Follow-up of offline activities

1. **Questions about Tsai *et al.***
   - **How many** compounds were screened? (20,000) What information is available about their **properties**? (kinase inhibition, molecular weight between 150 and 350 daltons)
   - **How** were the compounds **screened**? (single-dose 200 uM, crystallography with structurally divergent kinases)
   - What was the **initial chemical structure** that was found to bind to the ATP-binding site? (7-azaindole)
   - By overlapping structures, the team aimed to optimizing what **two properties of the compounds**? (potency and selectivity)
   - What types of compounds were tested in the **subsequent screening**? (mono- and di-substituted analogs built around the 7-azaindole core)
   - What properties does the PLX4720 compound have that make it **particularly attractive** as a drug? (affinity, selectivity, and a good safety profile)
2. **Questions from the anonymous survey:**
   - **Difference between divide-and-conquer and dynamic programming:** they are indeed different strategies (thanks David Sommer!). These discussions on StackOverflow may help you recognize the commonalities and differences
   - **How were papers selected?** Based on four considerations: (1) topic relevant for drug discovery, (2) reasonably well written, (3) balancing recent and classic literature, and (4) widely-used information resources. They remain however a limited and biased selection.

Exercises of lecture 3 and 4

	H	E	A	G	A	W	G	H	E	E
P	-2	-1	-1	-2	-1	-4	-2	-2	-1	-1
A	-2	-1	5	0	5	-3	0	-2	-1	-1
W	-3	-3	-3	-3	-3	15	-3	-3	-3	-3
H	10	0	-2	-2	-2	-3	-3	10	0	0
E	0	6	-1	-3	-1	-3	0	0	6	6
A	-2	-1	5	0	5	-3	0	-2	-1	-1
E	0	6	1	3	1	-3	-3	0	6	6

Adapted from *Biological Sequence Analysis* (R. Durbin, S. Eddy, A. Krogh, G. Mitchison), Figure 2.3. We assume that a gap cost per unaligned residue of *d=-8*. Try to use the information to perform global alignment between the two amino-acid sequences:

1. HEAGAWGHEE
2. PAWHEAE

		H	E	A	G	A	W	G	H	E	E
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80
P	-8	-2	-9	-17	-25	-33	-42	-49	-57	-65	-73
A	-16	-10	-3	-4	-12	-20	-28	-36	-44	-52	-60
W	-24	-18	-11	-6	-7	-15	-5	-13	-21	-29	-37
H	-32	-14	-18	-13	-8	-9	-13	-7	-3	-11	-19
E	-40	-34	-8	-16	-16	-9	-12	-15	-7	3	-5
A	-48	-42	-16	-3	-11	-11	-12	-12	-15	-5	2
E	-56	-50	-24	-11	-6	-12	-14	-15	-12	-9	1

```
HEAGAWGHE-E
--P-AW-HEAE
```

What does Fomivirsen target?

It is possible to search for local sequence matches in large databases of nucleotides, for instance using the BLAST algorithm. An implementation is freely available at National Institute of Health (NIH, US): https://blast.ncbi.nlm.nih.gov/Blast.cgi. Try to search for the RNA/protein targeted by fomivirsen, given its sequence 5'-GCG TTT GCT CTT CTT CTT CTT GCG-3'.

Human betaherpesvirus 5 strain SYD-SCT1, complete genome	42.1	42.1	100%	0.14	100.00%	MT044485.1
Human betaherpesvirus 5 strain HAN-SOT4, complete genome	42.1	42.1	100%	0.14	100.00%	MT044484.1
Human betaherpesvirus 5 strain HAN-SOT3, partial genome	42.1	42.1	100%	0.14	100.00%	MT044483.1
Human betaherpesvirus 5 strain GLA-SOT3, complete genome	42.1	42.1	100%	0.14	100.00%	MT044482.1
Human betaherpesvirus 5 strain GLA-SOT2, complete genome	42.1	42.1	100%	0.14	100.00%	MT044481.1
Human betaherpesvirus 5 strain SYD-SCT2, complete genome	42.1	42.1	100%	0.14	100.00%	MT044480.1
Human betaherpesvirus 5 strain HAN-SOT5, complete genome	42.1	42.1	100%	0.14	100.00%	MT044479.1
Human betaherpesvirus 5 strain HAN-SOT1, complete genome	42.1	42.1	100%	0.14	100.00%	MT044478.1
Human betaherpesvirus 5 strain GLA-SOT4, complete genome	42.1	42.1	100%	0.14	100.00%	MT044477.1

An example of HMM

```
library(HMM)
hmmModel <- initHMM(States=c("A", "D"), # A=Angel, D=Devil
                    Symbols=c("B", "R"), # B=Blessing, R=Rant
                    startProbs=c(0.5, 0.5),
                    transProbs = matrix(c(0.8, 0.2, 0.2, 0.8), nrow=2),
                    emissionProbs = matrix(c(0.9, 0.1, 0.1, 0.9), nrow=2))
simHmm <- simHMM(hmmModel, 100)
simStates <- paste(simHmm$states, collapse="")
simSymbols <- paste(simHmm$observation, collapse="")
```

ADDDDDDADDDDDDDDAAAA

BBRRBRRBRRBRRRRRBBBB

AAAAADAADDDDDDAADAAA

BBBBBRBRBRRRBRBBRBBB

AAADDDDDDADDDDADDAAA

BBBRRRRRBBRRRRBRRRBB

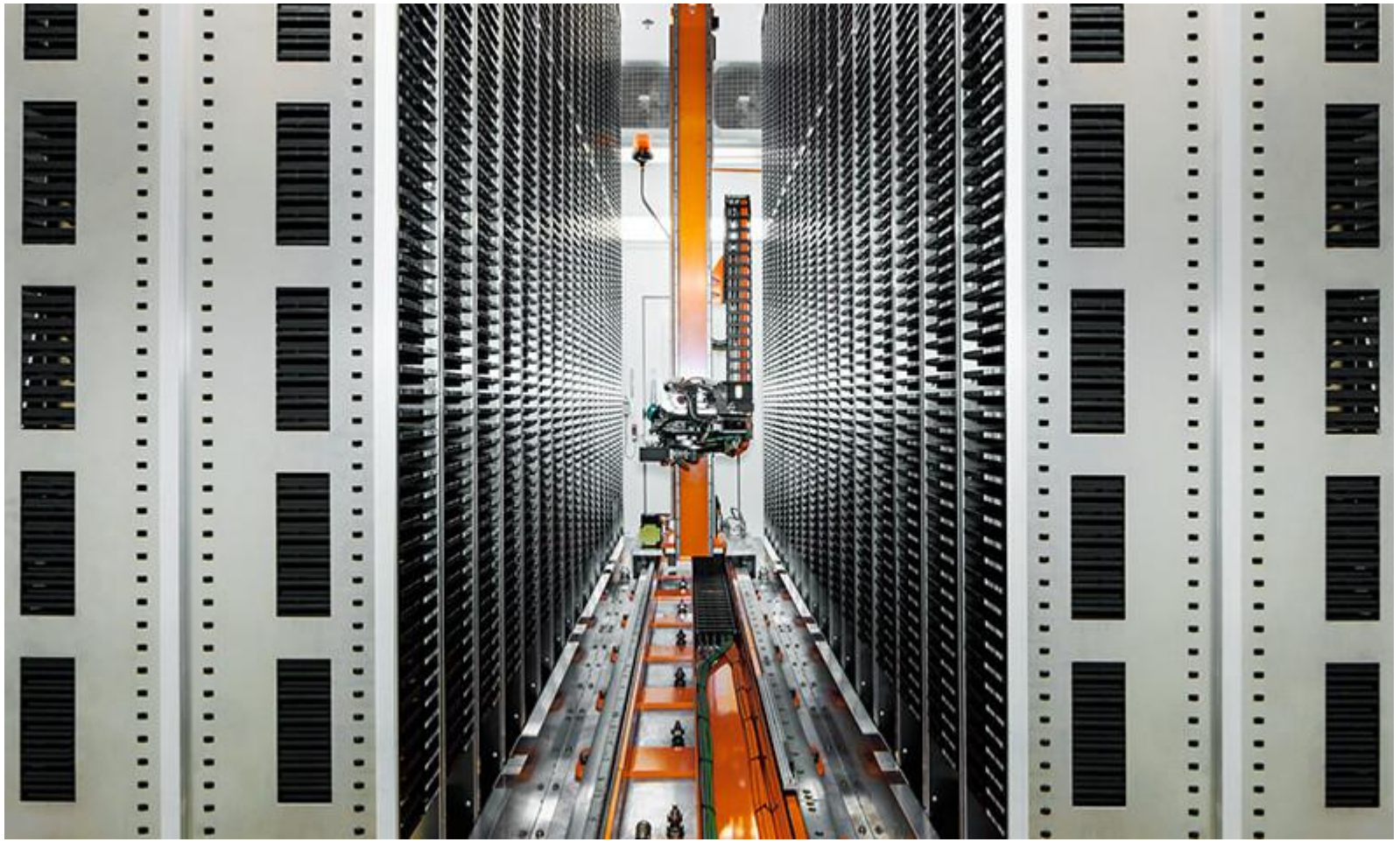
AAAADDAAAAAAADDDDDDD

BBBRRRBBBBBBBRRRRRRR

DDDDDDDDDDDDDAADAADD

RRRRRBRRRRBRRBBRRBRR

AMIDD Lecture 5: Proteins and Ligands



The chemical library at Novartis headquarters in Basel currently contains roughly 3 million molecules. We aim to expand that number radically within the next few years.

