Determination of Biological Structures

Workshop on Computer-Aided Drug Discovery
Graz, 5.9.2022

Karl Gruber
Institute of Molecular Biosciences, University of Graz

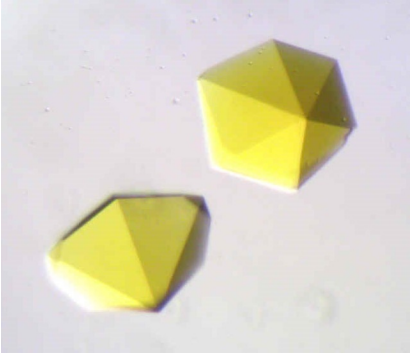
Experimental Determination of Molecular Biological Structures and their Validation

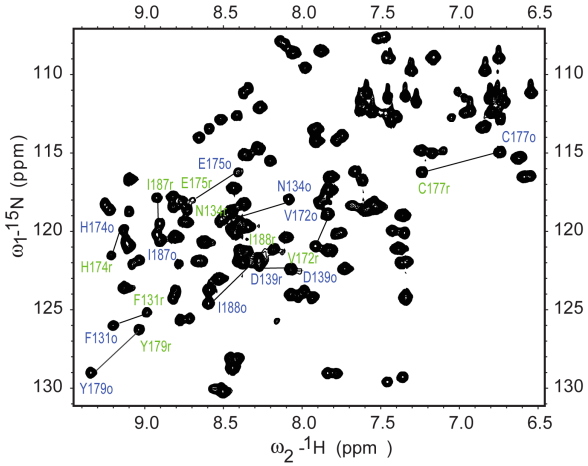
Workshop on Computer-Aided Drug Discovery
Graz, 5.9.2022

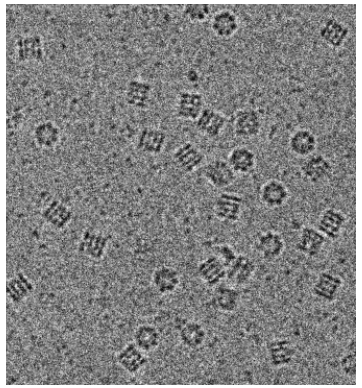
Karl Gruber
Institute of Molecular Biosciences, University of Graz

Structure determination of biological macromolecules

- X-ray crystallography
- NMR-spectroscopy
- Cryo-electron microscopy







Macromolecular structure databases

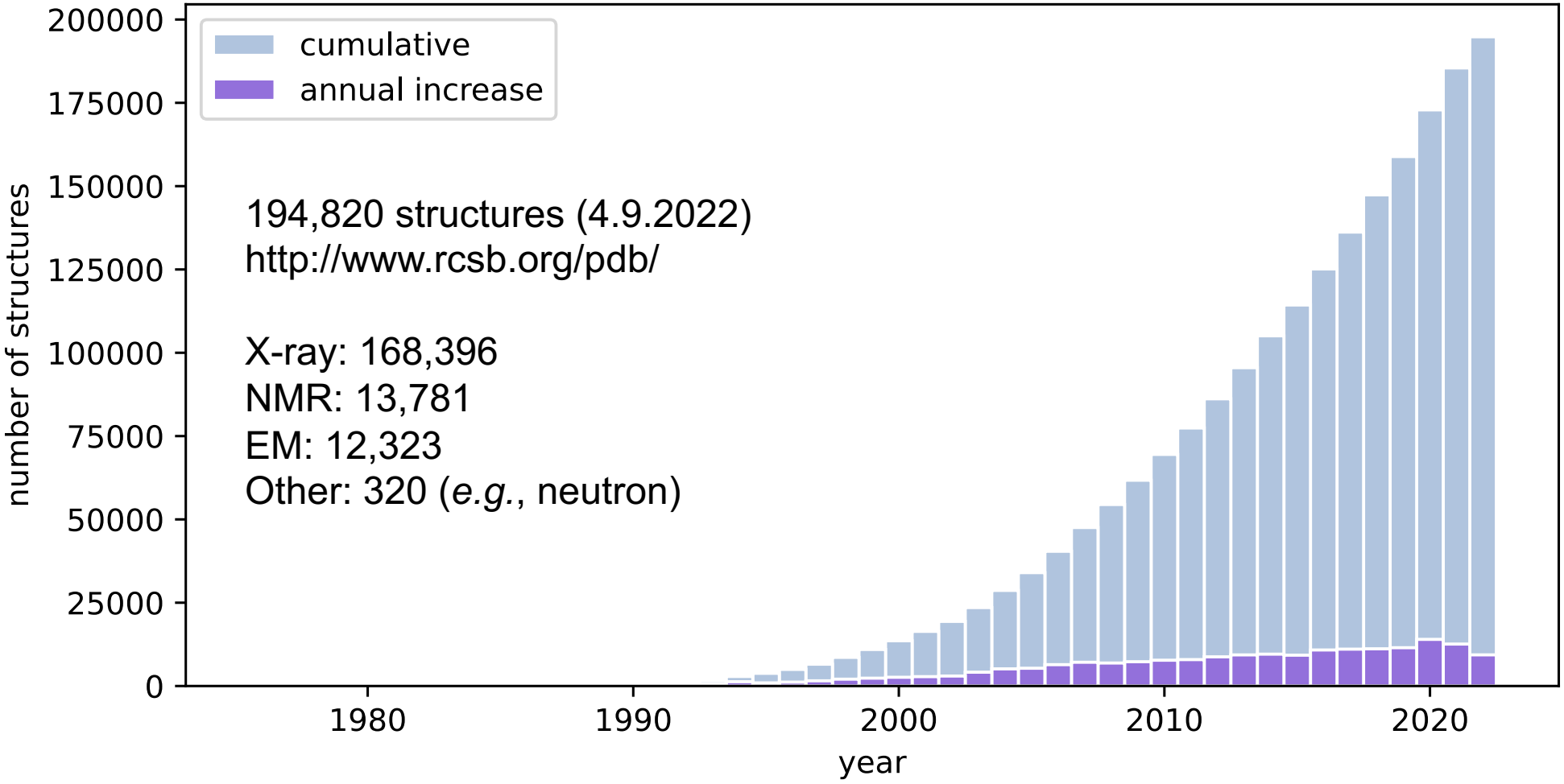
RCSB PDB
PROTEIN DATA BANK

http://www.rcsb.org

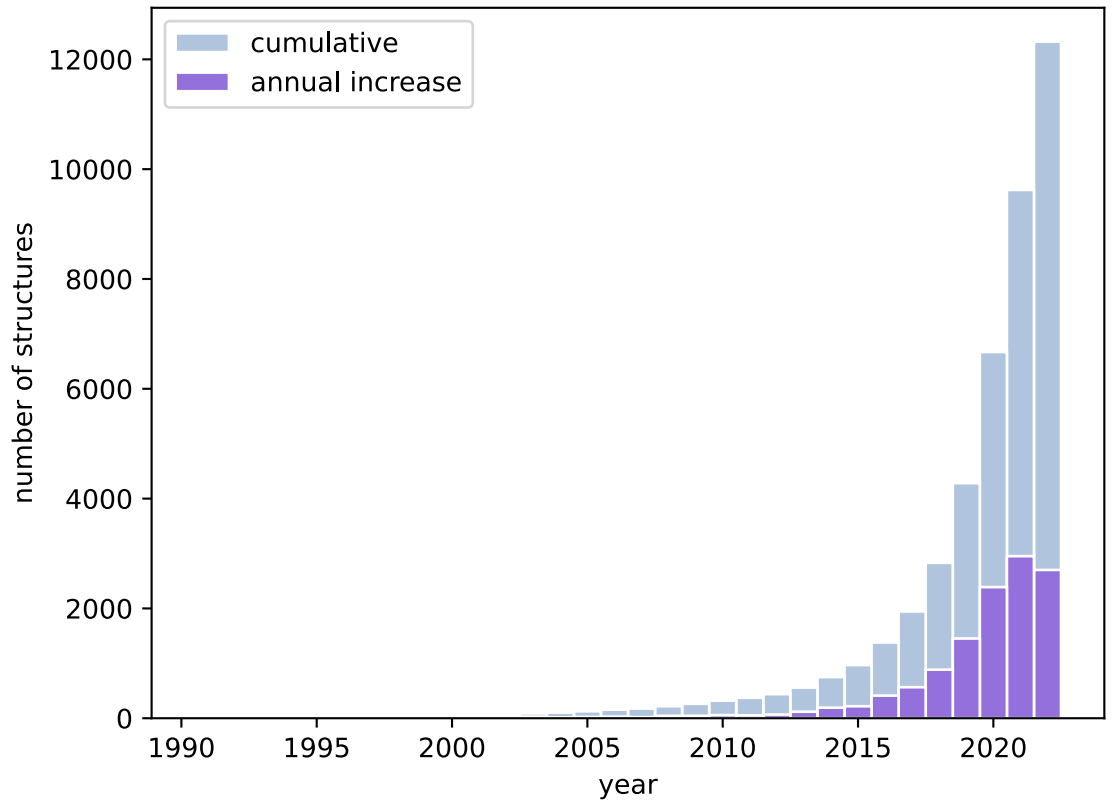
PDBe
Protein Data Bank in Europe

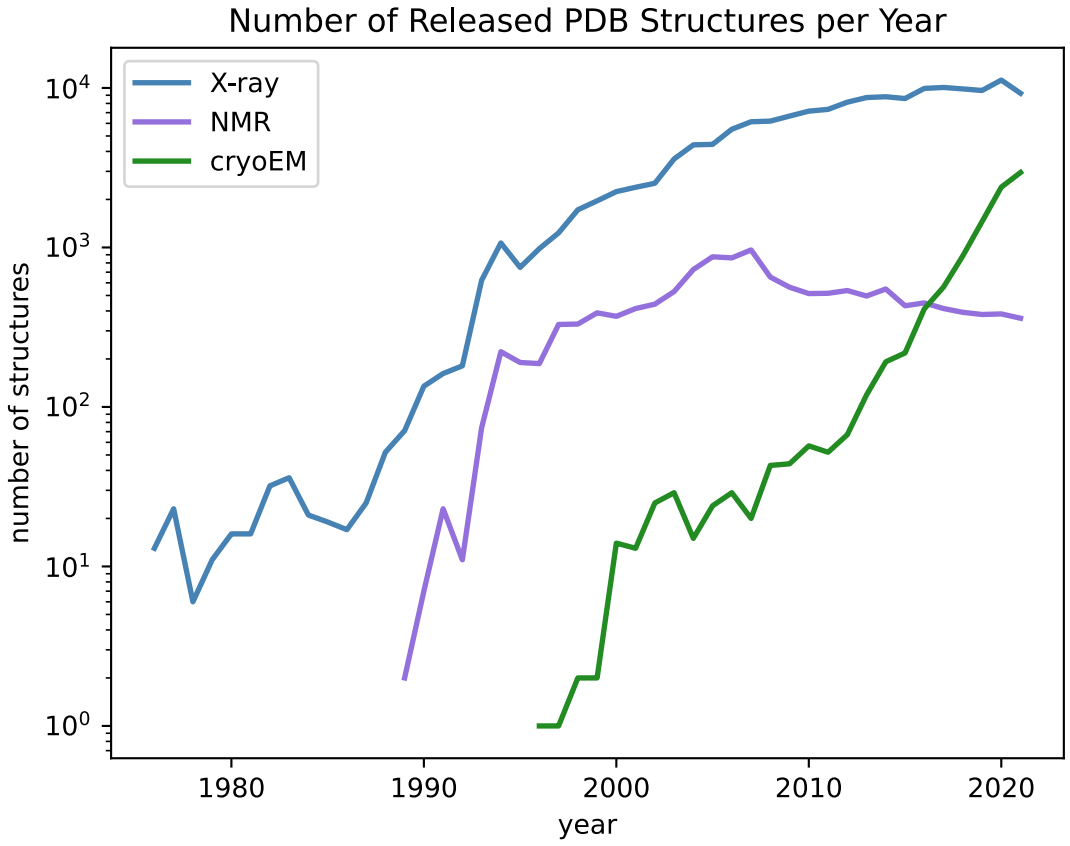
http://www.ebi.ac.uk/pdbe/

Protein Data Bank (PDB)



