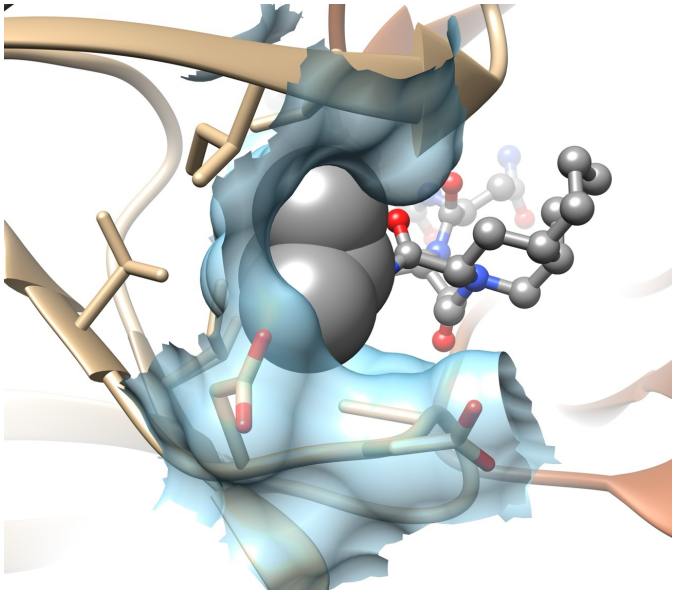






62

## **Molecular recognition** – Van der Waals interactions



Each atom tries to be positioned at optimal distance from its neighbors

van der Waals interactions contribute therefore to:

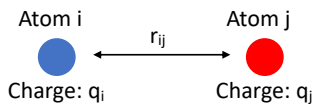
- packing of atoms (and macromolecule folding)
- shape complementarity between binding molecules (example: protein/protein or ligand/protéine complexes)





## **Molecular recognition** – Electrostatic interactions

The interaction between two point charges in a uniform medium is described by the **Coulomb law**




Coulomb energy

$$E_{Coul} = \frac{1}{4\pi\varepsilon_o\varepsilon}\frac{q_i q_j}{r_{ij}}$$

$\varepsilon_o$ : dielectric constant of vacuo

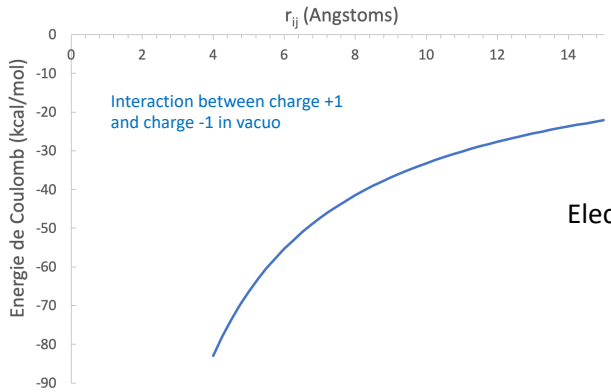
$\frac{1}{4\pi\varepsilon_o}$ =332 (kcal/mol) Å/$q_e^2$

$\varepsilon$: dielectric constant of medium

ex: $\varepsilon_{(vacuo)} = 1$ ; $\varepsilon_{(water)} = 80$

Interaction between charges +1 et -1 at 5 Å :

- -66 kcal/mol in vacuo
- -0.8 kcal/mol in water




Electrostatic interaction energy follows a 1/r expression

→ **Long range interaction**

## **Molecular recognition –** Electrostatic interactions

Electrostatic interactions can involve:

- Integer charge – integer charge

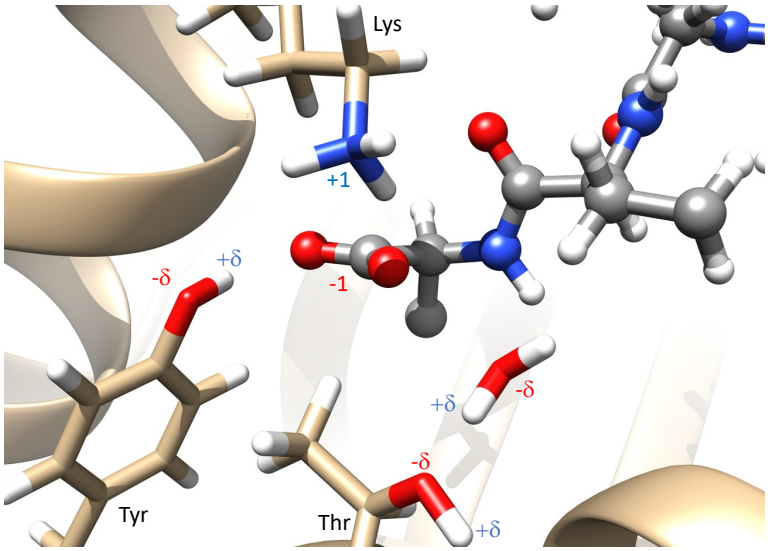
  Called ionic interactions.
  At short distance (~ 4/5 Å), ionic interactions are called salt bridges.

- Integer charge – permanent dipole

  Ex: charged assisted hydrogen bond

- Permanent dipole – permanent dipole

  Ex: hydrogen bond







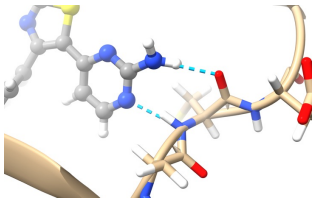

## **Molecular recognition –** Electrostatic interactions – Hydrogen bonds

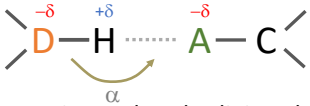
Typically between two dipoles:

- D-H where D is the hydrogen bond **donor**
- A-C where A is the hydrogen bond **acceptor** and C a carbon atom