Jay Bradner, President of NIBR, in an interview in 2017

Dr. Jitao David Zhang, Computational Biologist

[1] Pharmaceutical Sciences, Pharma Research and Early Development, Roche Innovation Center Basel, F. Hoffmann-La Roche

[2] Department of Mathematics and Informatics, University of Basel

Today's goals

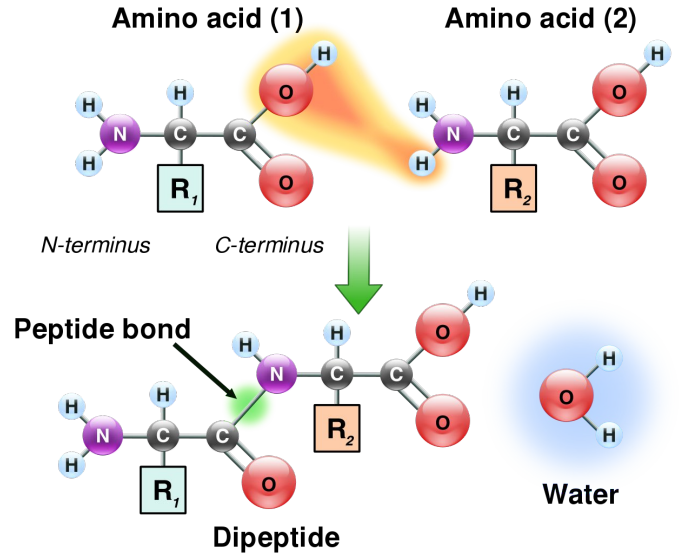
- Protein biology and structure determination
- Representation and molecular descriptors of small molecules
- Two views of ligand-target binding

What properties must a drug satisfy?

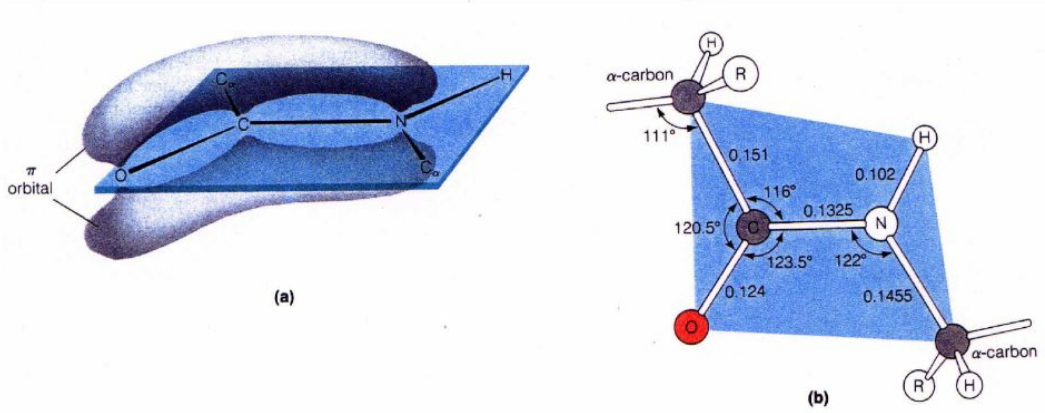
- Potency
- Selectivity
- Physico-chemical properties
- Administration, Distribution, Metabolism, Excretion (ADME)
- Safety
- Formulation
- Stability
- ...

From amino acids to proteins

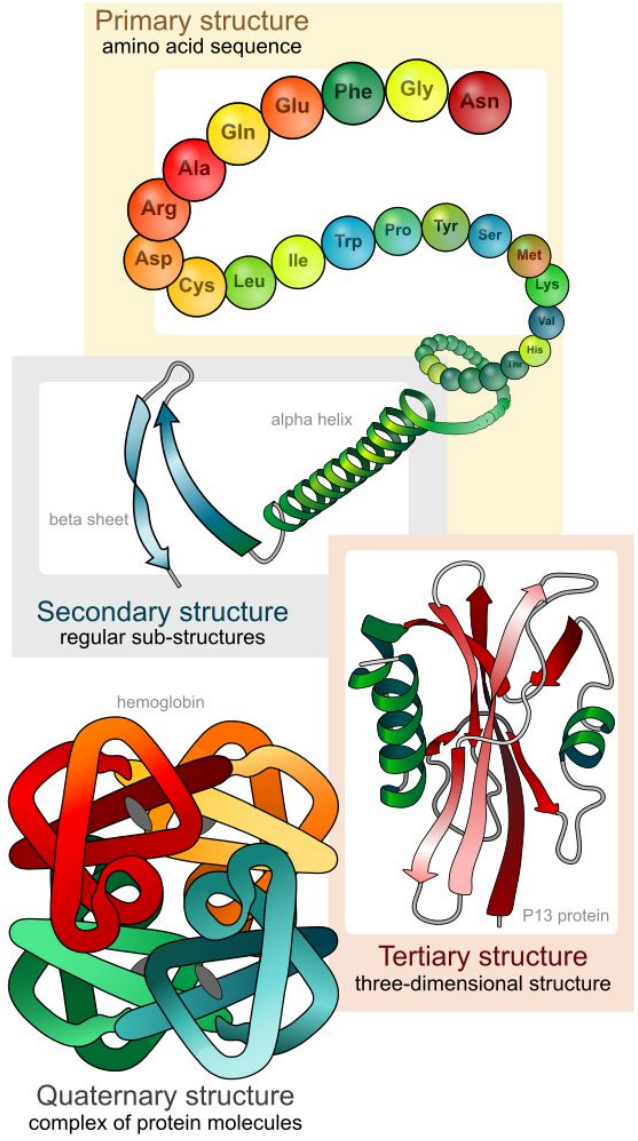
- Translation of mRNA means that two consecutive amino acids specified by 3-nucleotide codons form **peptide bonds** (top left panel). The peptide bonds concatenate amino acids together into *peptides* or *proteins.*
- The peptide plane geometry, determined by X-ray crystallography, is used to model structures and proteins. (bottom left panel).
- Protein structures can be thought of as hierarchical: primary amino-acid sequences form secondary structures (alpha helices and beta sheets), which form 3D structures of proteins, which can further form complexes (right panel).



Peptide bond formation

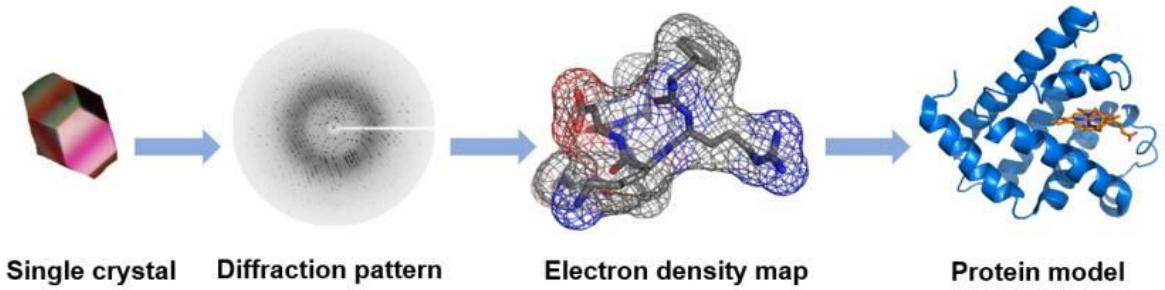


Peptide plane geometry. (Left) distribution of electrons in the bond (right) bond angles and distances by X-ray. Source

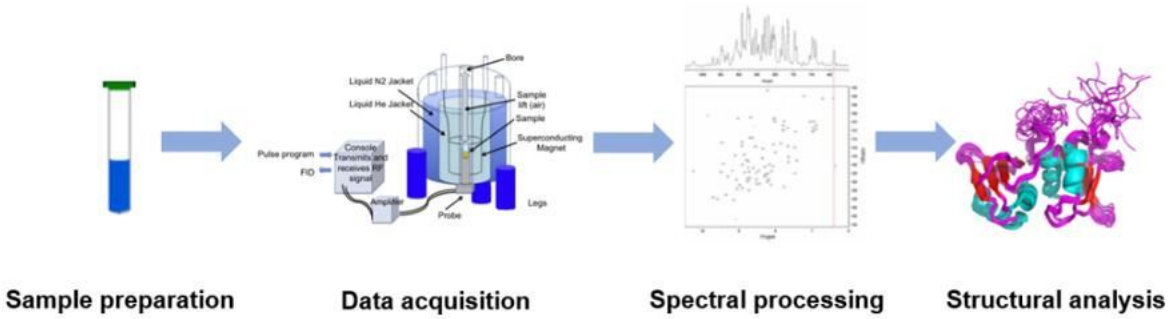


Four levels of protein structures

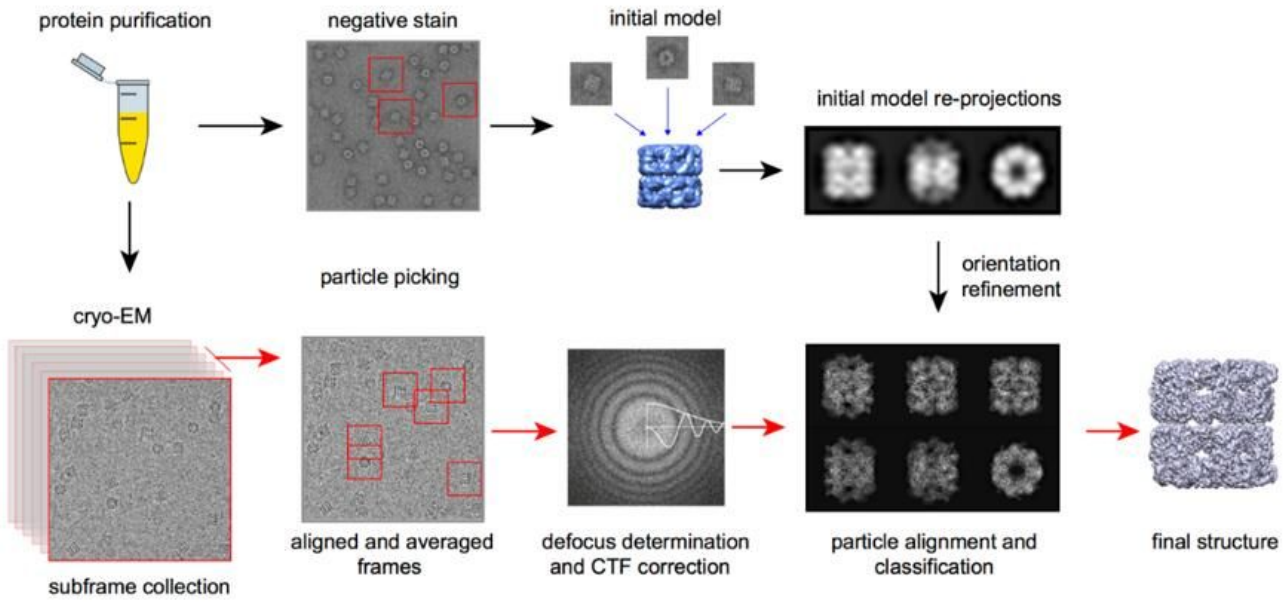
Three major experimental approaches to determining protein structures