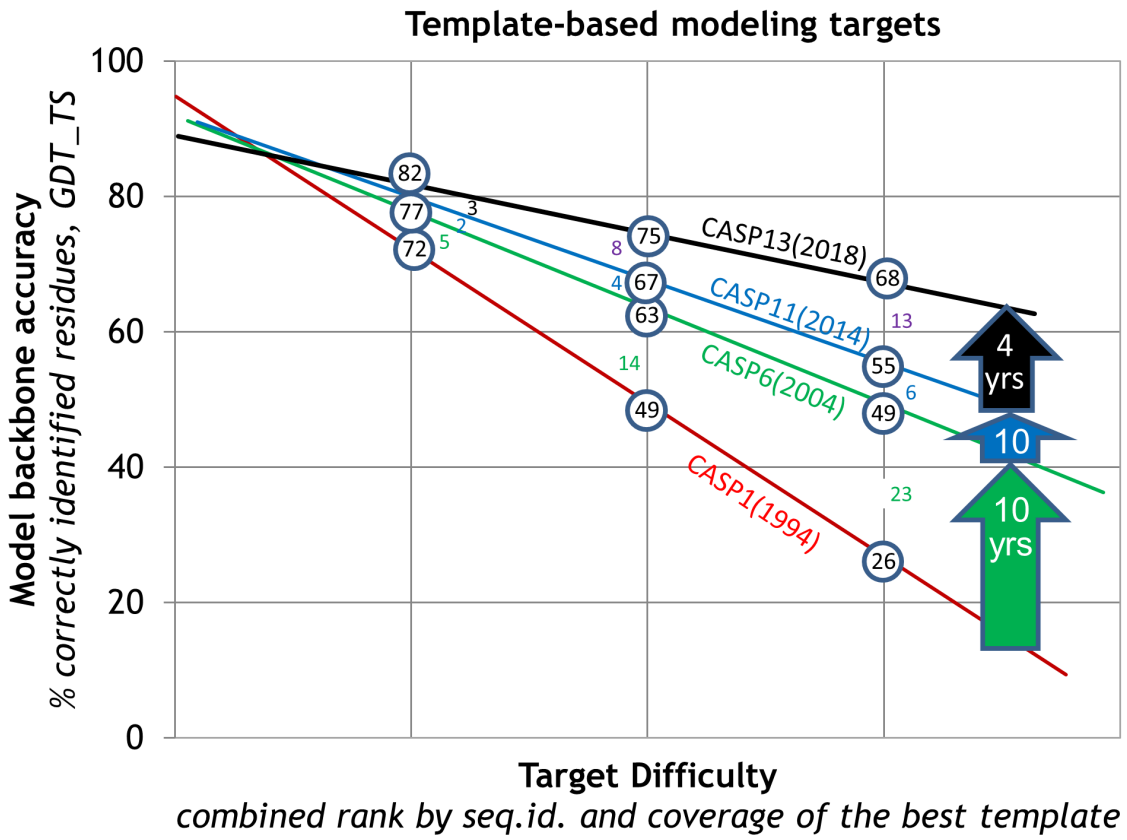
Cryo-electron microscopy (cryoEM)



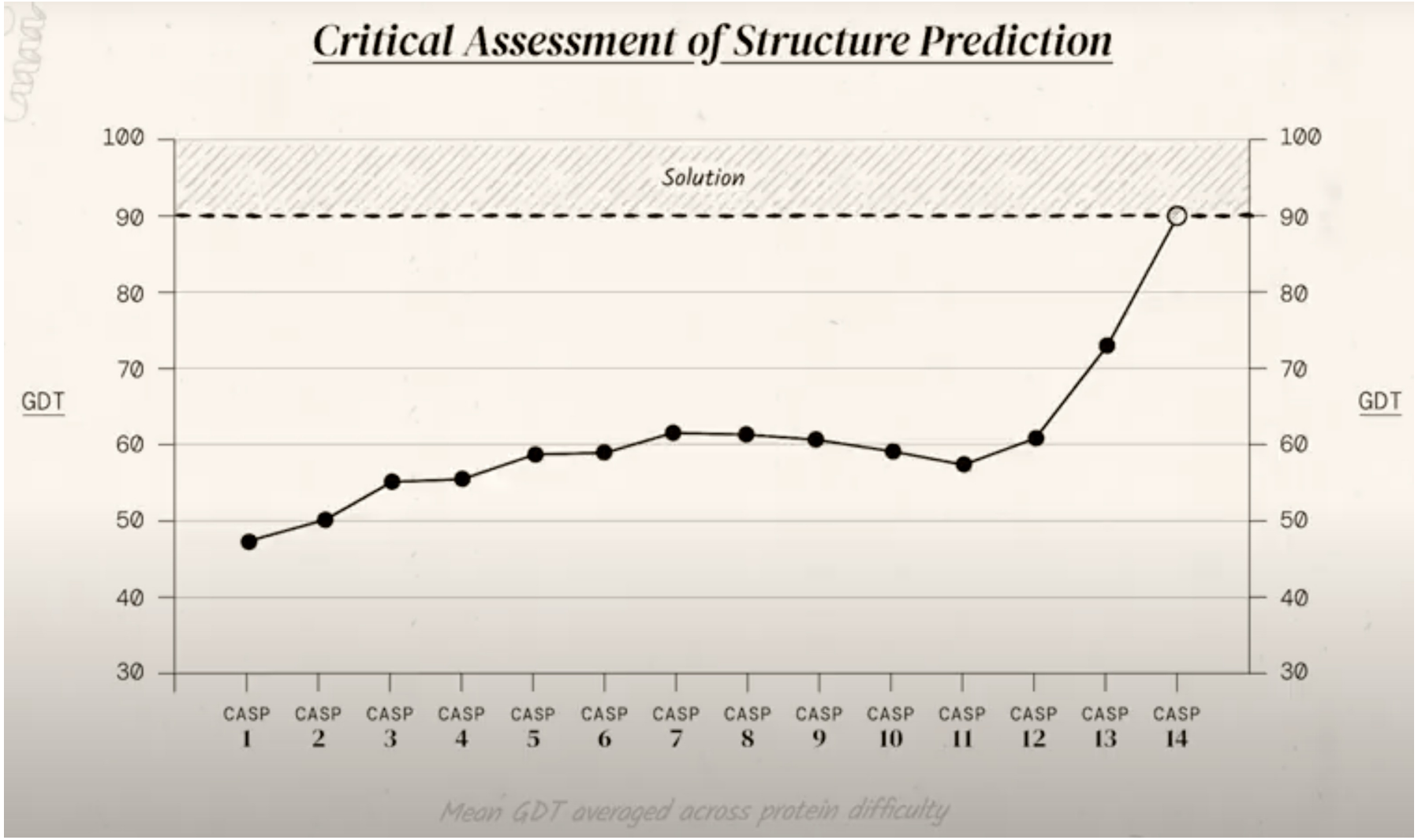


CASP – Critical assessment of protein structure prediction

- biennial competition for protein structure prediction taking place since 1994
- **targets:** structures solved experimentally but not yet published
- **contestants:** expert groups as well as automatic prediction servers
- benchmark of the quality of prediction-algorithms
- upcoming CASP-15, Dec. 2022



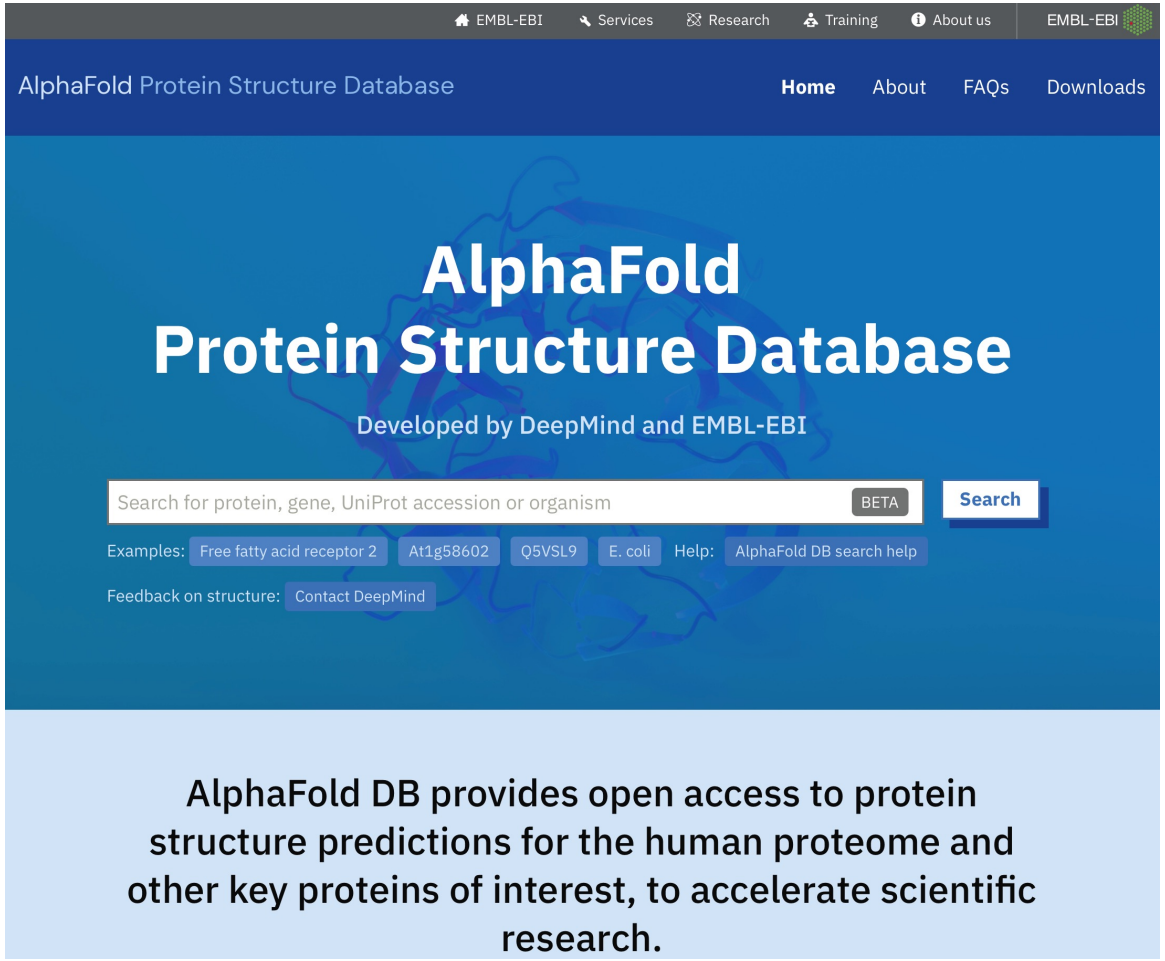
progress in the area of template-based modeling