**Extremely frequent** in proteins and nucleic acids

Important factor of the architecture of bio-macromolecules



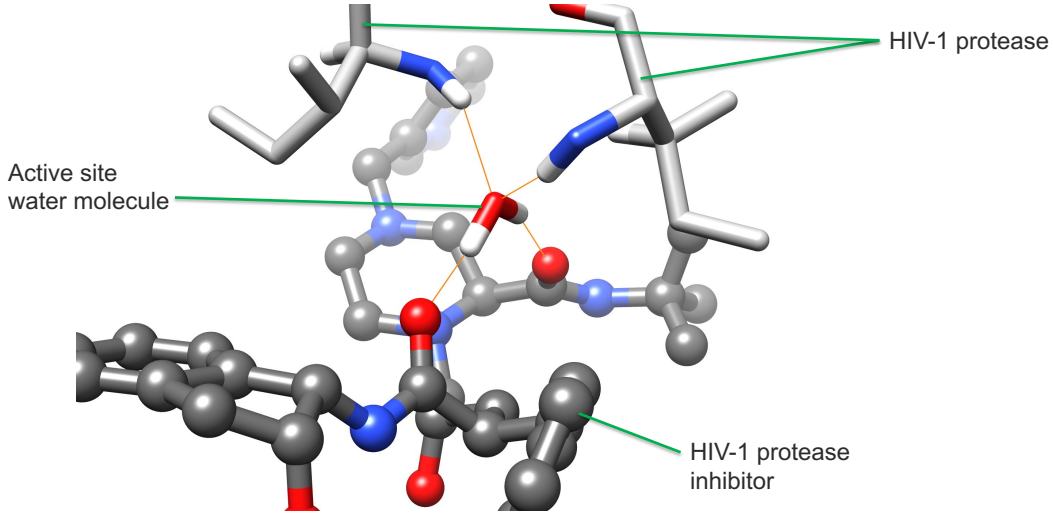



Distances typiques dans les liaison hydrogène :

- Entre H et A : ~ 1.95 Å
- Entre A et D : O – O : 2.50 – 2.70 Å
  O – N : 2.75 – 2.85 Å
  N – N : 2.70 – 3.00 Å

L'angle α dépend du type des atomes et de leur hybridation

## **Molecular recognition** – Electrostatic interactions – Hydrogen bonds




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

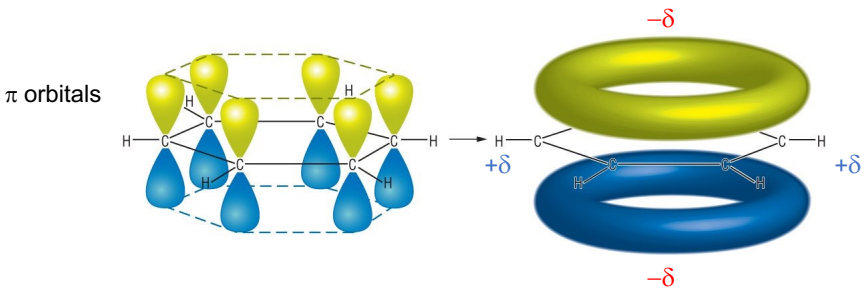
→ **Directionality / locality of interactions**
**Specificity of molecular recognition**



67

## **Molecular recognition** – π interactions

Electronic structure of benzene:



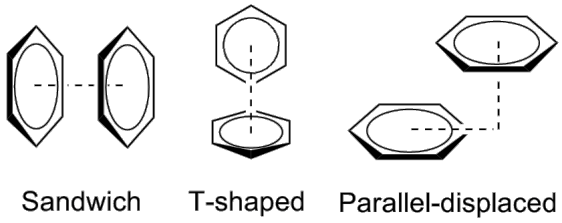

Aromatic cycles (Phenyl, Tyrosine, Tryptophan & Histidine) can interact with:

- Other aromatic cycles (stacking)
- Metals
- Polar groups
- Hydrogen bond donors



68

## Molecular recognition – π interactions



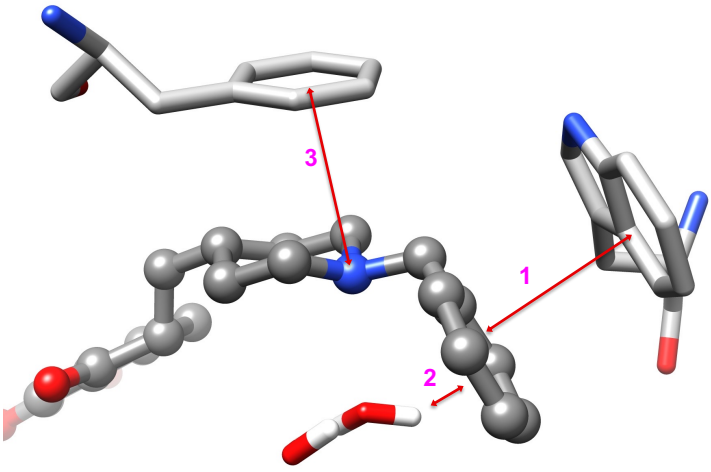

(source: Wikipedia)

T-shaped and parallel-displaced π-π interactions are the most frequent



69

## Molecular recognition – π interactions




1. π - stacking
2. OH-π interaction
3. Cation - π interaction

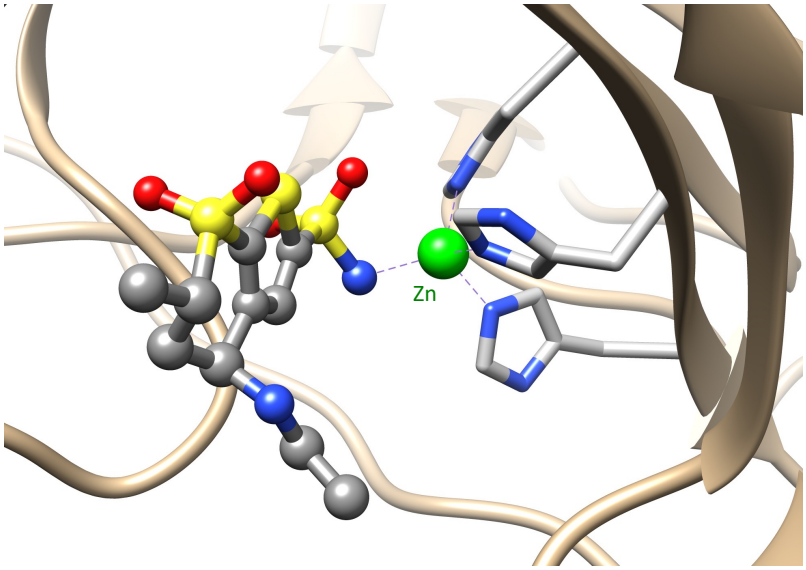
Ex: π interactions between Donepezil and acetylcholine esterase (PDB ID 1EVE)

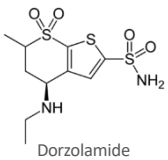


70

## **Molecular recognition** – Metal-ion interaction

Partially covalent




Dorzolamide

Ex: Dorzolamide, anti-glaucoma drug, in complex with carbonic anhydrase II (PDB ID: 3FW3)