X-ray crystallography



Nuclear Magnetic Resonance (NMR)



Cryo-electron microscopy (CryoEM)

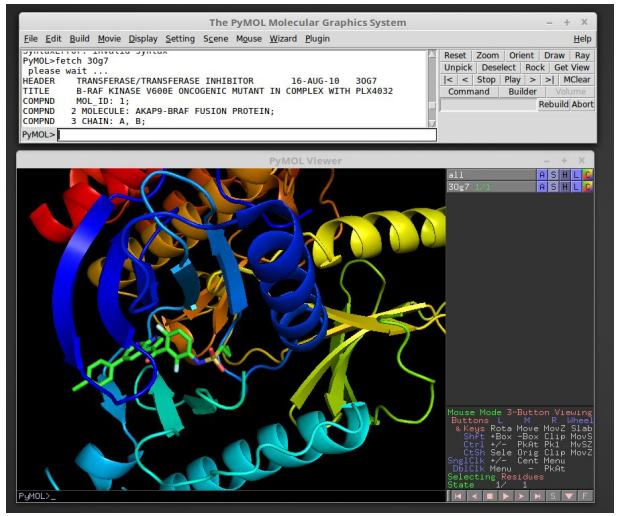
Figure sources:
https://www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm

Three major experimental approaches to determining protein structures

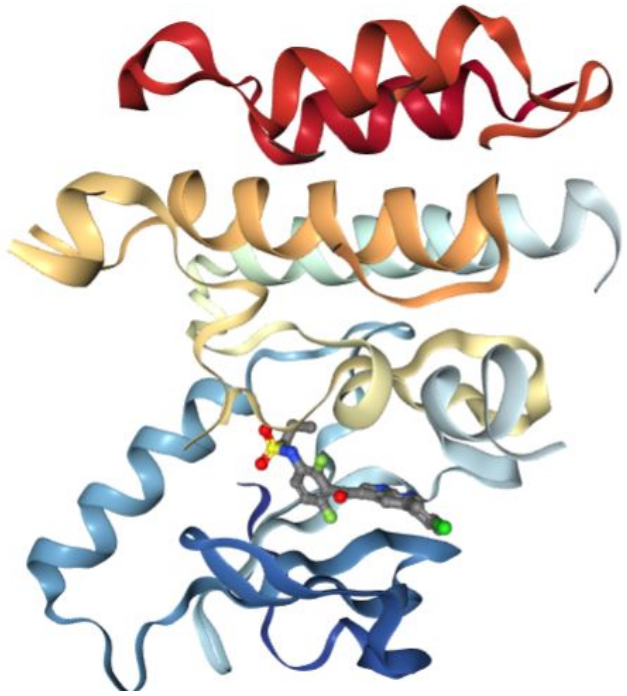
Method	Underlying physical properties	Main mathematical technique used	Advantages	Limitations
X-ray crystallography	The crystalline structure of a molecule causes a beam of incident X-rays to diffract into many specific directions.	Fourier series and Fourier transform	• Established • Broad molecular weight range • High resolution	• Crystallization • Static model
Nuclear Magnetic Resonance (NMR)	Nuclei with odd number of protons and/or neutrons in a strong constant magnetic field, when perturbed by a weak oscillating magnetic field, produce an electromagnetic signal with a frequency characteristic of the magnetic field at the nucleus.	Distance geometry (the study of matrices of distances between pairs of atoms) of and discrete differential geometry of curves	• 3D structure in solution • Dynamic study possible	• High sample purity needed • Molecular weight limit (~<40-50 kDa) • Sample preparation and computational simulation
Cryo-electron microscopy	An electron microscope using a beam of accelerated electrons (instead of protons) as a source of illumination. Samples are cooled to cryogenic temperatures and embedded in an environment of vitreous water (amorphous ice).	An inverse problem of reconstruction - the estimation of randomly rotated molecule structure from a projection with noise; Fourier transform; iterative refinement	• Easy sample preparation • Ntive-state structure • Small sample size	• Costly EM equipment • Challenging for small proteins

In silico presentation of protein structures: PDB

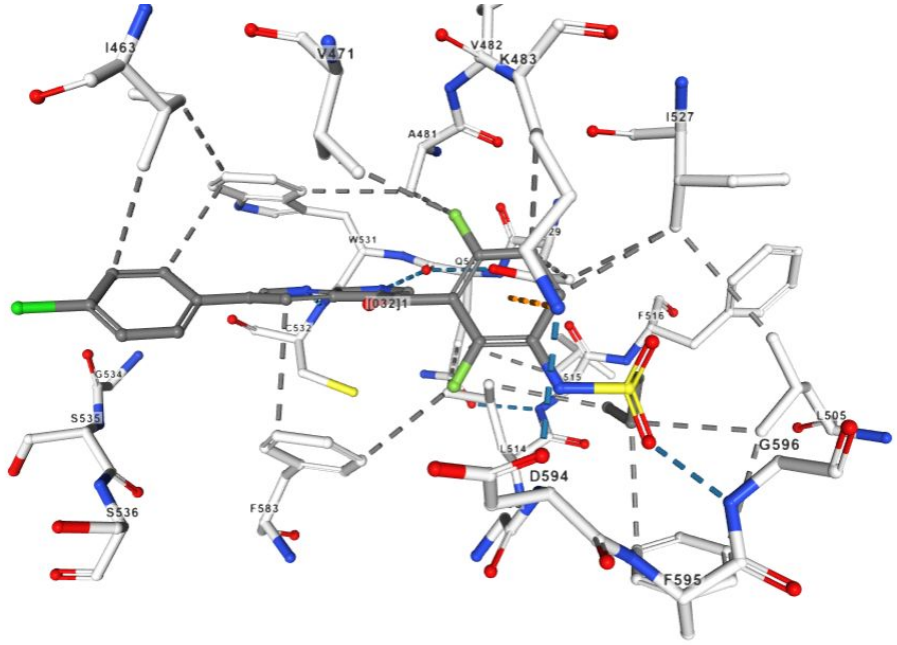
http://www.rcsb.org/3d-view/3OG7



Working with PDB files with **PyMol** from the command-line



Structural view



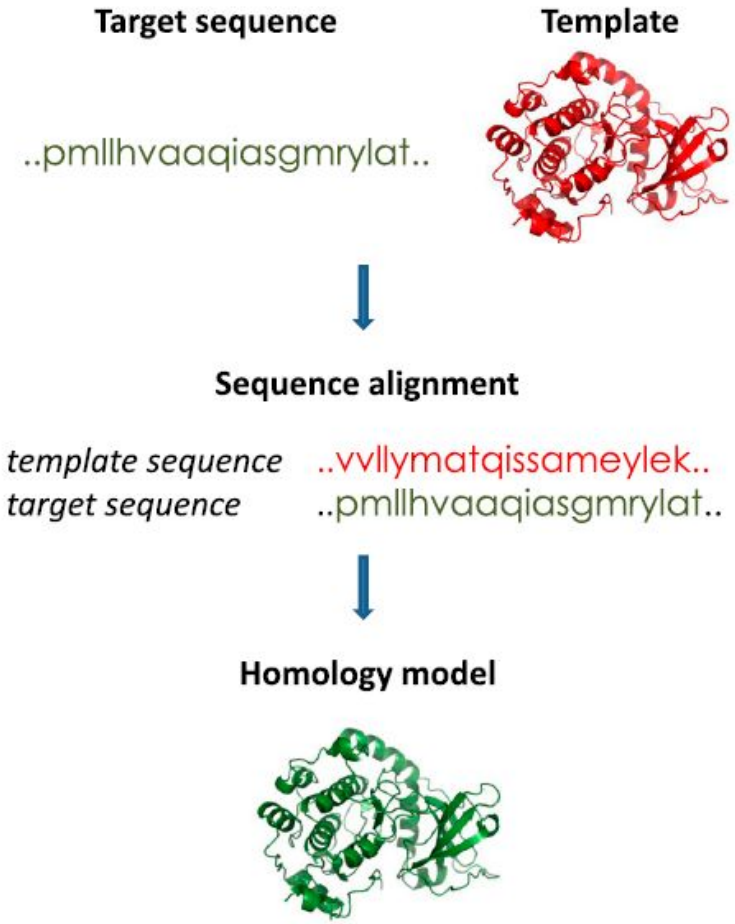
Ligand view

Balls and sticks: protein V600E and ligand (PLX4032)

Blue dashes: hydrogen bonds (<3.5 Angstrom)

Gray dashes: hydrophobic interactions (<4 Angstrom)

If no structure is available, homology model building and *in silico* prediction may help



Sliwoski, Gregory, Sandeepkumar Kothiwale, Jens Meiler, und Edward W. Lowe. „Computational Methods in Drug Discovery“. *Pharmacological Reviews* 66, Nr. 1 (1. Januar 2014): 334–95. https://doi.org/10.1124/pr.112.007336.