AlphaFold: The making of a scientific breakthrough

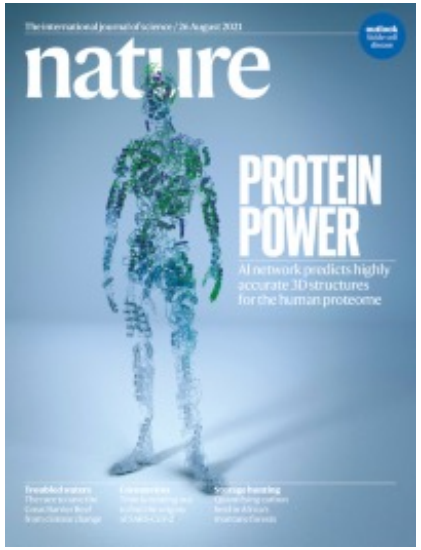
https://www.youtube.com/watch?v=gg7WjuFs8F4

Database of AlphaFold models



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

- precomputed models of protein structures
- whole proteomes: human, mouse,…
- complete Swissprot database (curated protein sequences)
- quality parameters useful for validation



Science magazine:
Breakthrough of the Year 2021

https://www.science.org/content/article/breakthrough-2021

README.md

ColabFold

New Updates

```
+ 11Mar2022 We use in default AlphaFold-multimer-v2 weights for complex modeling.
+           We also offer the old complex modes "AlphaFold-ptm" or "AlphaFold-multimer-v1"
+ 04Mar2022 ColabFold now uses a much more powerful server for MSAs and searches through the Cola
+           Please let us know if you observe any issues.
+ 26Jan2022 AlphaFold2_mmseqs2, AlphaFold2_batch and colabfold_batch's multimer complexes predict
+           now in default reranked by iptmscore*0.8+ptmscore*0.2 instead of ptmscore
```

Making Protein folding accessible to all via Google Colab!

Notebooks	monomers	complexes	mmseqs2	jackhmmer	templates
AlphaFold2_mmseqs2	Yes	Yes	Yes	No	Yes
AlphaFold2_batch	Yes	Yes	Yes	No	Yes
RoseTTAFold	Yes	No	Yes	No	No
AlphaFold2 (from Deepmind)	Yes	Yes	No	Yes	No

Run AlphaFold yourself

- GoogleColab notebooks: based on Jupyter Notebooks, Python
- runs on the Google cloud (free service, “restrictions apply”)
- can be installed locally (requires larger GPU)

https://github.com/sokrypton/ColabFold

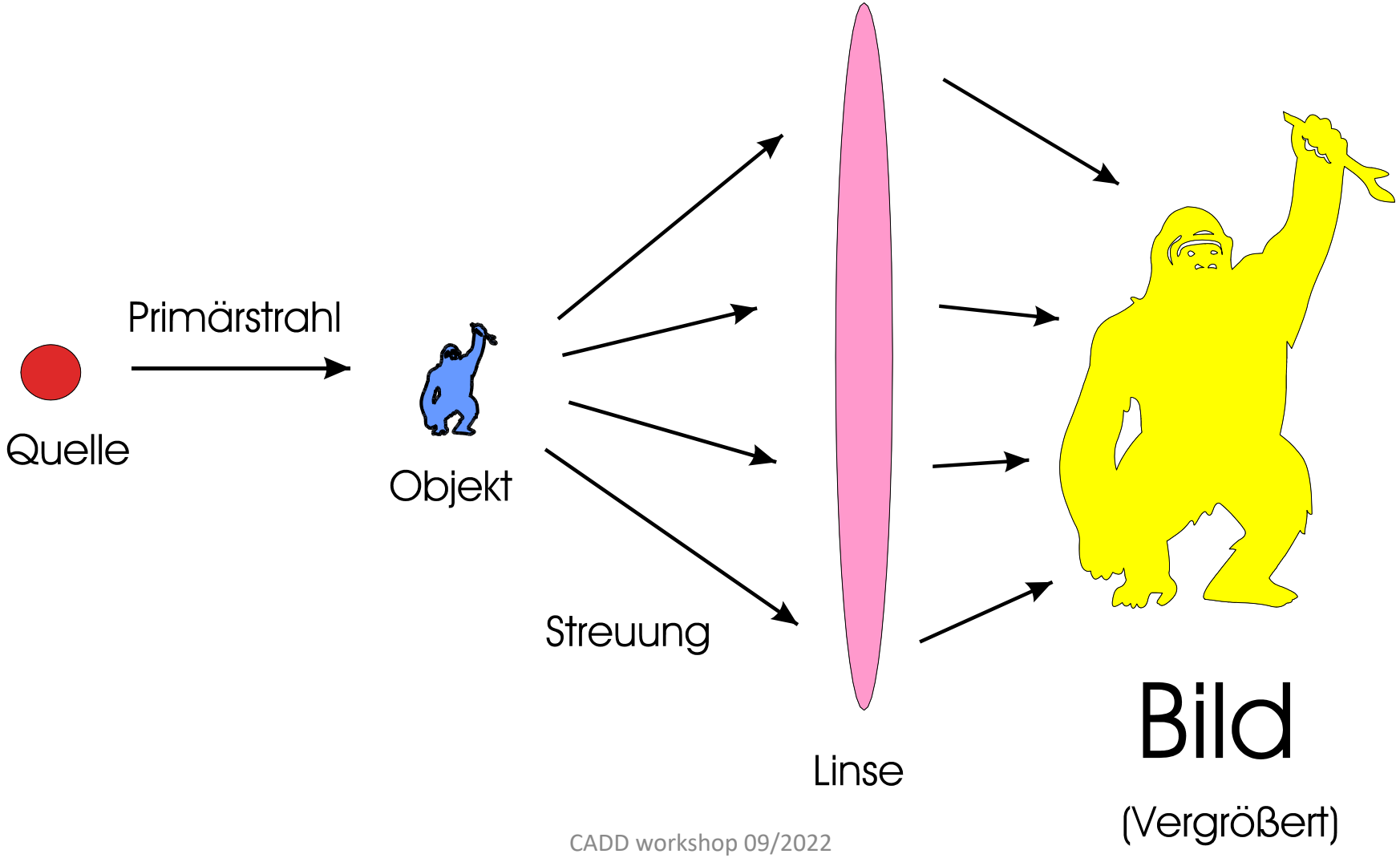
Ultimate aim

Determination of the 3-dimensional structure of a (macro)molecule with **atomic resolution**.

Know the positions of all atoms in the molecule. (At least have a reasonable idea of their average positions.)

“See” the atoms in the molecule as separate entities.

Principle of optical imaging (microscope)



Limits of resolution

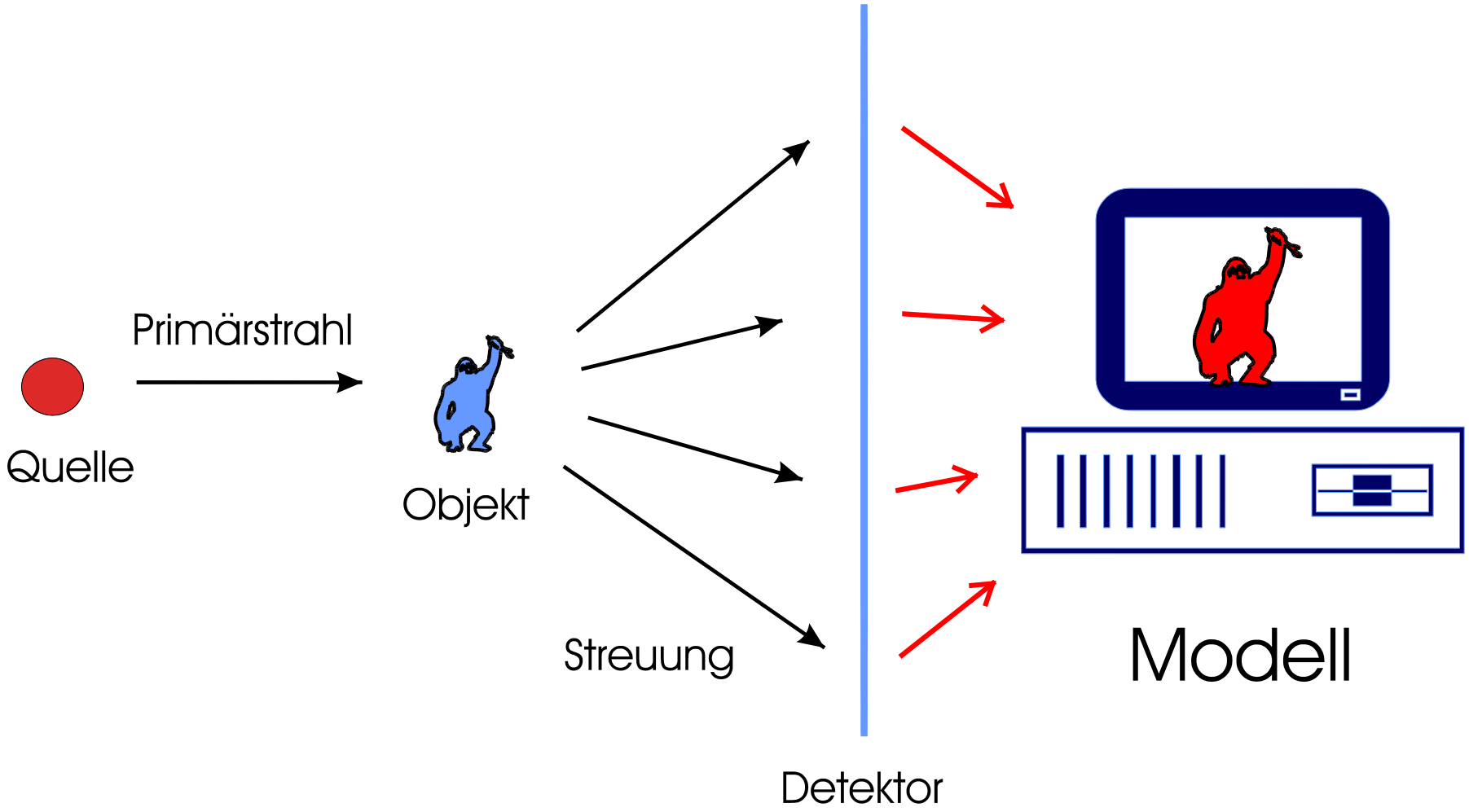
$$d_{\min} = \frac{\lambda}{2 \sin \vartheta_{\max}}$$