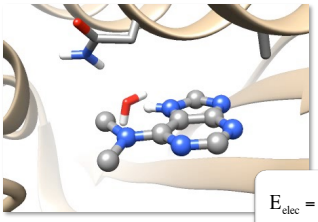


71

## **Molecular recognition** – Other factors

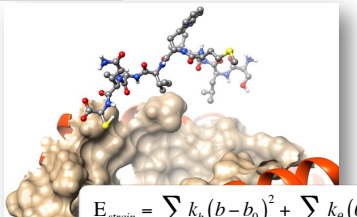
Many other factors impact the molecular recognition and binding affinity

Water bridges



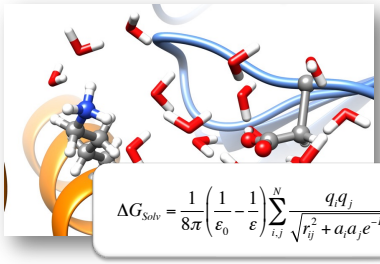
$$E_{elec} = \frac{q_i\, q_j}{4\pi\varepsilon_0\varepsilon\, r_{ij}}$$

Conformational changes



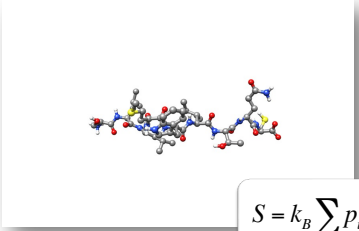
$$E_{strain} = \sum_{bonds} k_b (b - b_0)^2 + \sum_{angles} k_\theta (\theta - \theta_0)^2 + \ldots$$

Desolvation and elec. shielding



$$\Delta G_{Solv} = \frac{1}{8\pi}\left(\frac{1}{\varepsilon_0} - \frac{1}{\varepsilon}\right)\sum_{i,j}^{N} \frac{q_i q_j}{\sqrt{r_{ij}^2 + a_i a_j e^{-D}}} \quad , \quad D = \left(\frac{r_{ij}}{2\sqrt{a_i a_j}}\right)^2$$

Entropy changes



$$S = k_B \sum p_i \ln(p_i)$$

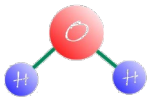


72

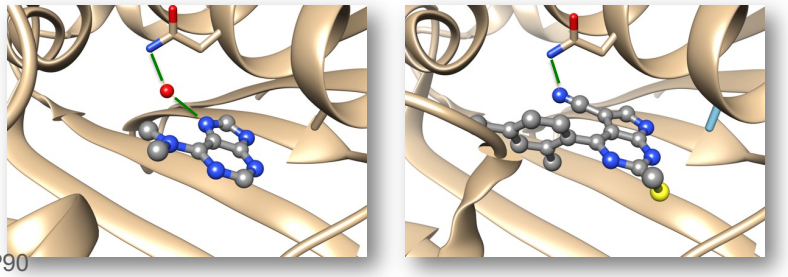
## **Molecular recognition** – Other factors – Water

Molecular recognition between small molecule and protein takes place in an **aqueous environment.**

**Discrete water molecules**

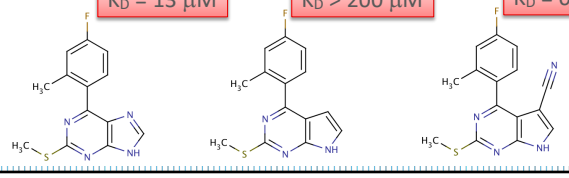


- Bridge interactions through H-bonds or OH... π ➔ favorable to binding.
- Displacement from the protein cavity ➔ favorable to binding.



HSP90

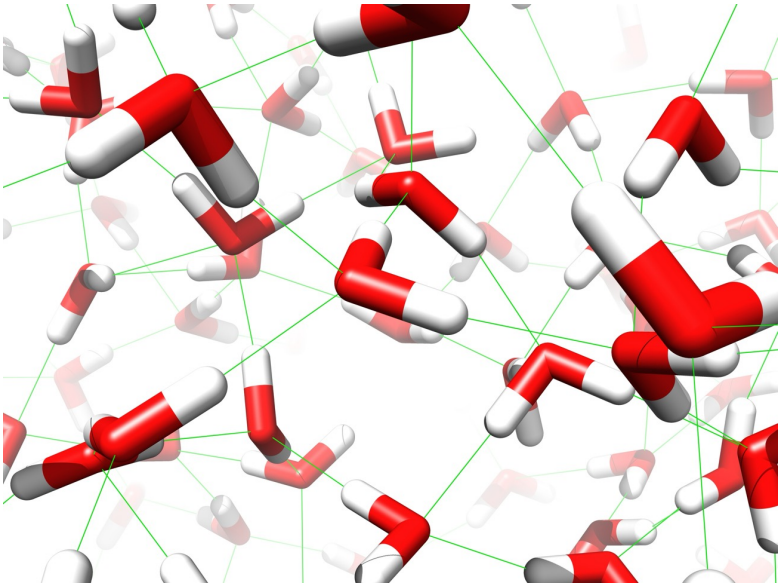
$K_D$ = 13 µM | $K_D$ > 200 µM | $K_D$ = 0.2 µM





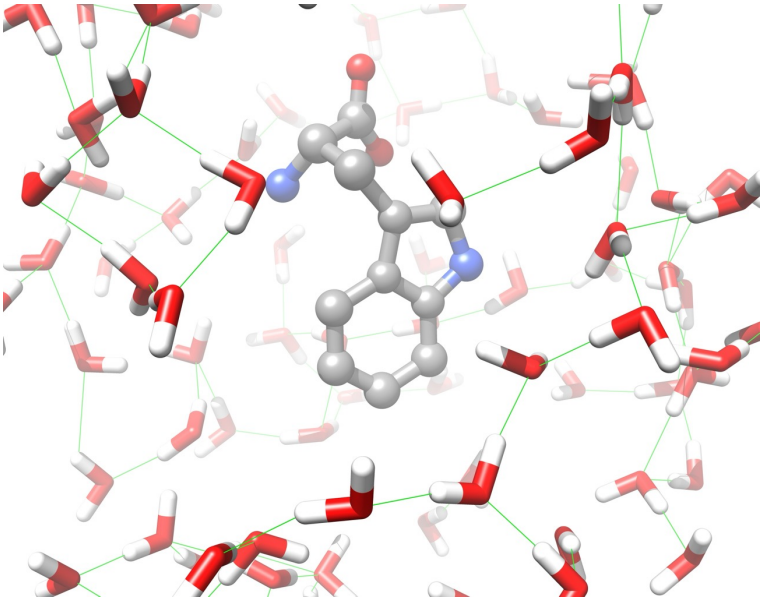


## **Molecular recognition** – Other factors – Water – Hydrophobic effect



Water structure is stabilized by hydrogen bonds and dipole interactions

## **Molecular recognition** – Other factors – Water – Hydrophobic effect

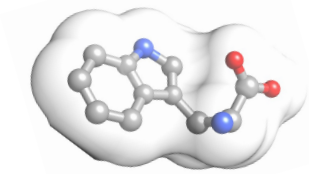


The presence of a solute decreases water-water interactions

Non-polar solvation energy is proportional to the solvent accessible surface area (SASA) for large molecules:

$$E = \sigma \times SASA$$

$\sigma$ = 0.025 kcal/Å$^2$

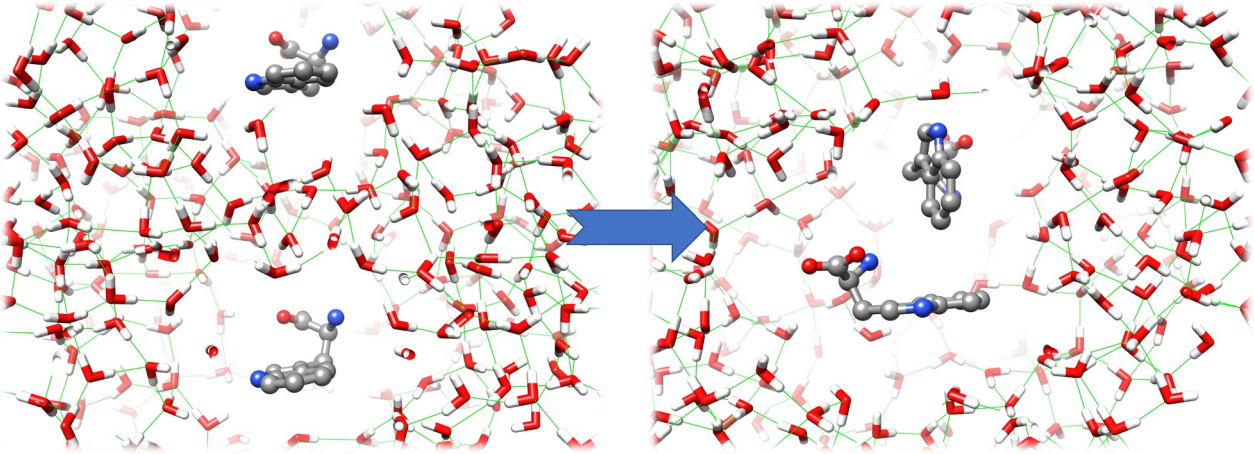




75

## **Molecular recognition** – Other factors – Water – Hydrophobic effect

Solutes aggregate to limit their deleterious on water structure



Energy of non-polar desolvation: $\Delta G_{np} = \sigma \times \Delta SASA$

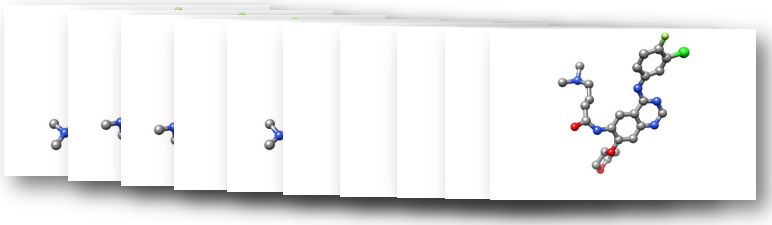
The solvent-accessible surface area of aggregated solutes is lower than the sum of those of the separated solutes ($\Delta$SASA<0). $\Delta G_{np}$ is therefore favorable to aggregation (binding of solutes)

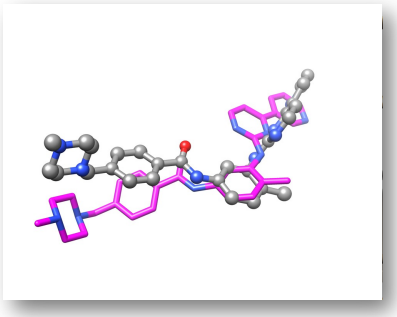


76

## **Molecular recognition** – Other factors – Conformational changes

**Molecules have many conformations (conformers)**





Ligand **bioactive** conformation **(geometry as bound to the protein)**

does **NOT** correspond to

**Lowest energy** conformation **(most stable geometry in solution)**

BUT is a low energy conformation (within 3 to 5 kcal/mol)

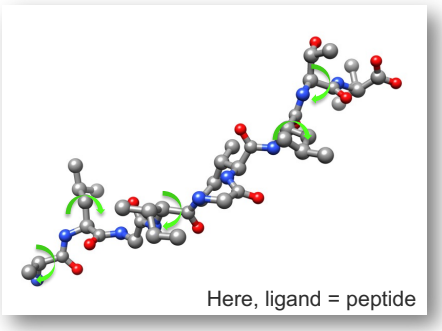
Bioactive conformation (in protein)
Lowest energy conformation (in solution)

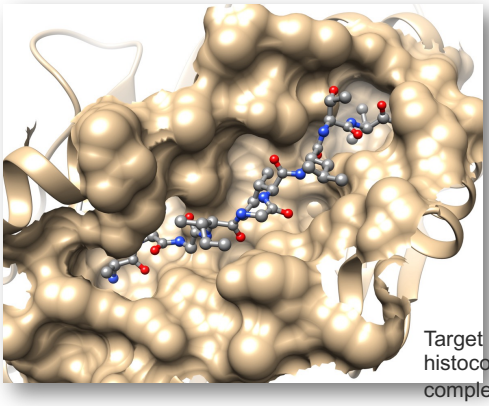




## **Molecular recognition** – Other factors – Entropy changes

**Entropy is a measure of disorder. Nature likes disorder!**
Loss of entropic energy when entropy (disorder) decreases.
Gain of entropic energy when entropy (disorder) increases.