W296–W303 Nucleic Acids Research, 2018, Vol. 46, Web Server issue *Published online 21 May 2018*
doi: 10.1093/nar/gky427

SWISS-MODEL: homology modelling of protein structures and complexes

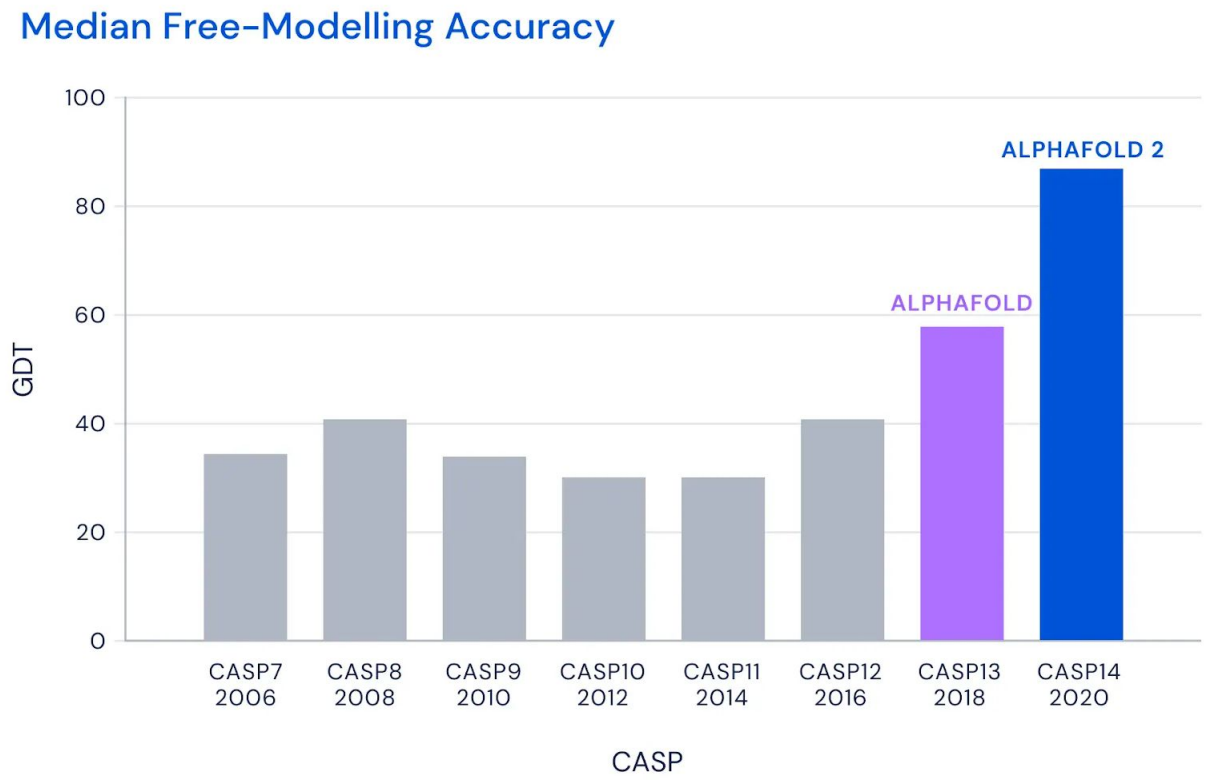
Andrew Waterhouse[1,2,†], Martino Bertoni[1,2,†], Stefan Bienert[1,2,†], Gabriel Studer[1,2,†], Gerardo Tauriello[1,2,†], Rafal Gumienny[1,2], Florian T. Heer[1,2], Tjaart A. P. de Beer[1,2], Christine Rempfer[1,2], Lorenza Bordoli[1,2], Rosalba Lepore[1,2] and Torsten Schwede[1,2,*]

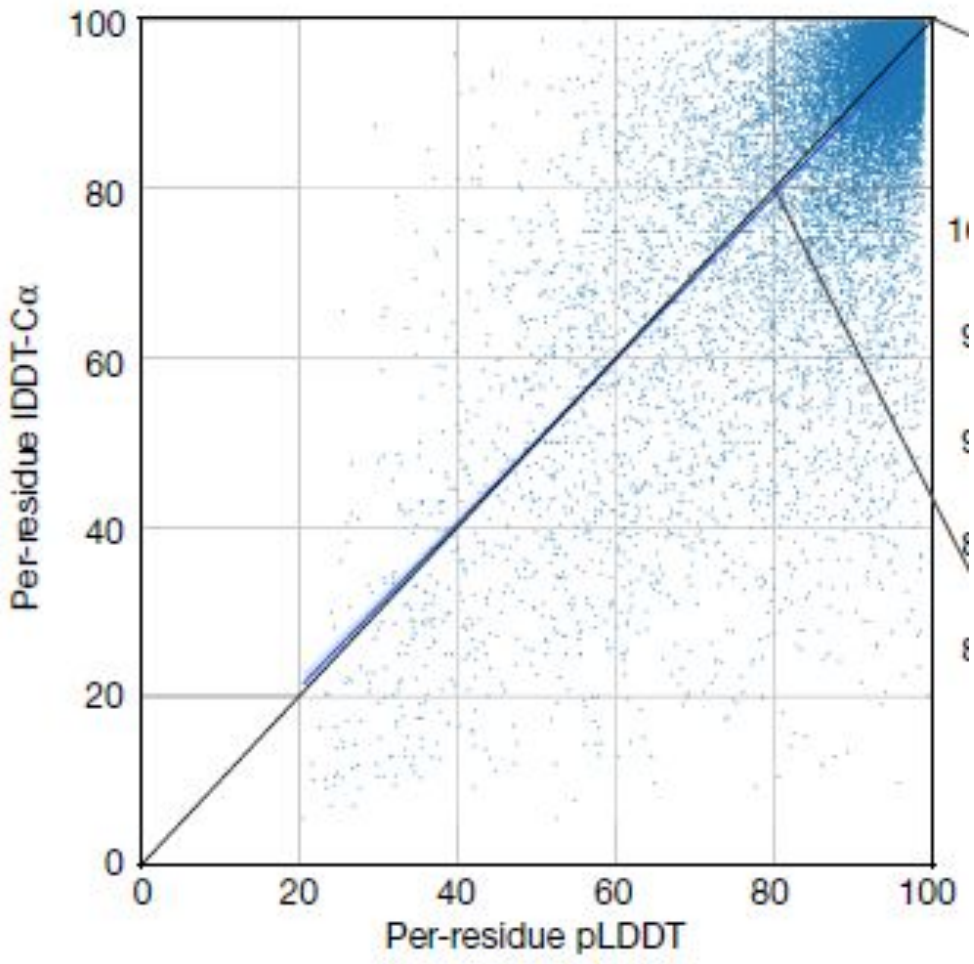
[1]Biozentrum, University of Basel, Klingelbergstrasse 50–70, CH-4056 Basel, Switzerland and [2]SIB Swiss Institute of Bioinformatics, Biozentrum, University of Basel, Klingelbergstrasse 50–70, CH-4056 Basel, Switzerland

Received February 09, 2018; Revised May 01, 2018; Editorial Decision May 02, 2018; Accepted May 07, 2018

- **Levinthal's paradox:** *It would take a protein the present age of the universe to explore all possible configurations and find the minimum energy configuration. Yet proteins fold in microseconds*.
- **CASP: Critical Assessment of Techniques for Protein Structure Prediction**
- A thought-provoking blog from Mohammed AlQuraishi: AlphaFold @ CASP13: “What just happened?”, with an informal but good overview of history of protein structure prediction, and his indictment (criminal accusations) of both academia and pharma.
- By 2021 AlphaFold2 and RoseTTAfold reach experiment-level accuracy in some predictions of protein static structure

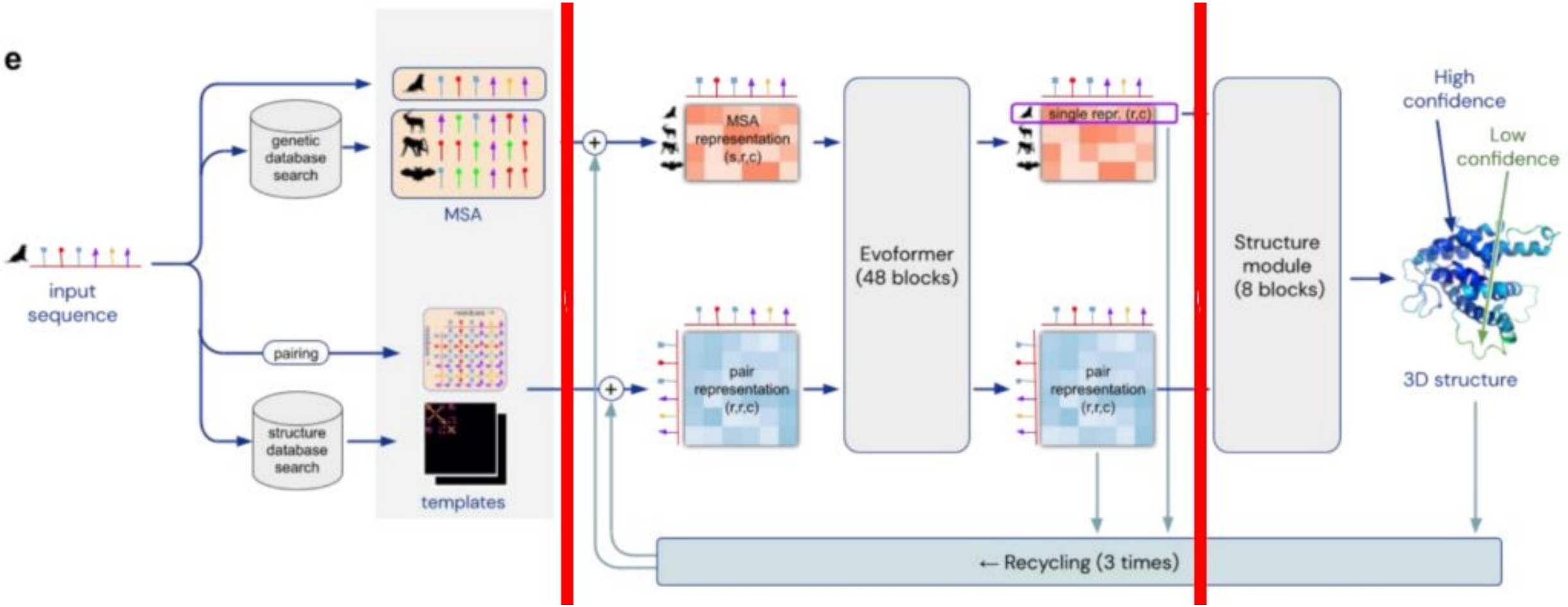
AlphaFold2 reaches prediction accuracy comparable to experimental approaches