d_{min} … smallest distance between two points that can be imaged separately

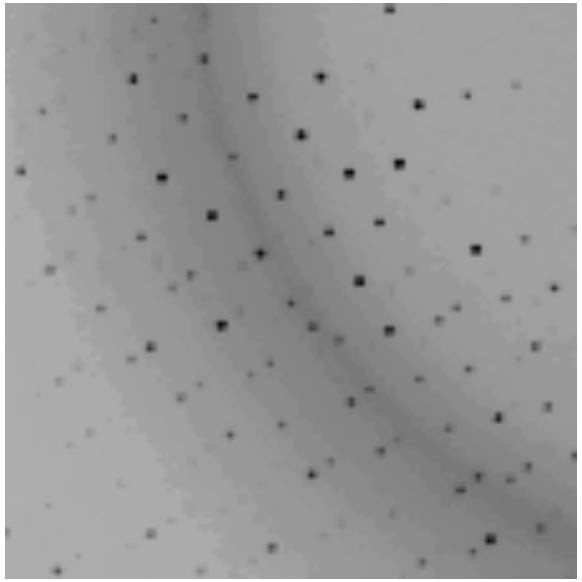
λ … wavelength

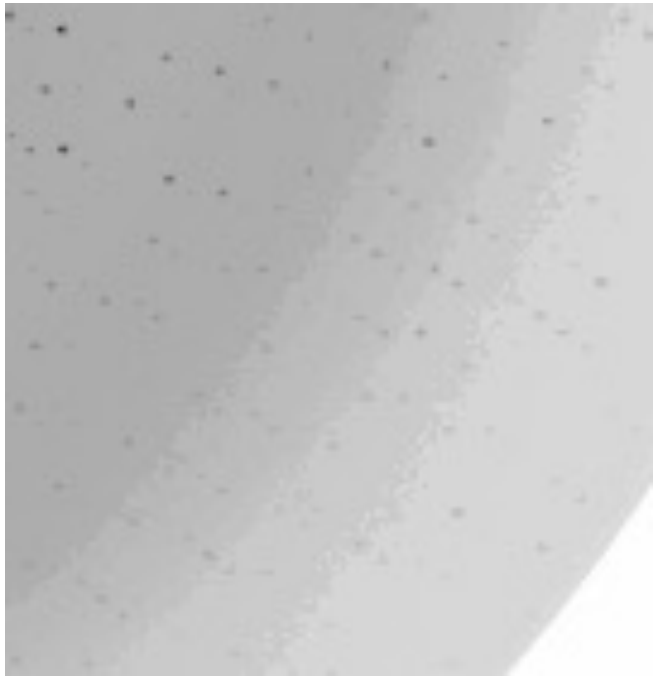
ϑ_{max} … maximum diffraction angle

Principle of a diffraction experiment

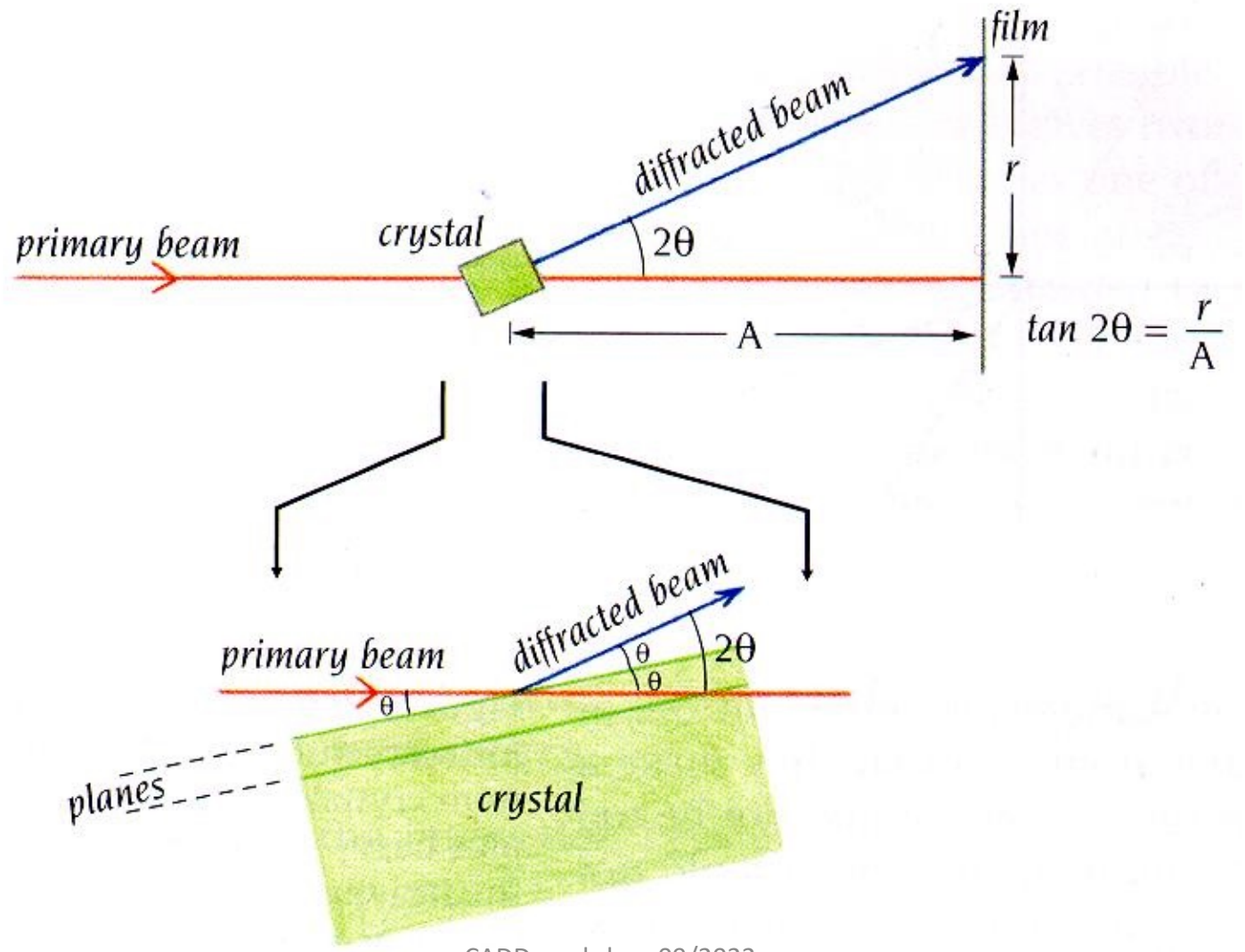


Diffraction image (protein crystal)