Target = MHC (Major histocompatibility complex)

Two main events **upon ligand binding** to protein:

- **Conformational** degrees of **freedom** (rotatable bonds) are **blocked**: **unfavorable!**
- **Water** molecules are **kicked-out** from the protein binding site **to bulk**: **favorable!**

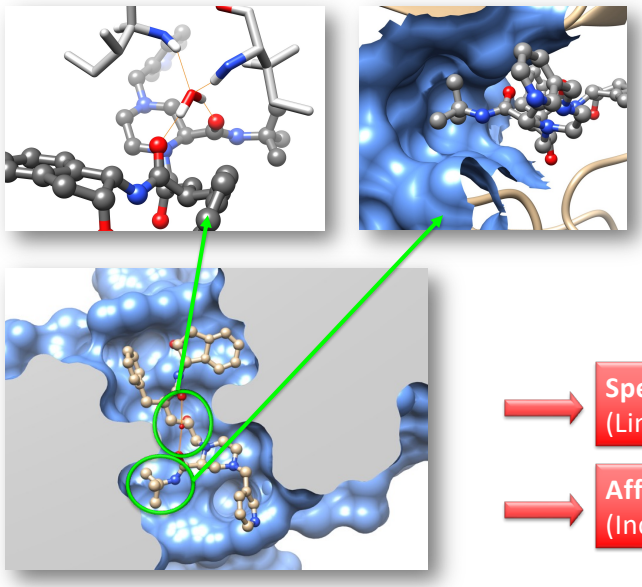
## **Molecular recognition** – Summary

| Category | Interaction | Distance | Residues involved | Remarks |
|---|---|---|---|---|
| Electrostatic | Ionic (charge-charge) | Long range | Arg, Lys, Asp, Glu<br>His (if charged) | Called salt bridge at short distance |
| | Hydrogen bond | Short range | Arg, Lys, Asp, Glu<br>His, Tyr<br>Ser, Thr, Asn, Gln<br>Cys | Directionality / locality of interactions<br>Specificity of molecular recognition |
| | $\pi$ interaction | Short range | Phe, Tyr, Trp, His | |
| Electrostatic/Non-polar | Van der Waals | Short range | Ala, Val, Ile, Leu, Pro, Cys, Met<br>Phe, Tyr, Trp, His | Packing of atoms<br>Shape complementarity |
| Non-polar | Hydrophobic effect | - | All | Solute aggregation |





## **Molecular recognition** – Potency and specificity



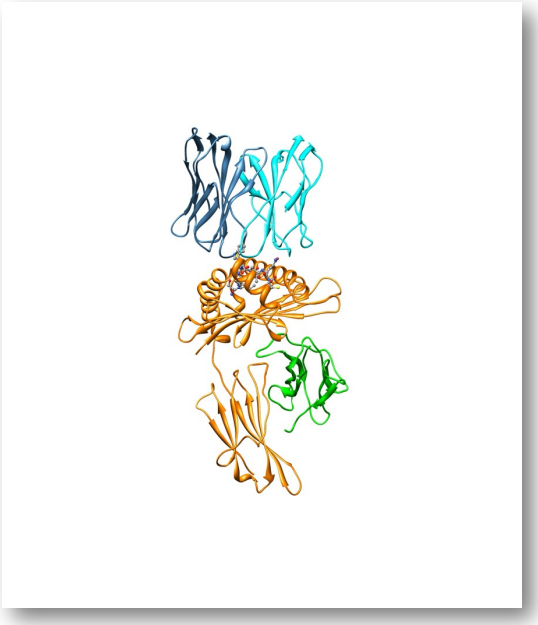
Various and numerous ligand-protein interactions:

- local and directional interactions
- shape complementarity

→ **Specificity**
(Limits number/nature of possible epitopes)

→ **Affinity/potency**
(Increased epitope recognition)

## **Molecular recognition -** Molecular Motions - 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**.

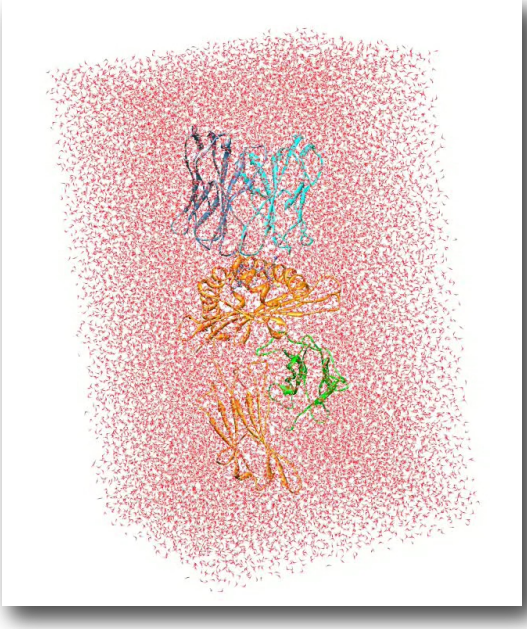
Typical simulation times: from **0.5 ns to ~ 1000 ns** (1 ns = $10^{-9}$ s).

➔ **Simulation closer to physiological reality, but more computationally intensive**





## **Molecular recognition -** Molecular Motions - 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**)

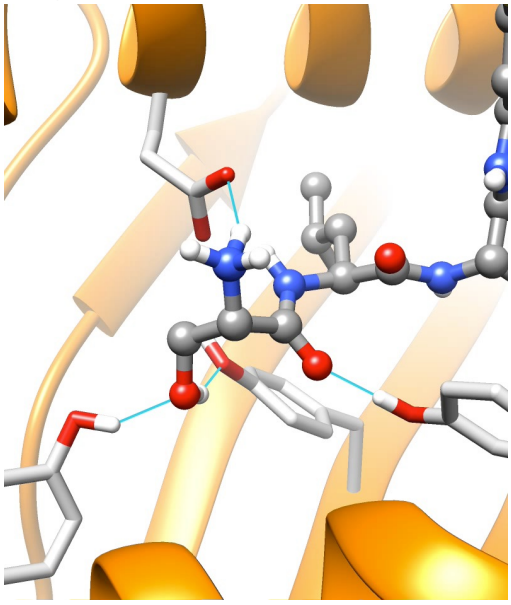
Atom motions are calculated to follow Newton's equation of motion, at **300 K** and **1 atm**.

Typical simulation times: from **0.5 ns to ~ 1000 ns** (1 ns = $10^{-9}$ s).