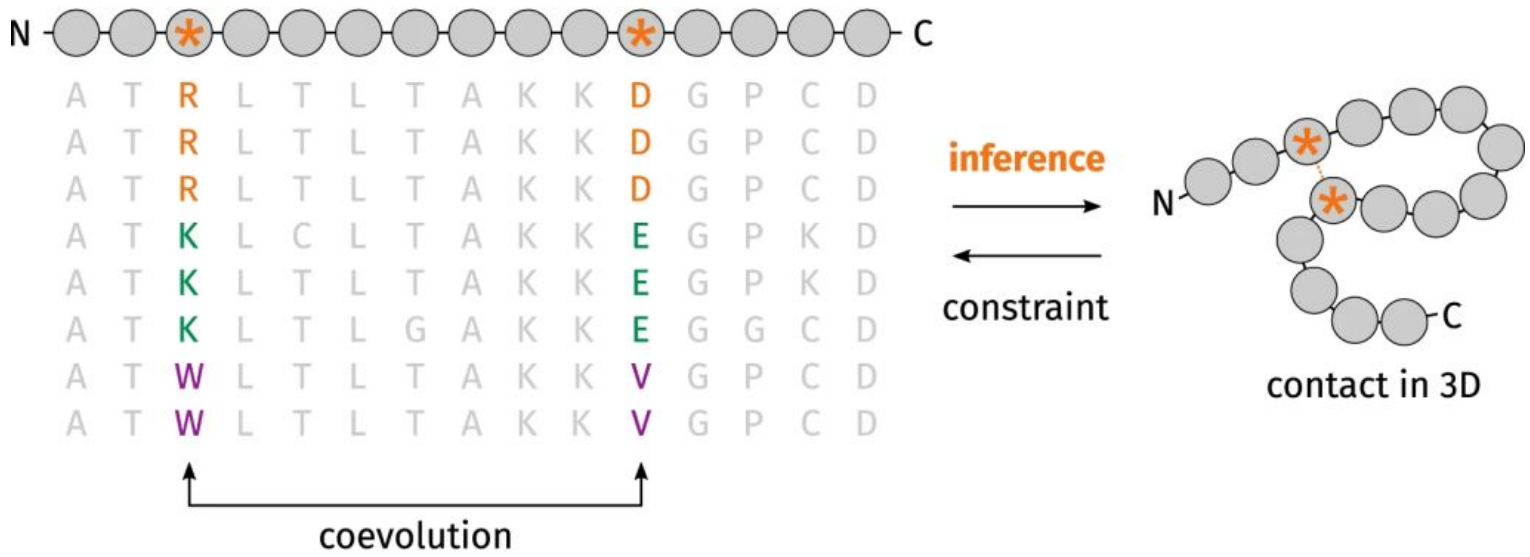
pLDDT: Predicted local distance difference test, estimating how the prediction differs from the experimental structure based on the local distance difference test (C-alpha, lDDT)

AlphaFold2 uses co-evolution of residues, determined structures, and neural networks to achieve the high performance



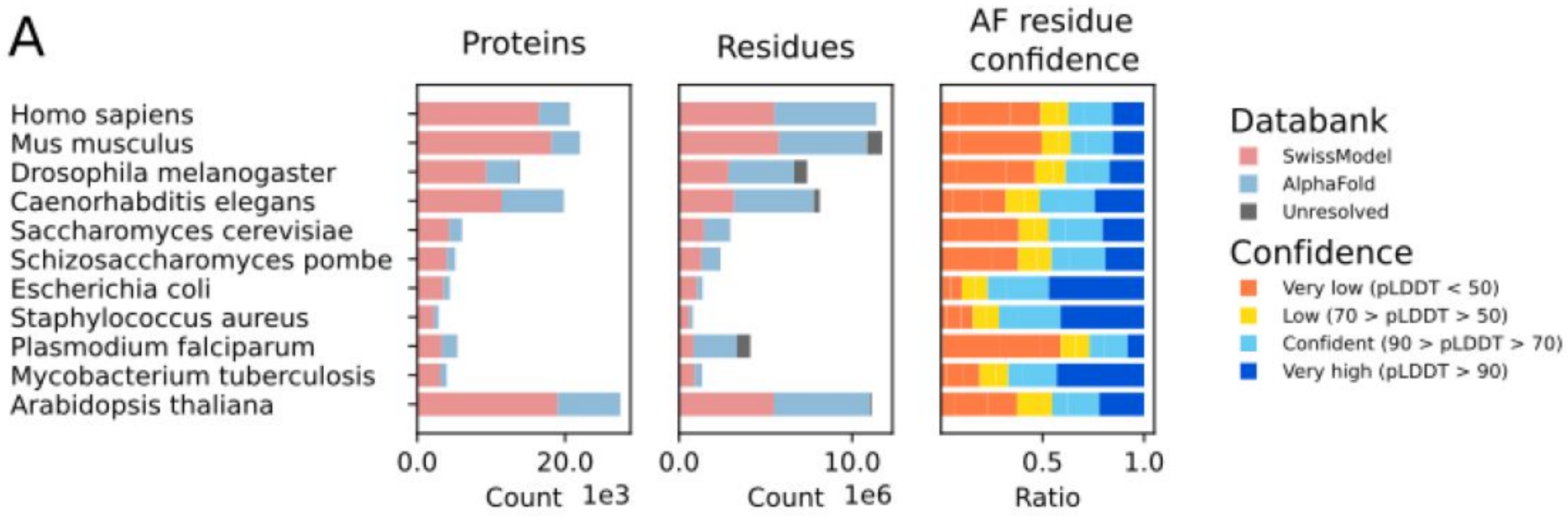
- Jumpe et al. “Highly Accurate Protein Structure Prediction with AlphaFold.” Nature 596, no. 7873 (August 2021): 583–89. https://doi.org/10.1038/s41586-021-03819-2.
- A blog post that explains how AlphaFold2 works: blogpig.com

The key idea (beyond using 2D and 3D structure mapping): learning from evolutionary constraints

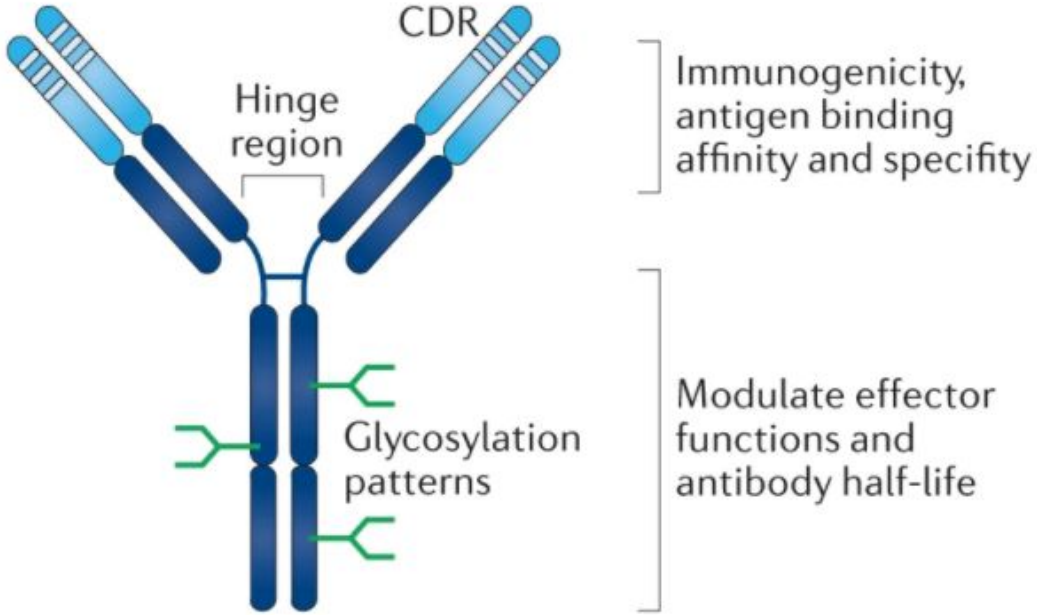


Marks, Debora S., Lucy J. Colwell, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. “Protein 3D Structure Computed from Evolutionary Sequence Variation.” PLOS ONE 6, no. 12 (December 7, 2011): e28766. https://doi.org/10.1371/journal.pone.0028766.