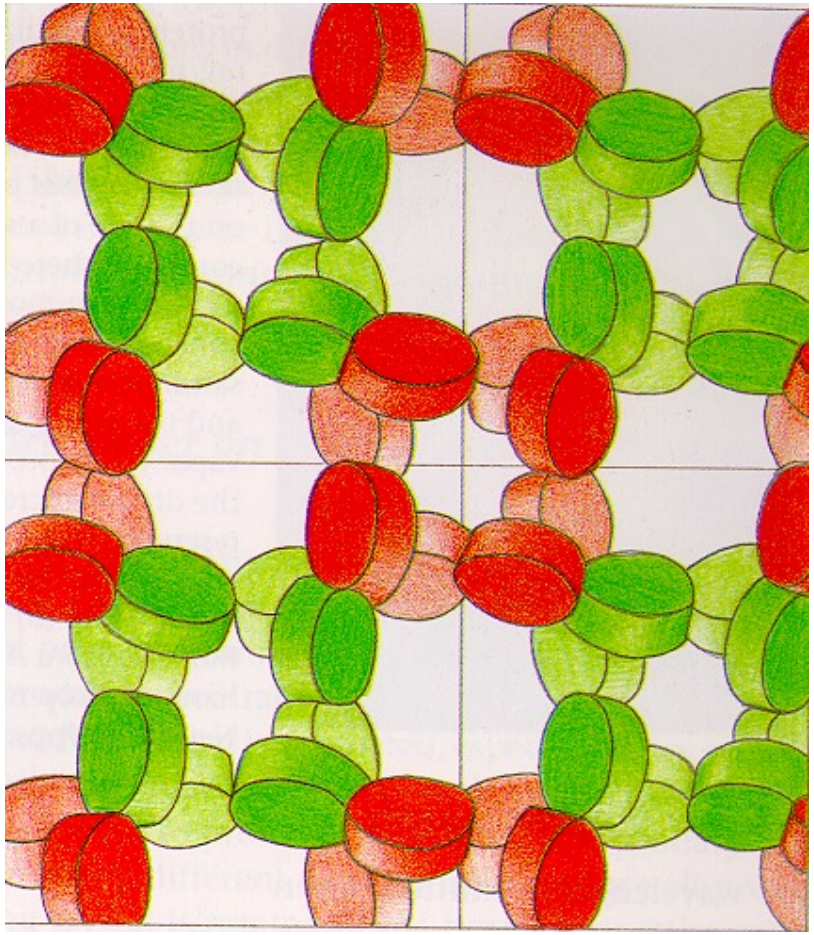




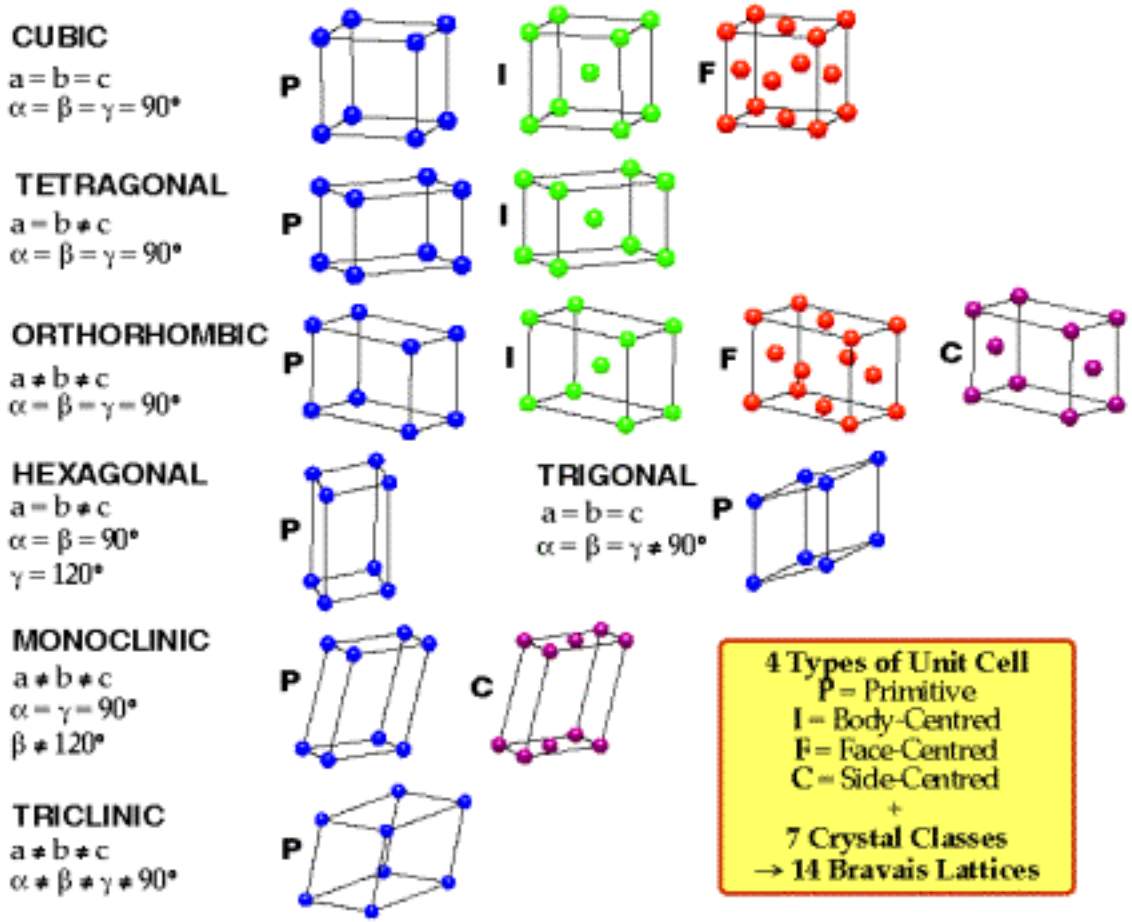
Why are the spots called “reflections”?



Crystals of biological objects



Crystal classes, crystal symmetry



Space group symmetry elements:

symmetry axes: only 2-, 3-, 4-, and 6-fold rotations are possible!!

mirror planes

center of inversion

inversion axes: rotation plus center of inversion

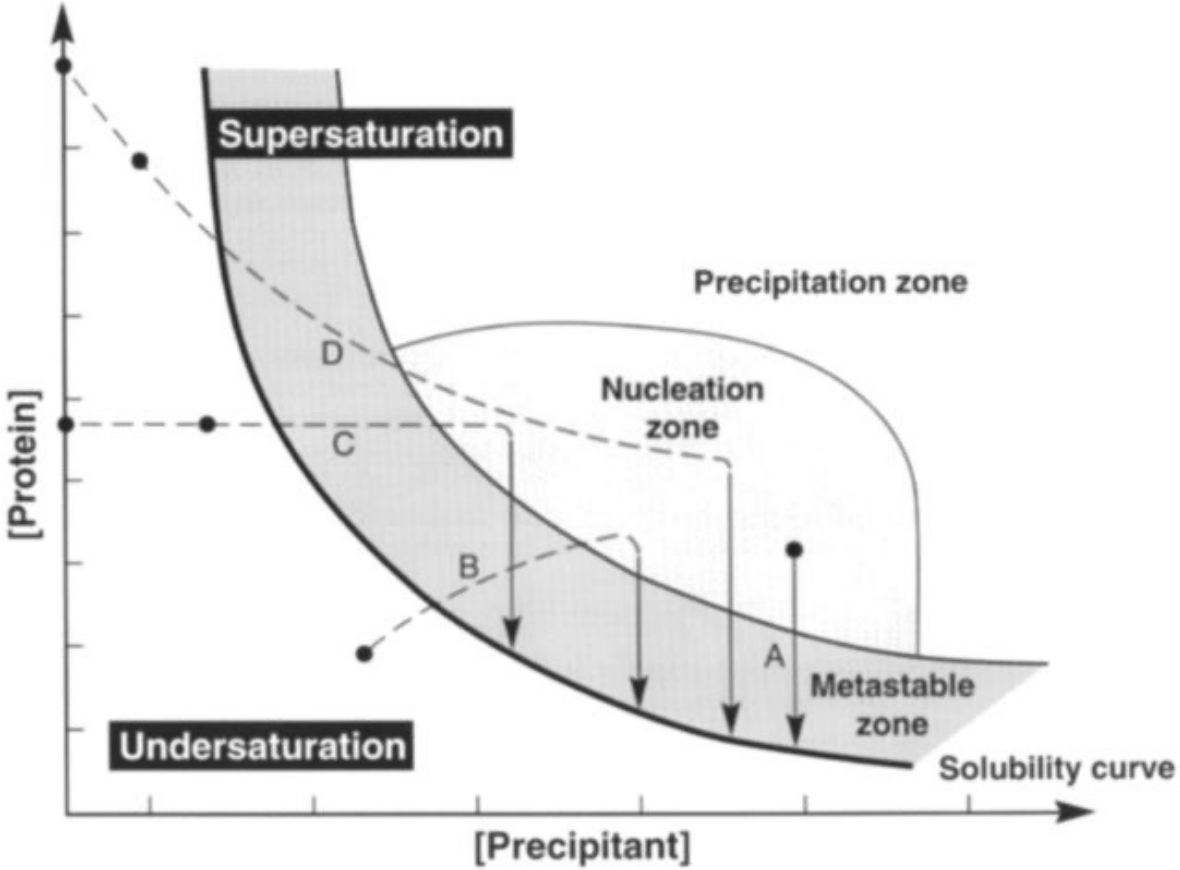
srew axes: rotation plus translation parallel to the axis

glide planes: mirror plane plus translation parallel to the plan

Workflow of an X-ray crystal structure determination

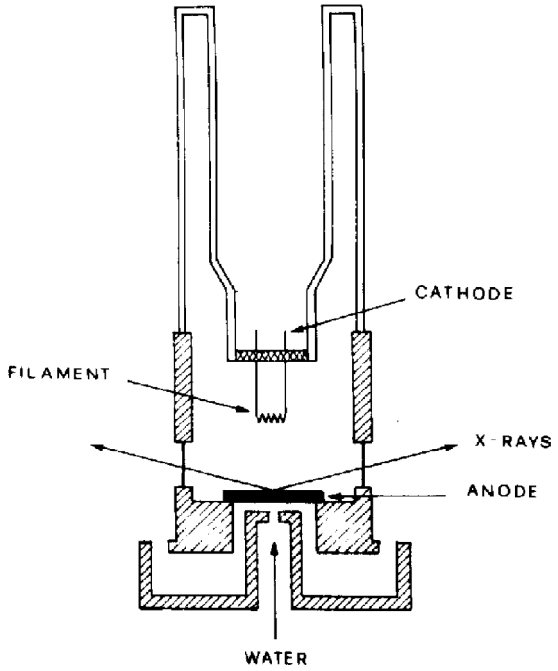
- Preparation and purification (of the protein)
- Crystallization
- Collection of a diffraction dataset
- Solution of the phase problem
- Structure refinement, validation
- Interpretation, publication of the results,…

Crystallization



- **Crystallization = ”controlled precipitation”**
- mostly trial and a lot of error/failure
- **important factors**:
  purity, homogeneity of the sample
  pH
  temperature (value, consistency)
  nature of the precipitating agent
  (salts, PEG,…)
  …

Data collection, X-ray sources



Wilhelm Conrad Röntgen, 1895

Synchrotron radiation source (ESRF)

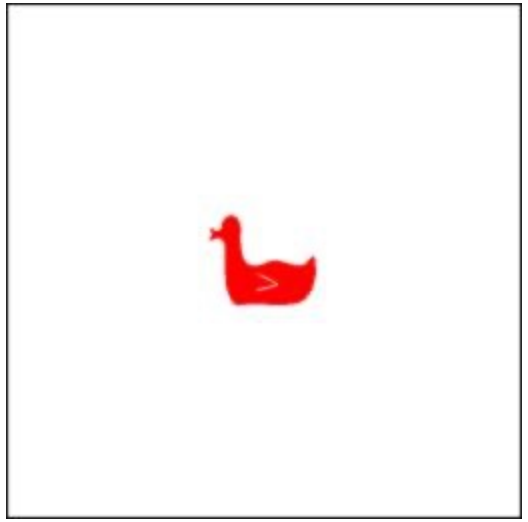
Phase problem

$$\rho(x,y,z) = \sum_{h,k,l} \left| F_{h,k,l} \right| e^{2\pi i \varphi_{h,k,l}} e^{-2\pi i (hx+ky+lz)}$$

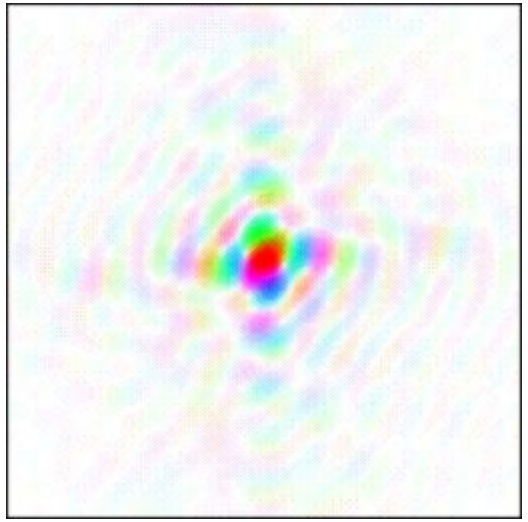
Calculation of the electron density in “real” space (x, y, z) using diffraction data (“reciprocal” space) by a Fourier transform.