➔ **Simulation closer to physiological reality, but more computationally intensive**

## **Molecular recognition -** Molecular Motions - Molecular Dynamics

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



Peptide epitopte in ball and stick representation

MHC protein in ribbon representation with some side chains in stick representation

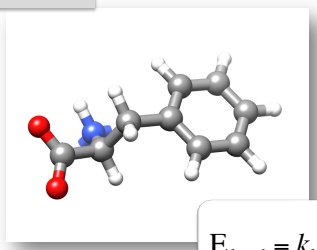




## Molecular recognition – introduction to molecular mechanics

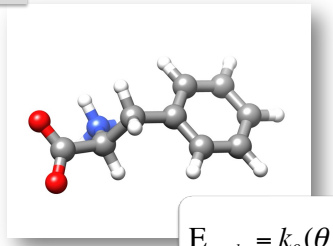
**Molecular dynamics is decomposed into elementary motions**

Bond length



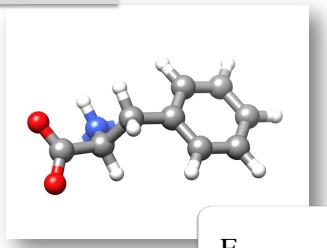
$$E_{bond} = k_b(b - b_0)^2$$

Bond angle



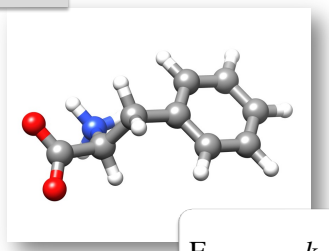
$$E_{angle} = k_\theta(\theta - \theta_0)^2$$

Dihedral angle



$$E_{dihedral} = k_\varphi\left(1 + \cos(n\varphi - \delta)\right)$$

Improper angle

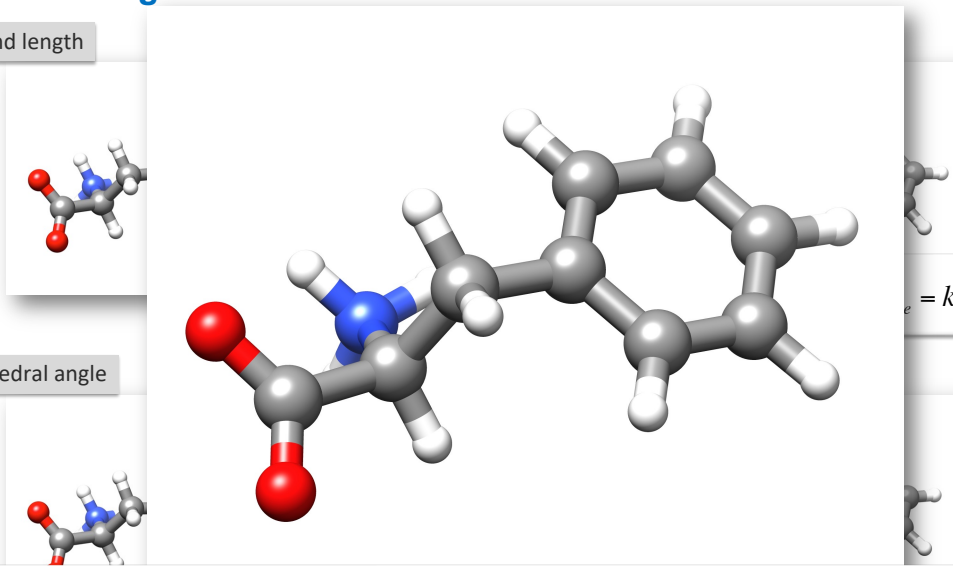


$$E_{improper} = k_\omega(\omega - \omega_0)^2$$

## Molecular recognition – introduction to molecular mechanics

Molecular dynamics is decomposed into elementary motions

Bond length



$= k_\theta(\theta - \theta_0)^2$

Dihedral angle

$$E_{bonded} = \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} k_\varphi(1 + \cos(n\varphi - \delta)) + \sum_{impropers} k_\omega(\omega - \omega_0)^2$$

$\omega_0)^2$

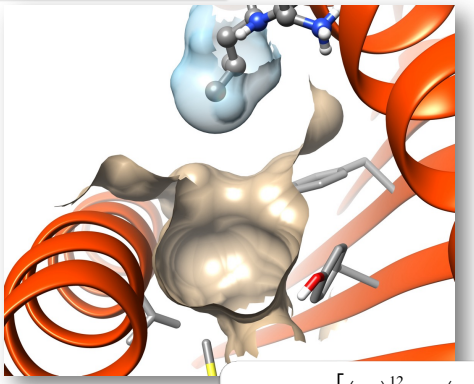




## Molecular recognition – Molecular interactions

Molecular recognition is driven by non-polar and electrostatic interactions

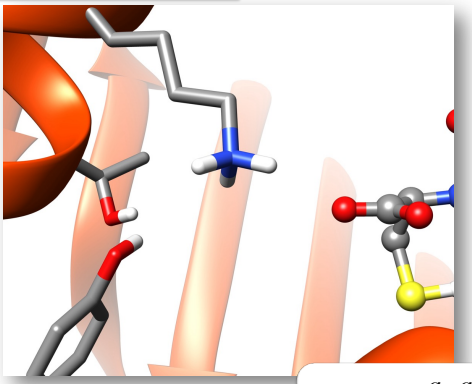
Non-polar interactions



$$E_{vdW} = \varepsilon\left[\left(\frac{r_m}{r_{ij}}\right)^{12} - 2\left(\frac{r_m}{r_{ij}}\right)^{6}\right]$$