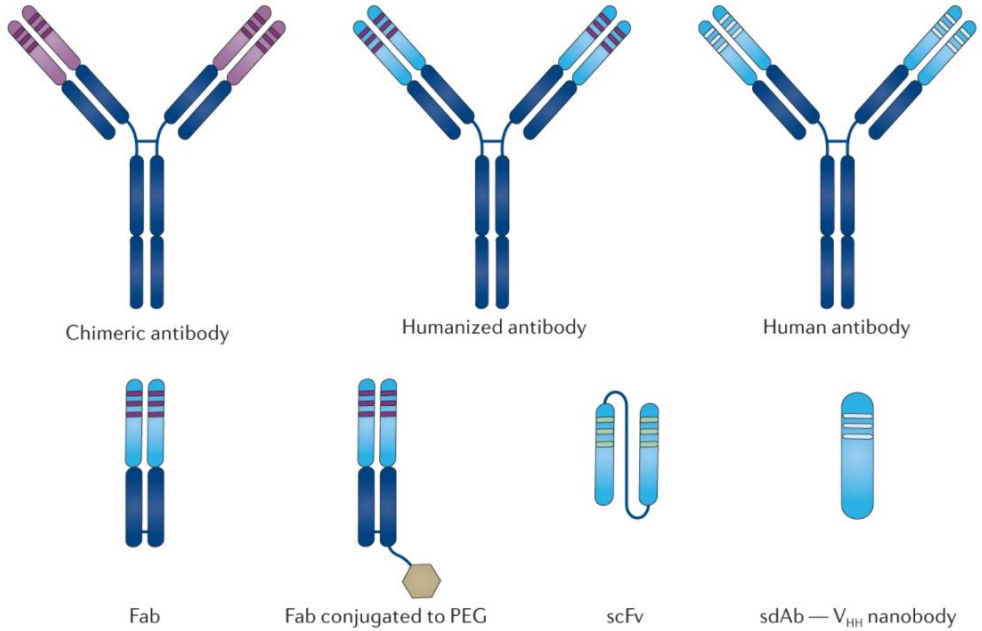
AlphaFold2 & RoseTTAfold extend our understanding of protein biology, while their impact on drug discovery remains to be seen

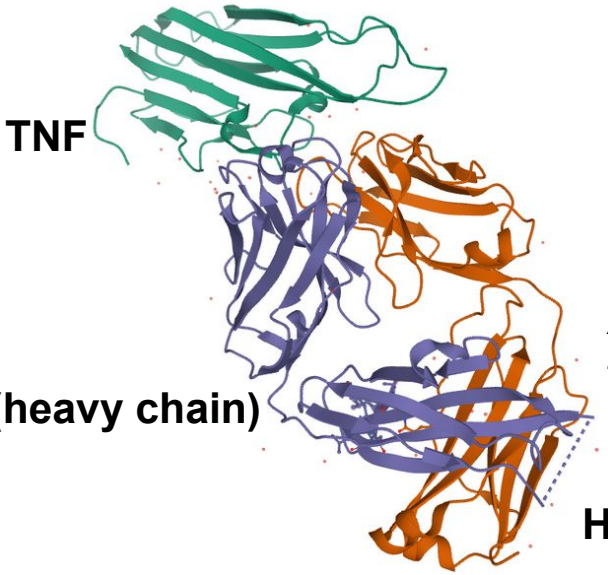


Antibodies are also proteins



Attwood, Misty M., Jörgen Jonsson, Mathias Rask-Andersen, and Helgi B. Schiöth. 2020. “Soluble Ligands as Drug Targets.” Nature Reviews Drug Discovery 19 (10): 695–710. https://doi.org/10.1038/s41573-020-0078-4.





PDB 3WD5, Crystal structure of *TNF-alpha* in complex with *Adalimumab* (*Humira*) Fab fragment, PubMed: 23943614

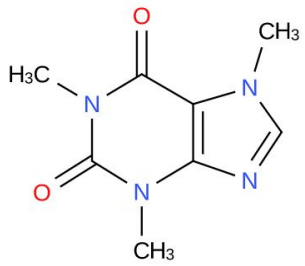
ChEMBL as information source of small molecules

Nomenclature

caffeine
1,3,7-trimethylxanthine
methyltheobromine

Bioactivity

Affinity to human proteins and drug targets



Chemical data

Formula: $C_8H_{10}N_4O_2$
Charge: 0
Mass: 194.19

Database Xrefs

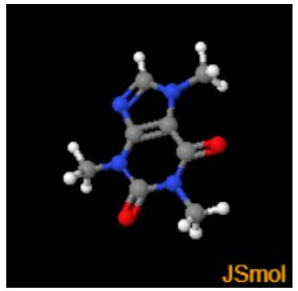
PubChem: CID2519
BindingDB: 1849

Chemical Informatics

***InChI**=1/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3*

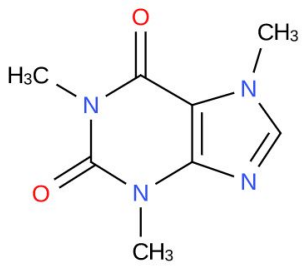
***SMILES**: CN1C(=O)N(C)c2ncn(C)c2C1=O*

Visualisation



A subset of available information from EBI ChEBI/ChEMBL, inspired by EBI's roadshow *Small Molecules in Bioinformatics*

Representation of small molecules



Molfile: </> View Raw | Download | Editor | Copy

Canonical SMILES: CN1C(=O)N(C)c2ncn(C)c2C1=O

Standard InChI: InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

Standard InChI Key: RYYVLZVUVIJVGH-UHFFFAOYSA-N

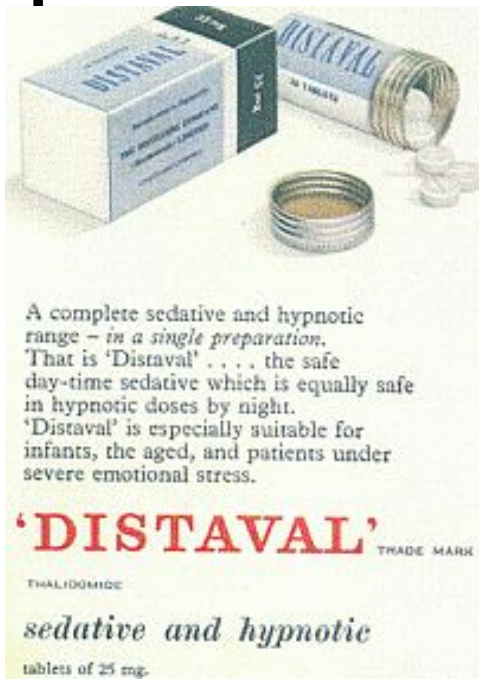
```
CHEMBL113

SciTegic12231509382D

14 15 0 0 0 0 999 V2000
-1.1875 -9.6542 0.0000 C 0 0
-1.1875 -8.9625 0.0000 C 0 0
-1.8125 -10.0292 0.0000 N 0 0
-2.4167 -8.9625 0.0000 N 0 0
-2.4167 -9.6542 0.0000 C 0 0
-1.8125 -8.6000 0.0000 C 0 0
-0.5000 -9.8917 0.0000 N 0 0
-0.5000 -8.7625 0.0000 N 0 0
-0.1125 -9.3042 0.0000 C 0 0
-3.0250 -10.0375 0.0000 O 0 0
-1.8125 -7.8917 0.0000 O 0 0
-1.8125 -10.7417 0.0000 C 0 0
-3.0250 -8.6000 0.0000 C 0 0
-0.2917 -8.0750 0.0000 C 0 0
2 1 2 0
3 1 1 0
4 5 1 0
5 3 1 0
6 2 1 0
7 1 1 0
8 2 1 0
9 7 2 0
10 5 2 0
11 6 2 0
12 3 1 0
13 4 1 0
```

- Simplified Molecular-Input Line-Entry System (SMILES)
- IUPAC International Chemical Identifier (InChI)
- InChiKey: a 27-character, hash version of InChI
- Molfile: a type of chemical table files

The tragedy of thalidomide and the importance of representation

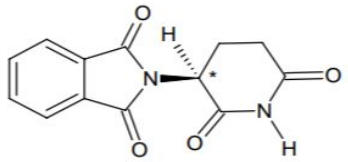


(1957)

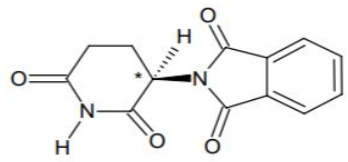
Frances Oldham Kelsey received the President's Award for Distinguished Federal Civilian Service from President John F. Kennedy, 1962

Canonic SMILES of thalidomide

C1CC(=O)NC(=O)C1N2C(=O)C3=CC=CC=C3C2=O



(-)(*S*)-thalidomide



(+)(*R*)-thalidomide

Isomeric SMILES of (-)(S)-thalidomide

C1CC(=O)NC(=O)[C@H]1N2C(=O)C3=CC=CC=C3C2=O

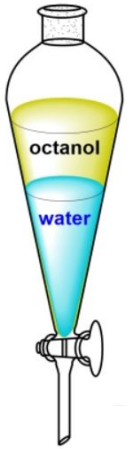
Isomeric SMILES of (+)(R)-thalidomide

C1CC(=O)NC(=O)[C@@H]1N2C(=O)C3=CC=CC=C3C2=O

I thank Manuela Jacklin for her help preparing this slide.