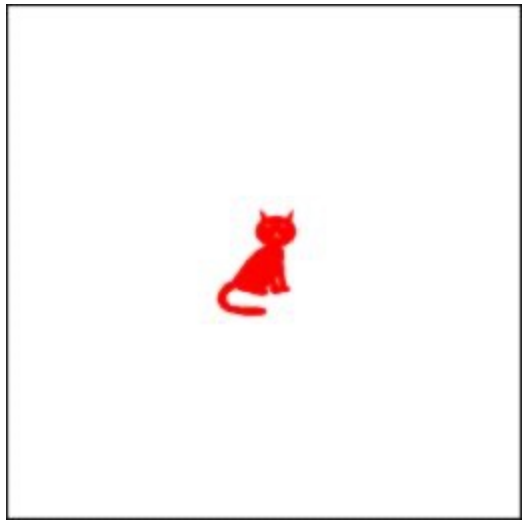
Each reflection (h, k, l) is characterized by its amplitude (F) and its phase (φ), and **both are required** for calculating the electron density.

From the diffraction experiment, we only obtain the amplitudes (square roots of the intensities) but **no phase information**. Hence, the **phase problem**.

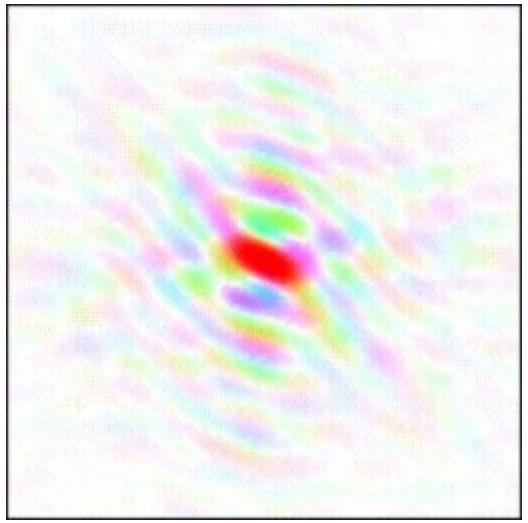


Fourier transform

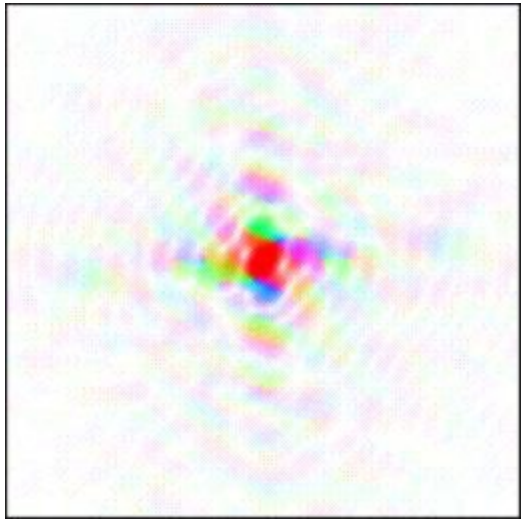




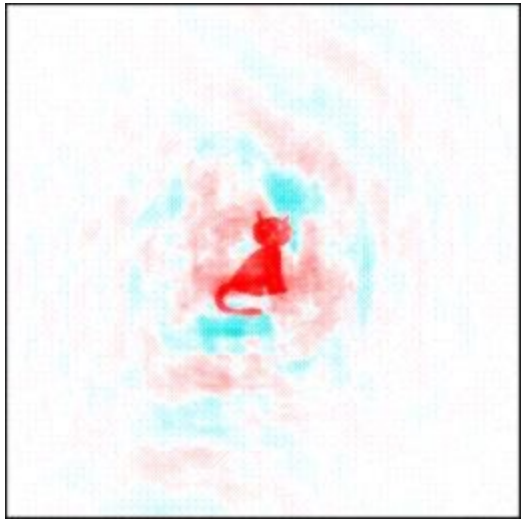
Fourier transform



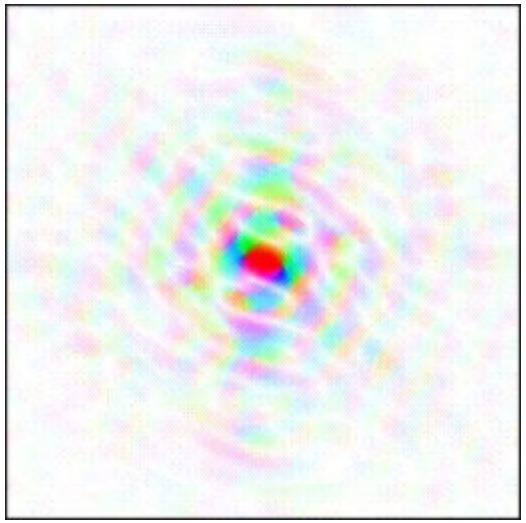
amplitudes: duck
phases: cat



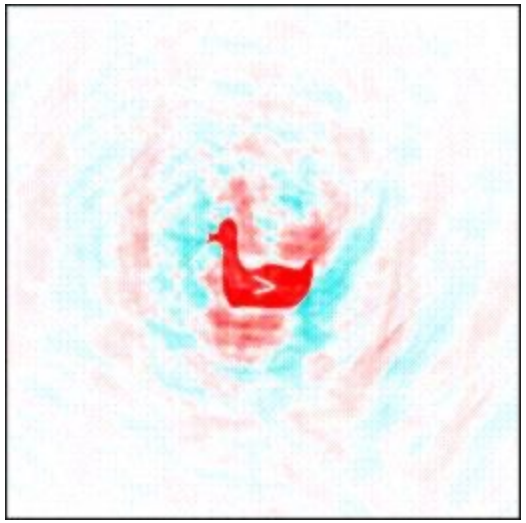
back FT



amplitudes: cat
phases: duck



back FT



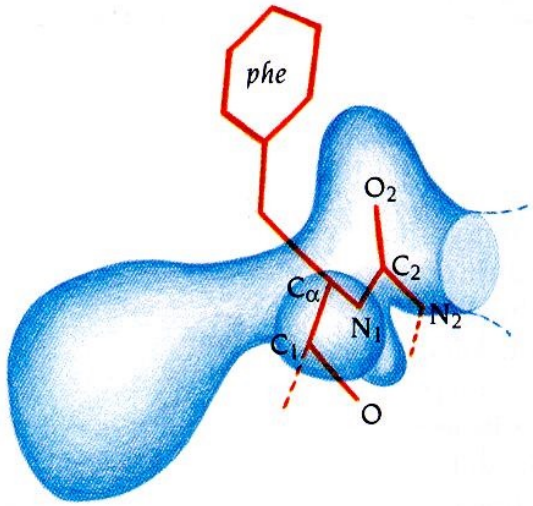
Solving the phase problem

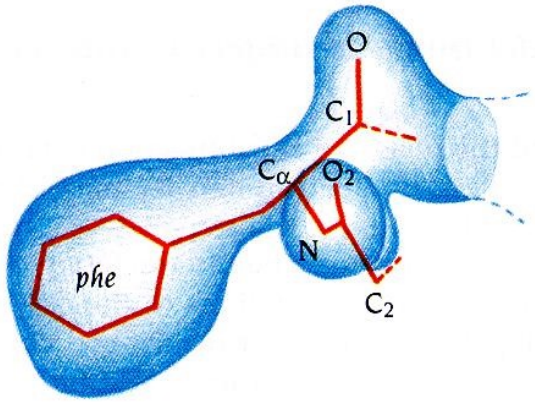
- **MR** **M**olecular **R**eplacement
- **MIR** **M**ultiple **i**somorphous **r**eplacement
- **MAD** **M**ultiple wavelengths **a**nomalous **d**ispersion

$$\Rightarrow \varphi_{h,k,l}^{calc}$$

$$\rho(x,y,z) = \sum_{h,k,l} \left| F_{h,k,l}^{obs} \right| e^{2\pi i \varphi_{h,k,l}^{calc}} e^{-2\pi i (hx+ky+lz)}$$

Model building





We do not observe "molecular structure" directly!

The direct result, after solving the phase problem, is **electron density**. The electron-density map is then interpreted by fitting into it (pieces of) a polypeptide chain.

Individual atoms are not resolved at the resolution typically obtained in protein crystallography. Instead, there are **lumps of density** corresponding to **groups of atoms**.

Structure refinement

$$F_{h,k,l}^{calc} = \left|F_{h,k,l}^{calc}\right| e^{2\pi i \varphi_{h,k,l}} =$$

$$= \sum_{i=1}^{N} f_i \, e^{2\pi i \left(hx_i + ky_i + lz_i\right)}$$

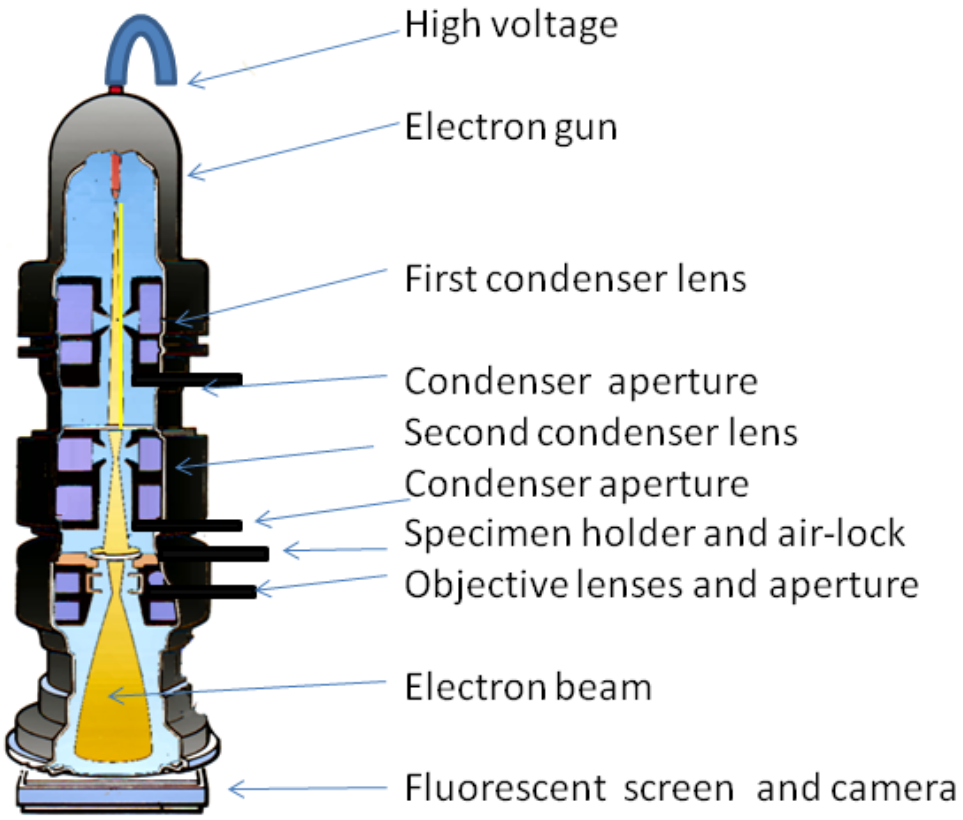
Structure factor calculation relies on the "crystallographic structure model": the unit cell contains **N distinct atoms** at positions (x, y, z) with structure factors f.