Shape complementarity

Electrostatic interactions



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

Specificity

## Molecular recognition - type of interactions

**Non Polar:**

Ala, Val, Leu, Ile, Pro, Met, ~Cys

**Polar:**

Ser, Thr, Asn, Gln, Tyr, His, Trp, ~Cys

**Aromatic:**

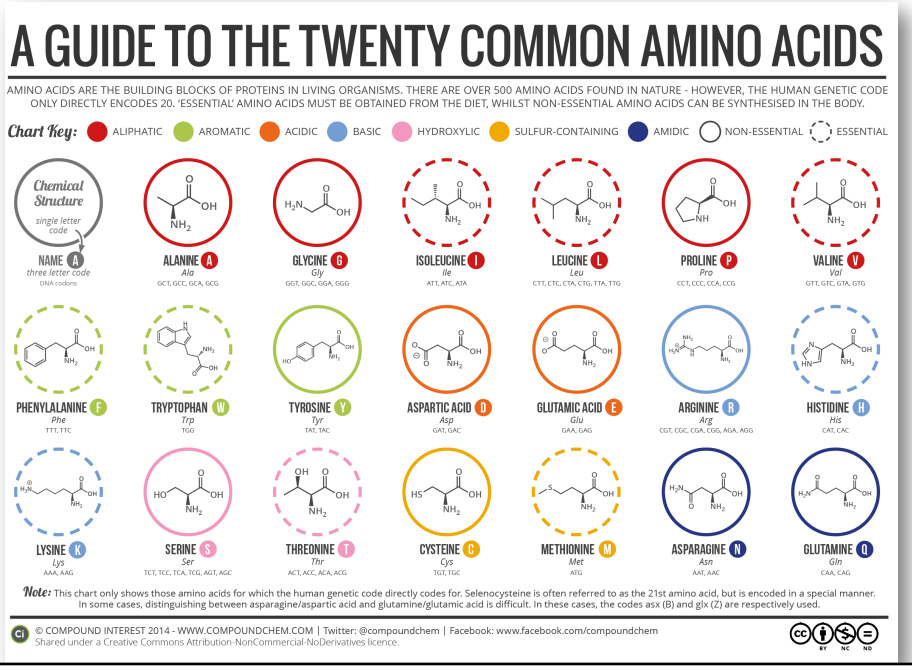
Phe, Tyr, Trp, His

**Negatively charged:**

Asp, Glu

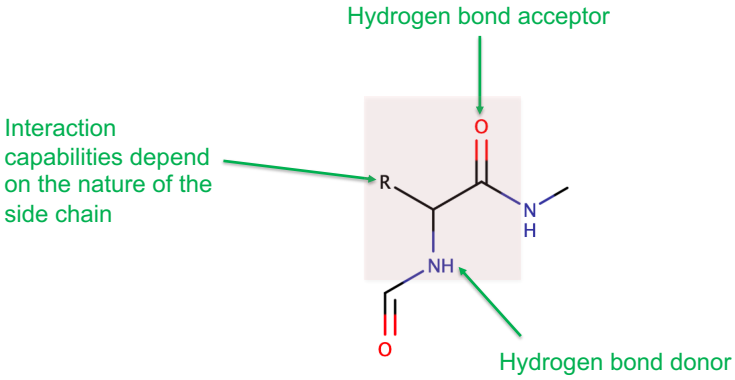
**Positively charged:**

Arg, Lys, ~His




88

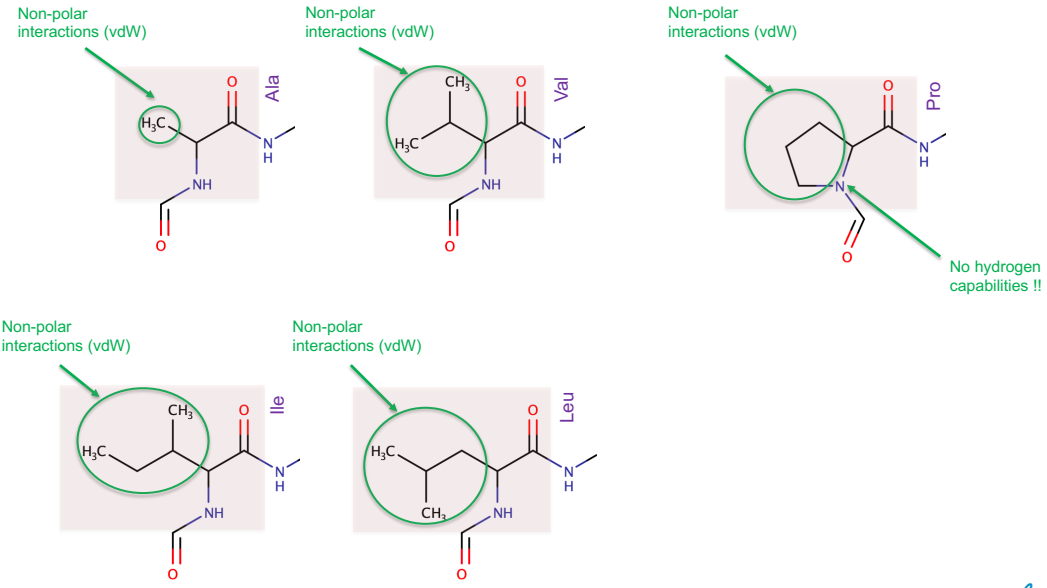
## Interactions Moléculaires – Backbone des acides aminés






89

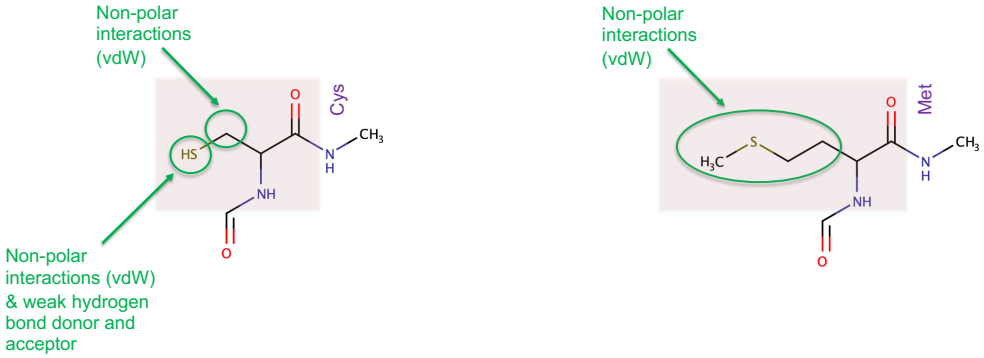
## Interactions Moléculaires – Chaînes latérales des acides aminés






90

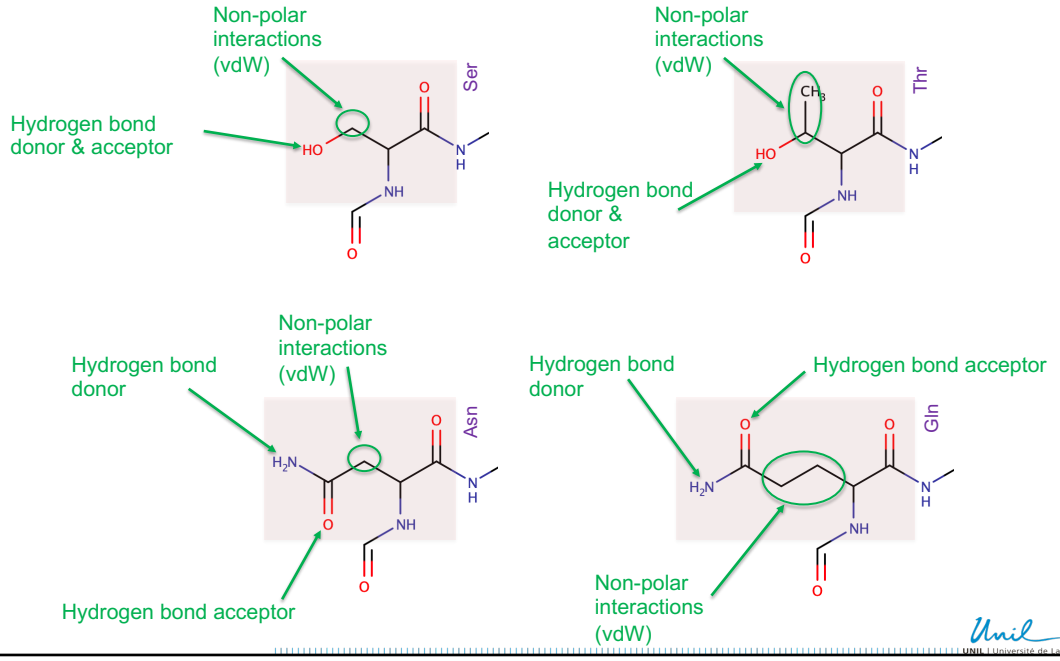
## Interactions Moléculaires – Chaînes latérales des acides aminés