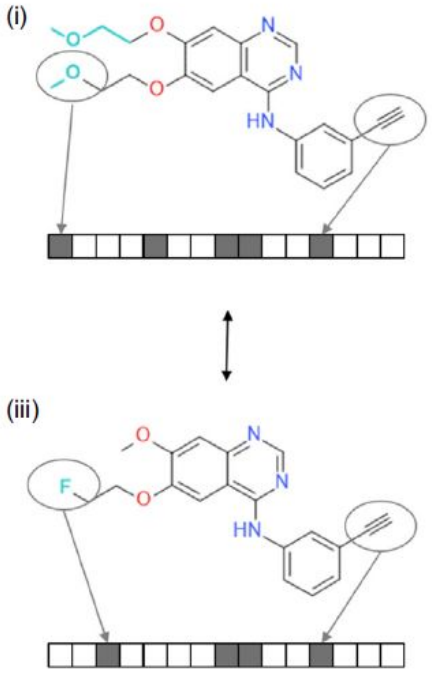
Molecular descriptors: numeric values that describe chemical molecules.

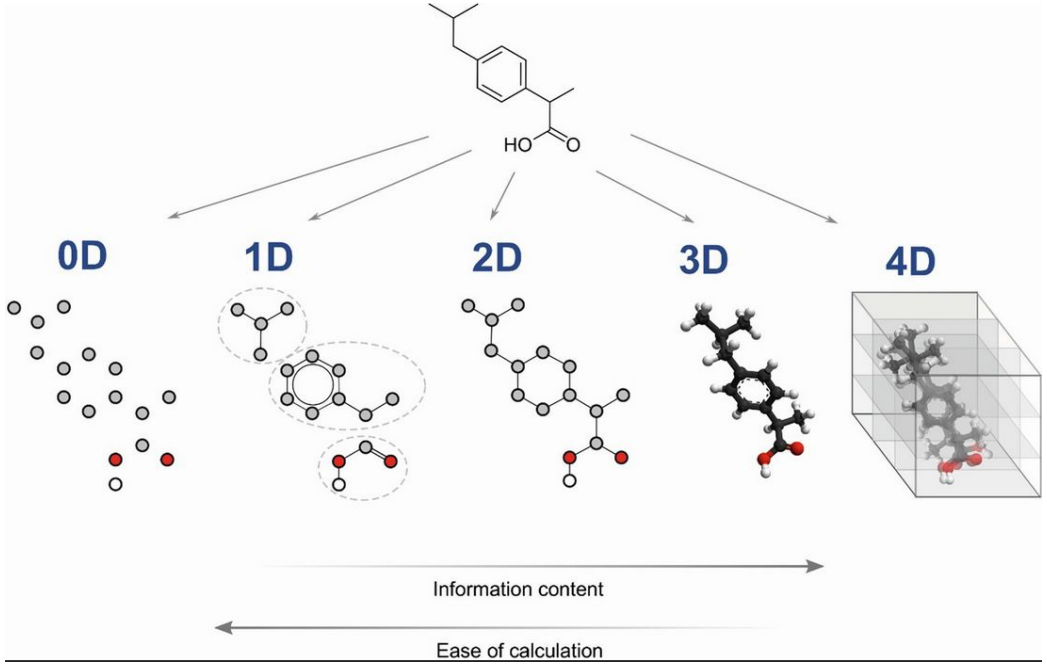
In contrast to symbolic representations, molecular descriptors enable **quantification of molecular properties**. It allows mathematical operations and statistical analysis that associate biophysical/biochemical properties with molecule structures.



$$\log P_{\mathrm{oct/wat}} = \log\left(\frac{[\mathrm{solute}]^{\text{un-ionized}}_{\text{octanol}}}{[\mathrm{solute}]^{\text{un-ionized}}_{\text{water}}}\right)$$

logP is an experimental molecular descriptor. Calculated version (cLogP) exists as well.





- Atom count
- Molecular weight
- Sum of atomic properties

- Fragment counts, *e.g.* # of -OH
- Fingerprints

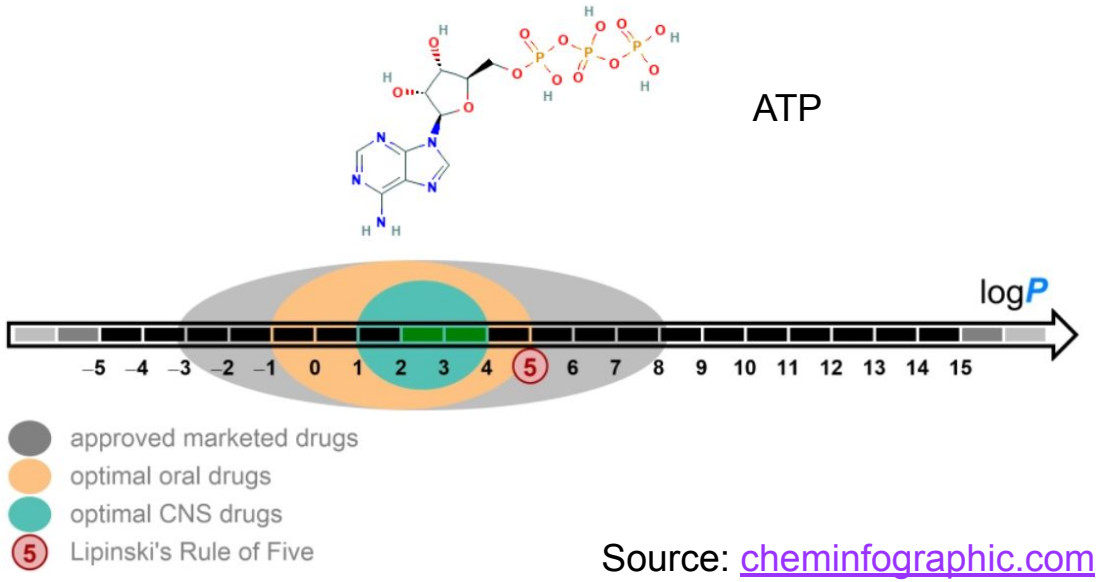
- Topological descriptors, *e.g.* the Wiener Index, sum of lengths of the shortest paths between all non-H atoms

- Geometrical
- Atomic coordinates
- Energy grid

- Combination of atomic coordinates and sampling of possible conformations

Lipinski's Rule of Five of small-molecule drugs

- **HBD<=5:** No more than **5 hydrogen-bond donors,** *e.g.* the total number of nitrogen–hydrogen and oxygen–hydrogen bonds.
- **HBA<=10**: No more than **10 hydrogen-bond acceptors**, *e.g.* all nitrogen or oxygen atoms
- **MW<500**: A **molecular weight** less than **500 Daltons, or** 500 g/mol. Reference: ATP has a molecular mass of ~507.
- **logP<=5**: An **octanol-water partition coefficient (log P)** that does not exceed **5.** (10-based)



Source: cheminfographic.com

Table 1. New FDA Approvals (2014 to Present)a of Oral bRo5 Drugs

drug	year approved	therapeutic area	MW	cLogP	HBD	N+O
velpatasvir	2016	HCV	883.02	2.5	4	16
venetoclax	2016	oncology	868.44	10.4	3	14
elbasvir	2016	HCV	882.0	2.6	4	16
grazoprevir	2016	HCV	766.90	−2.0	3	15
cobimetinib	2015	oncology	531.31	5.2	3	5
daclatasvir	2015	HCV	738.88	1.3	4	14
edoxaban	2015	cardiovascular	548.06	−0.9	3	11
ombitasvir	2014	HCV	894.13	1.3	4	15
paritaprevir	2014	HCV	765.89	1.1	3	14
netupitant	2014	nausea from chemotherapy	578.59	6.8	0	5
ledipasvir	2014	HCV	889.00	0.9	4	14
ceritinib	2014	oncology	558.14	6.5	3	8

DeGoey, *et al.*. 2018. "Beyond the Rule of 5: Lessons Learned from AbbVie's Drugs and Compound Collection." Journal of Medicinal Chemistry 61 (7): 2636–51.