F^{calc} corresponds to the hypothetical diffraction data (**including phases**) we would obtain based on the current model.

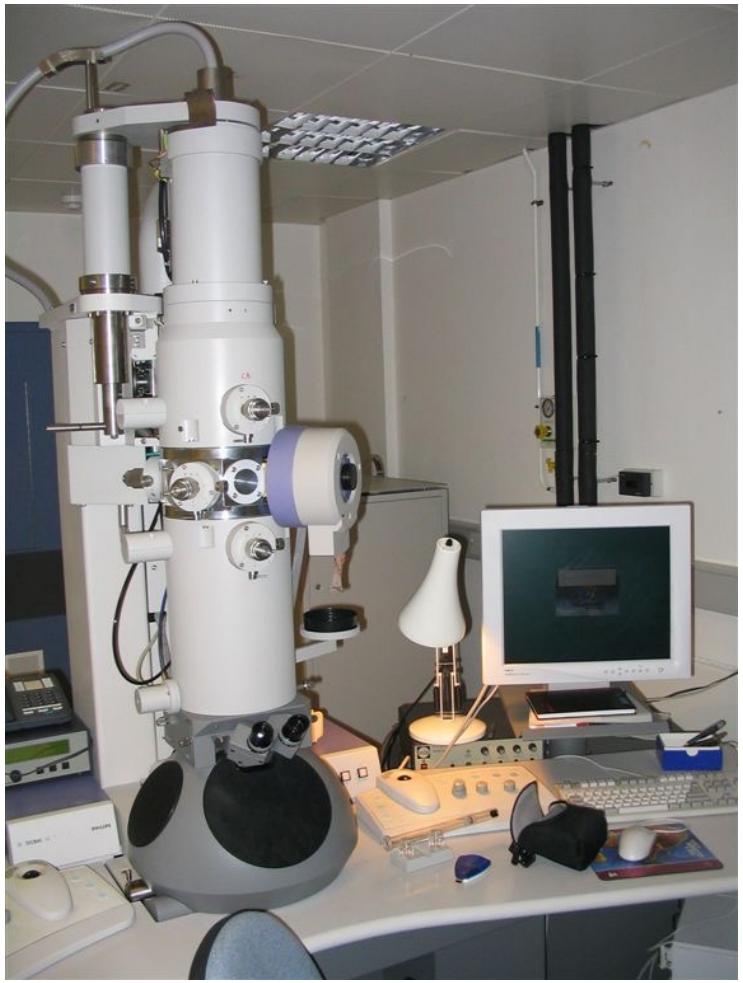
The model is continuously adjusted during structure refinement to **minimize the differences** between observed (**F^{obs}**) and calculated (**F^{calc}**) reflection amplitudes.

$$R = \frac{\sum_{h,k,l} \left| \, |F_{obs}(hkl)| - |F_{calc}(hkl)| \, \right|}{\sum_{h,k,l} |F_{obs}(hkl)|}$$