91

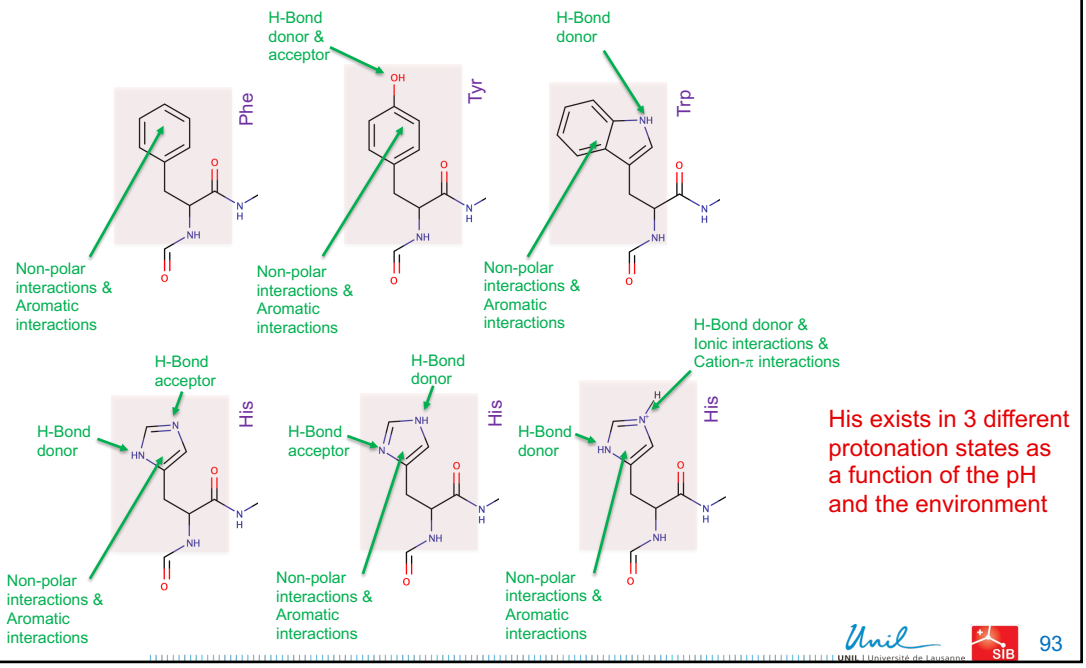
## Interactions Moléculaires – Chaînes latérales des acides aminés






92

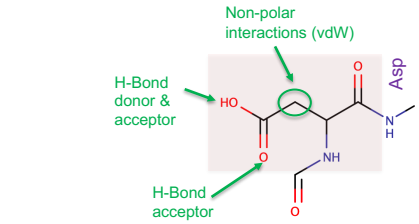
## Interactions Moléculaires – Chaînes latérales des acides aminés

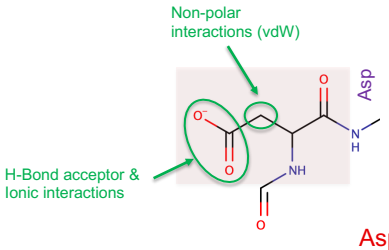




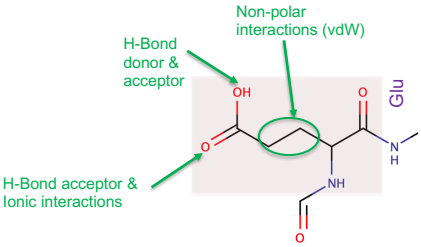

93

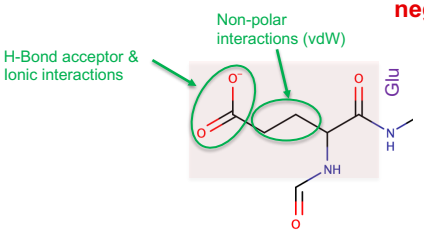
## **Molecular recognition** – Possible interactions per amino acids




Asp and Glu exist in 2 different protonation states as a function of the pH and the environment
But they are **generally negatives**

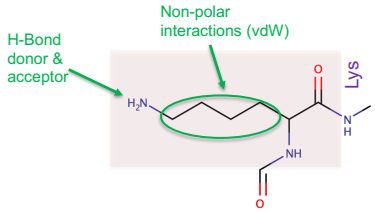


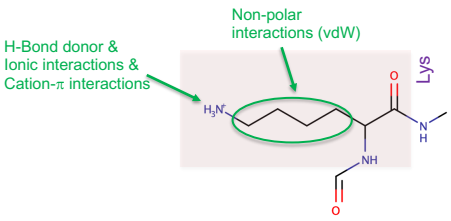





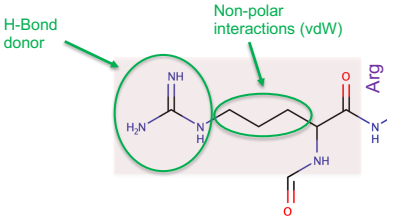

94

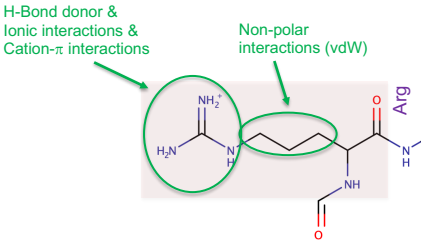
## **Molecular recognition** – Possible interactions per amino acids




Arg and Lys exist in 2 different protonation states as a function of the pH and the environment
But they are **generally positives**






95

## Molecular recognition

Let's start with UCSF Chimera !!!

Contacts: vincent.zoete@unil.ch , antoine.daina@sib.swiss