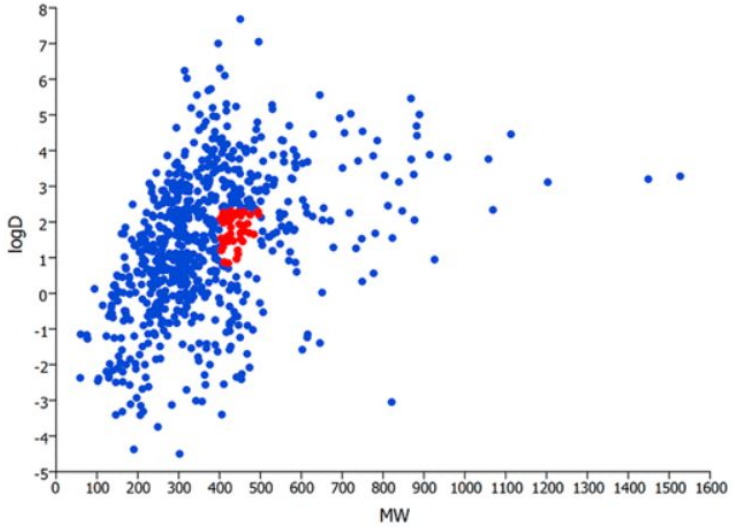
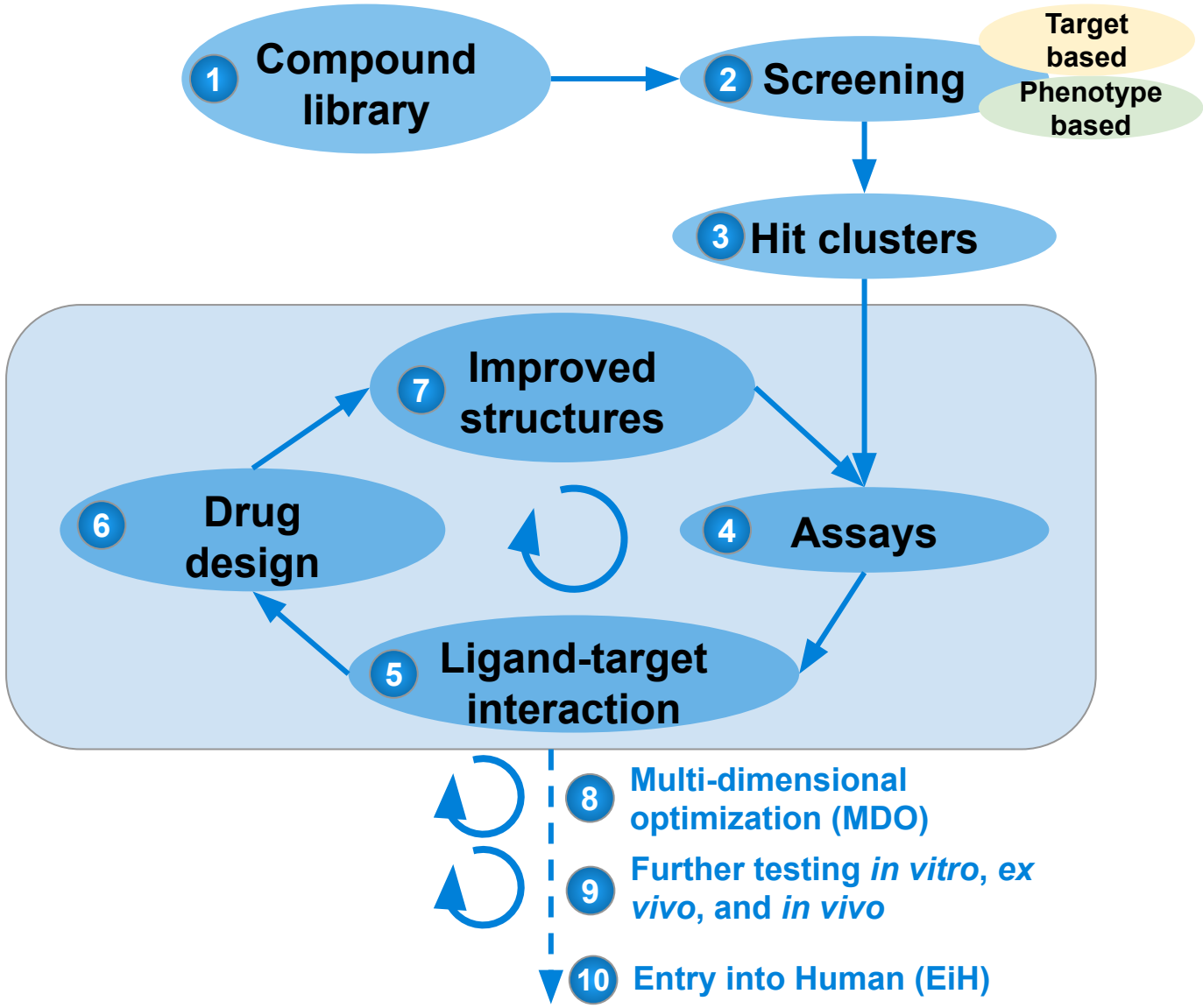


Figure 7: Plot of MW vs cLogD of FDA approved oral drugs. Red points: 'high probability area' supposed by (questionable) data analysis. Shultz, Michael D. 2019. "Two Decades under the Influence of the Rule of Five and the Changing Properties of Approved Oral Drugs." Journal of Medicinal Chemistry 62 (4): 1701–14.

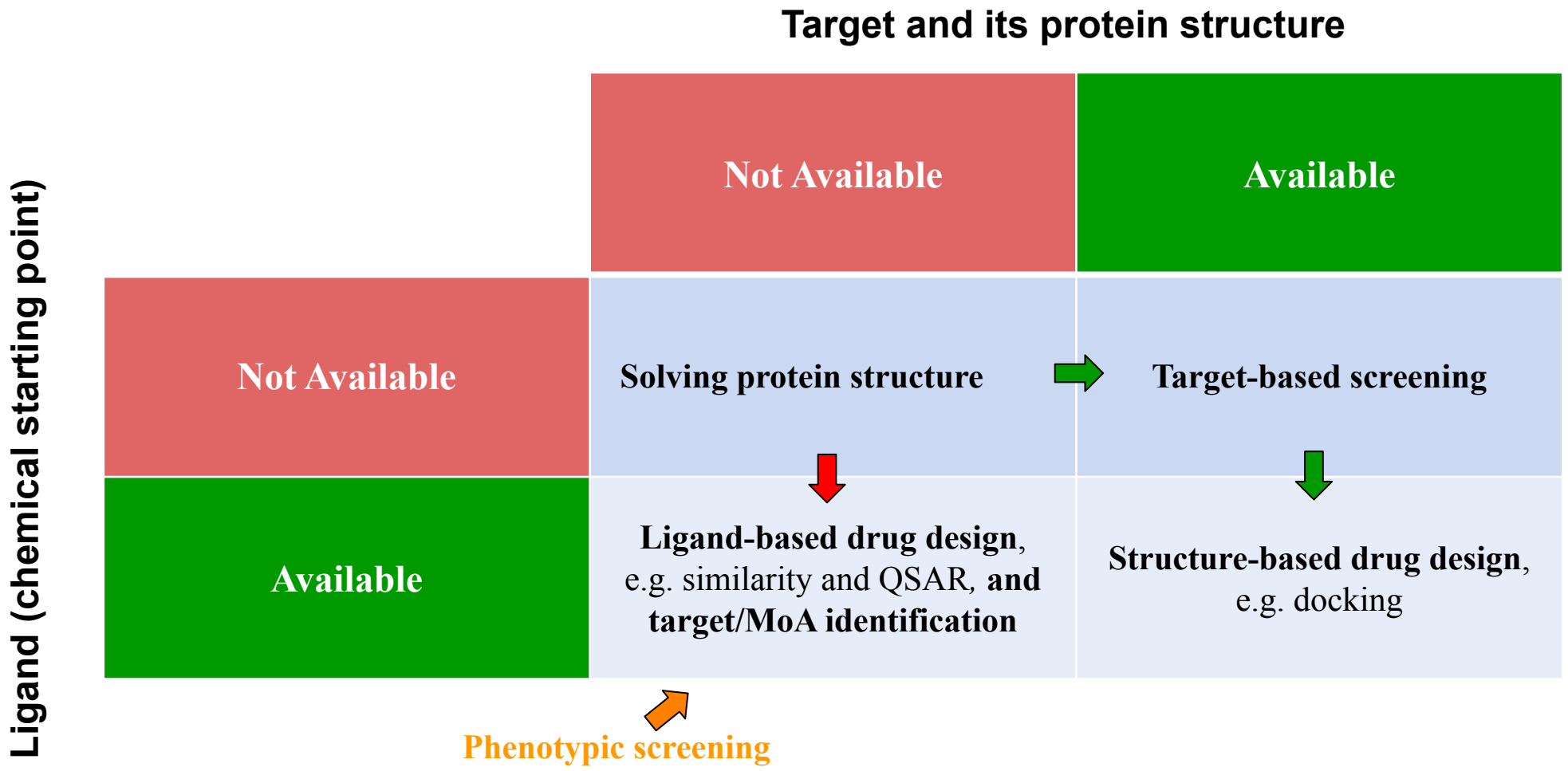
Workflow in a typical drug-discovery program

1. Compound library construction;
2. Screening compounds with ***bioassays***, or ***assays***, which determine potency of a chemical by its effect on biological entities: proteins, cells, *etc*;
3. Hit identification and clustering;
4. More assays, complementary to the assays used in the screening, maybe of lower throughput but more biologically relevant;
5. Analysis of ligand-target interactions, for instance by getting the co-structure of both protein (primary target, and off-targets if necessary) and the hit;
6. *Drug design,* namely to modify the structure of the drug candidate;
7. Analog synthesis and testing (back to step 4);
8. Multidimensional Optimization (MDO), with the goal to optimize potency, selectivity, safety, bioavailability, *etc;*
9. Further *in vitro*, *ex vivo*, and *in vivo* testing, and preclinical development;
10. Entry into human (Phase 0 or phase 1 clinical trial).



A schematic presentation of structure-based drug discovery

Ligand-based and structure-based drug design



QSAR= quantitative structure activity relationship; MoA= mechanism of action, or mode of action

Conclusions

- **A successful drug must possess many properties**, among others potency, selectivity, physico-chemical/ADME properties, and safety profiles. These need to be considered in the screening process.
- **Drug screening means to identify drug candidates (small molecules, antibodies, oligonucleotides, *etc.*) to modulate target function.** We need to understand the target (mostly proteins), the ligand (small molecules, antibodies, oligonucleotides), and the interaction between them (binding mode, affinity, consequence of modulation, *etc.*).
- **Protein structures can be determined experimentally (X-ray, NMR, CyroEM) or by *in silico* prediction (homology modelling, AlphaFold2/RoseTTAfold).**
- **Small molecules can be presented by symbols and by molecular descriptors.**

Offline activities

- **Anonymous post-lecture survey of Lecture #5**: https://forms.gle/BVdgcSbyJYmG8SSSA
- **Required reading:** selected pages of *Evaluation of the Biological Activity of Compounds: Techniques and Mechanism of Action Studies* by Dougall and Unitt and answer questions (see the next slide). Please submit your results to the Google Form.
- **Optional reading** based on your interests**:**
  - [Machine learning and drug discovery] Mullard, Asher. “What Does AlphaFold Mean for Drug Discovery?” Nature Reviews Drug Discovery 20, no. 10 (September 14, 2021): 725–27. https://doi.org/10.1038/d41573-021-00161-0.
  - [Mathematics and structural biology] Mathematical techniques used in biophysics by J. R. Quine.