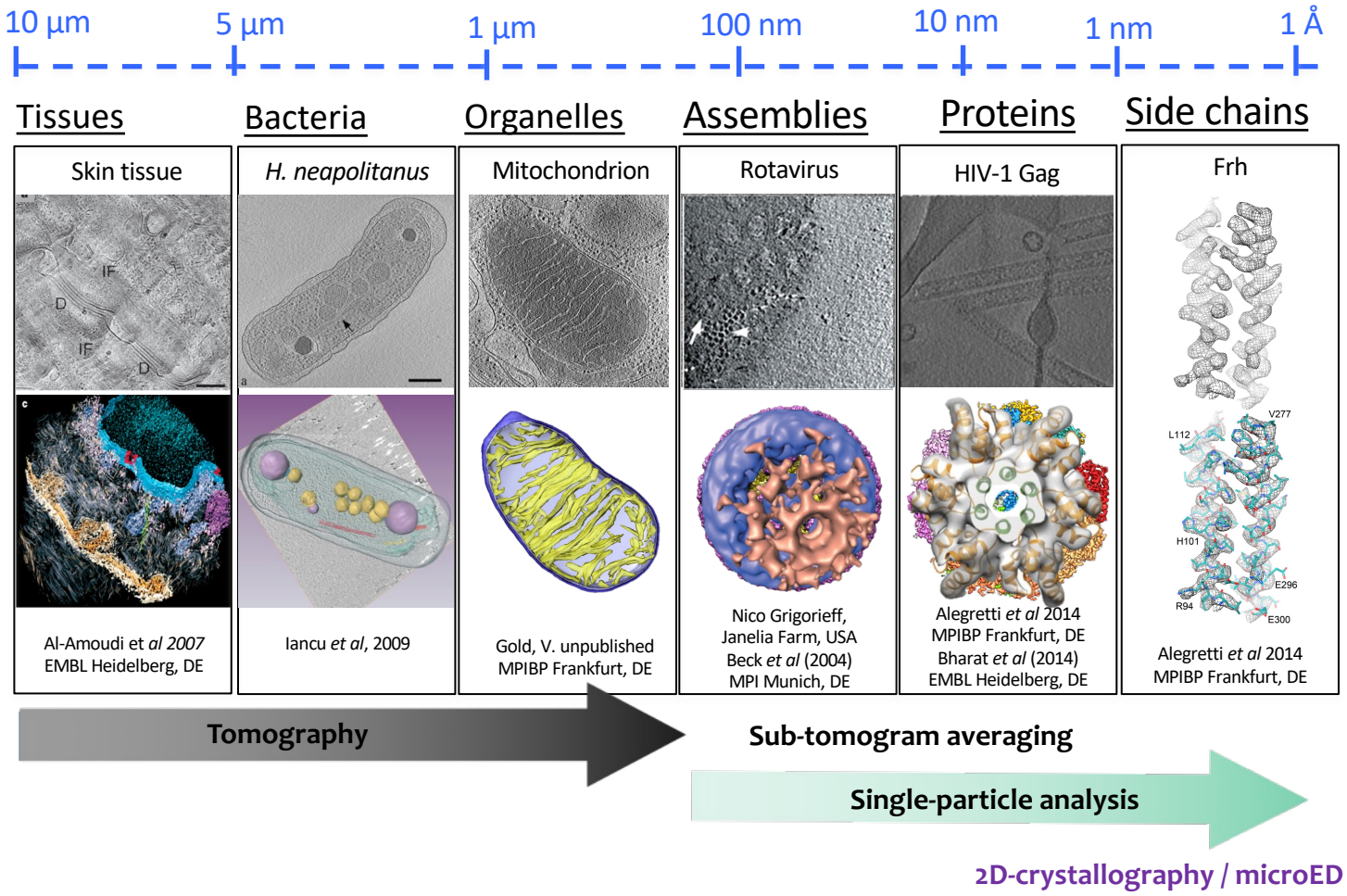
Electron microscopy



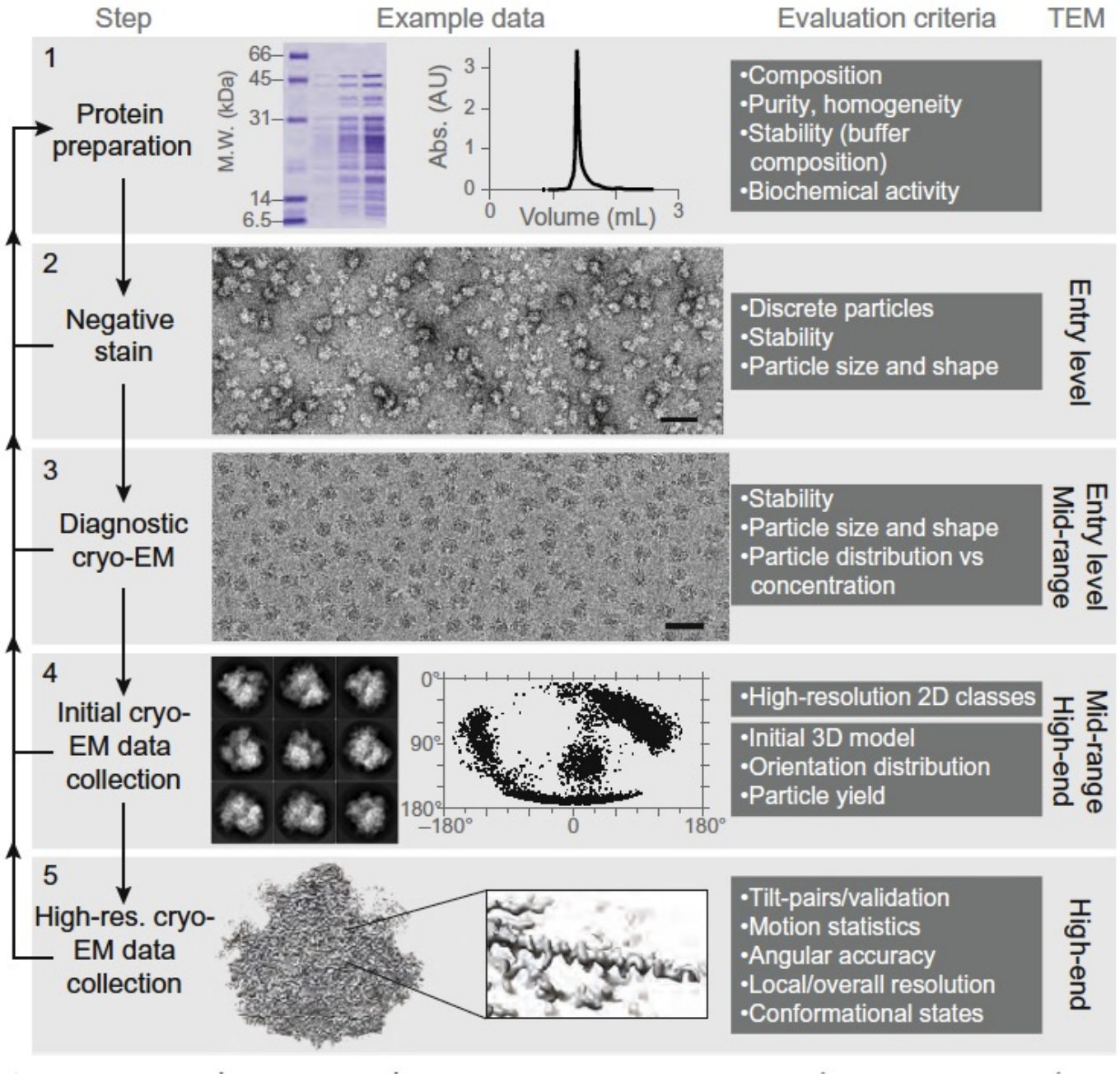
Transmission Electron Microscope



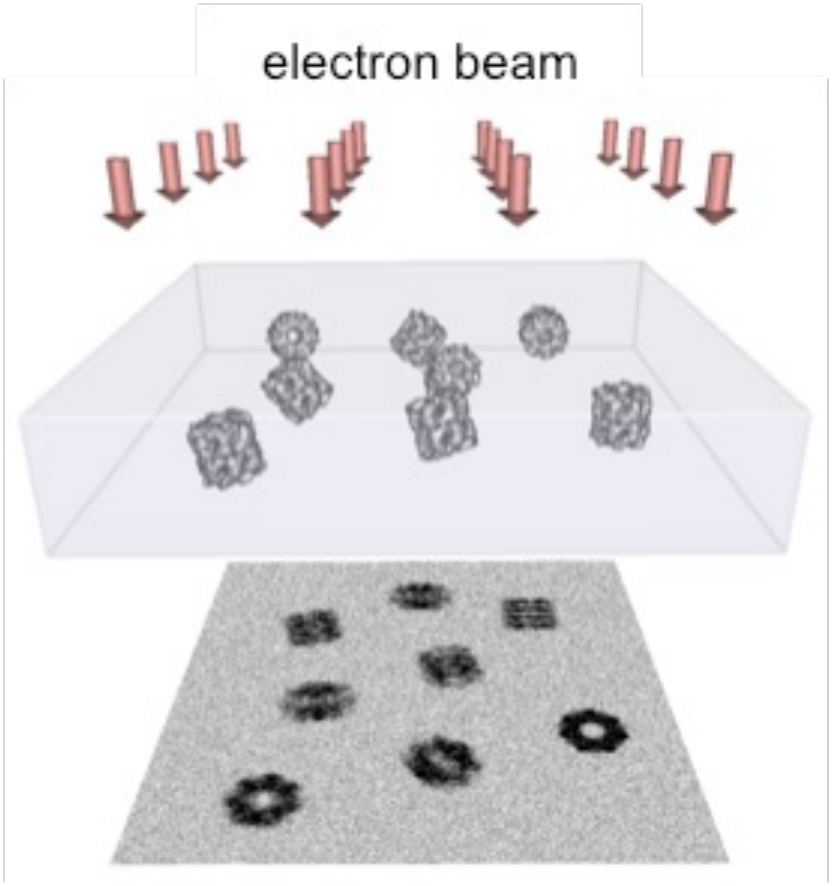
Applications of cryoEM

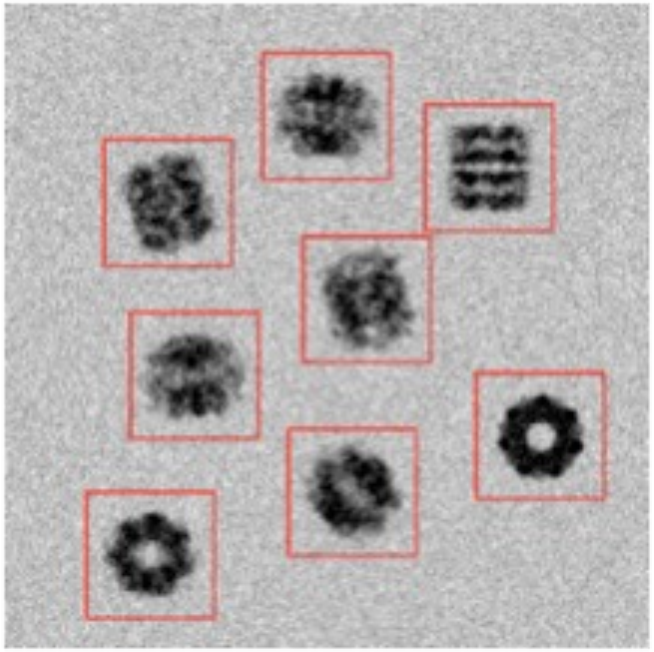


3D structure determination by cryo-EM



Particle picking

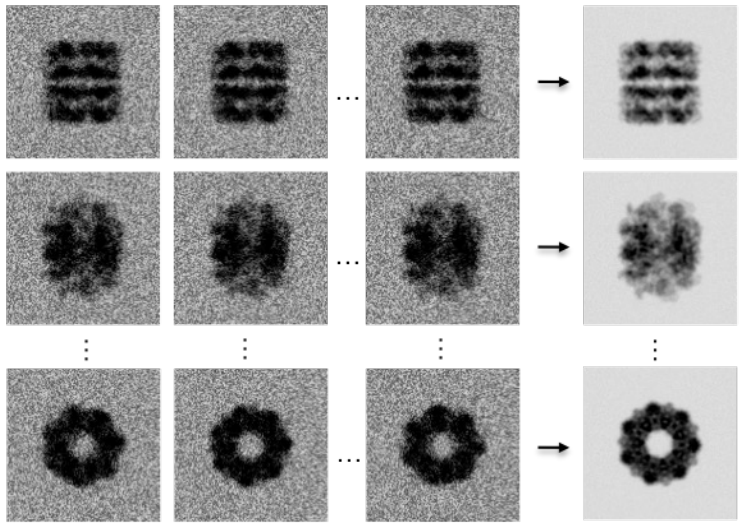




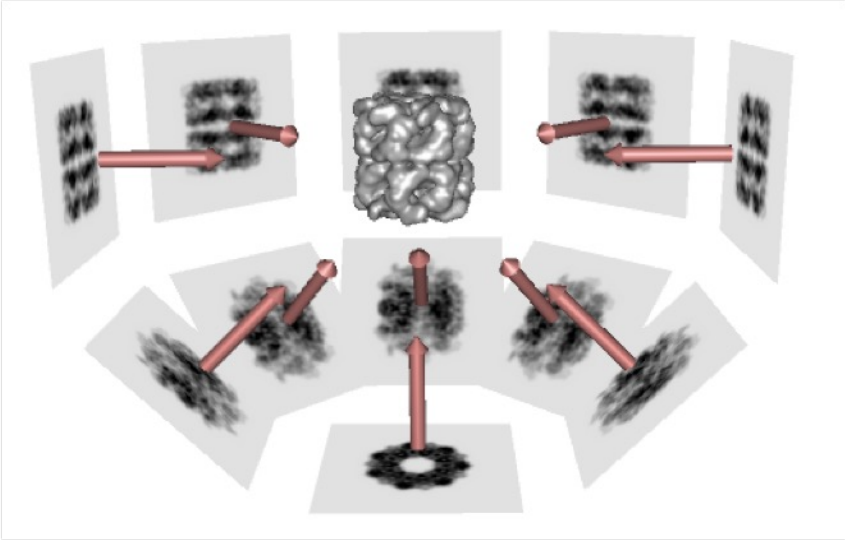
http://people.csail.mit.edu/gdp/cryoem.html

- All the boxed images are then **clustered** - The images in each cluster are **averaged**, to **<u>improve the signal-to-noise ratio</u>**. The image below shows three clusters, with images from each cluster and the averaged image on the same row.

- The 3D orientation of each of these average images is then found. Using these orientations, the images are **back-projected** to produce a **3D density map**. This 3D density map captures the electron density throughout the macromolecular complex.



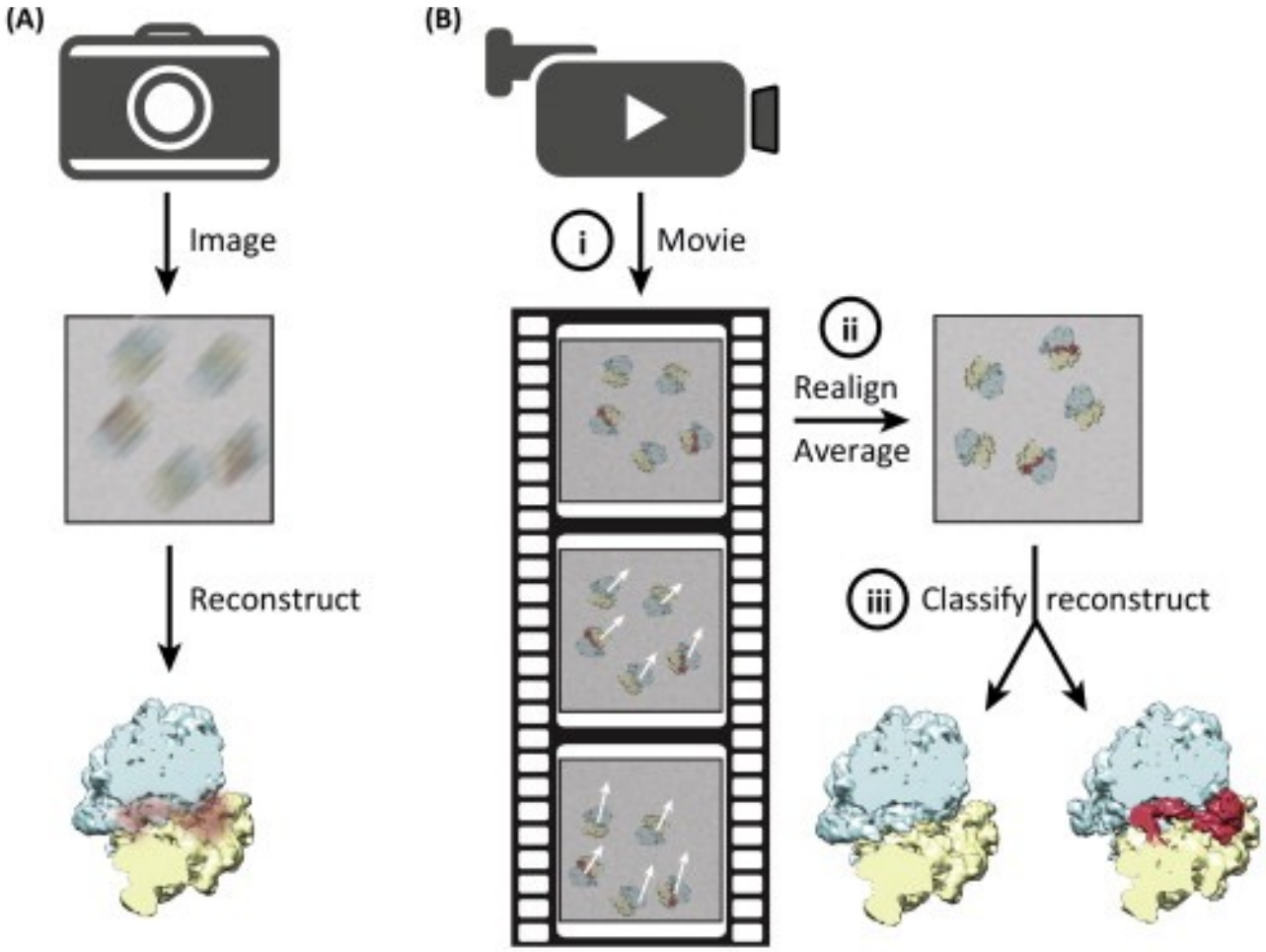
2D-class averages



Back-projection

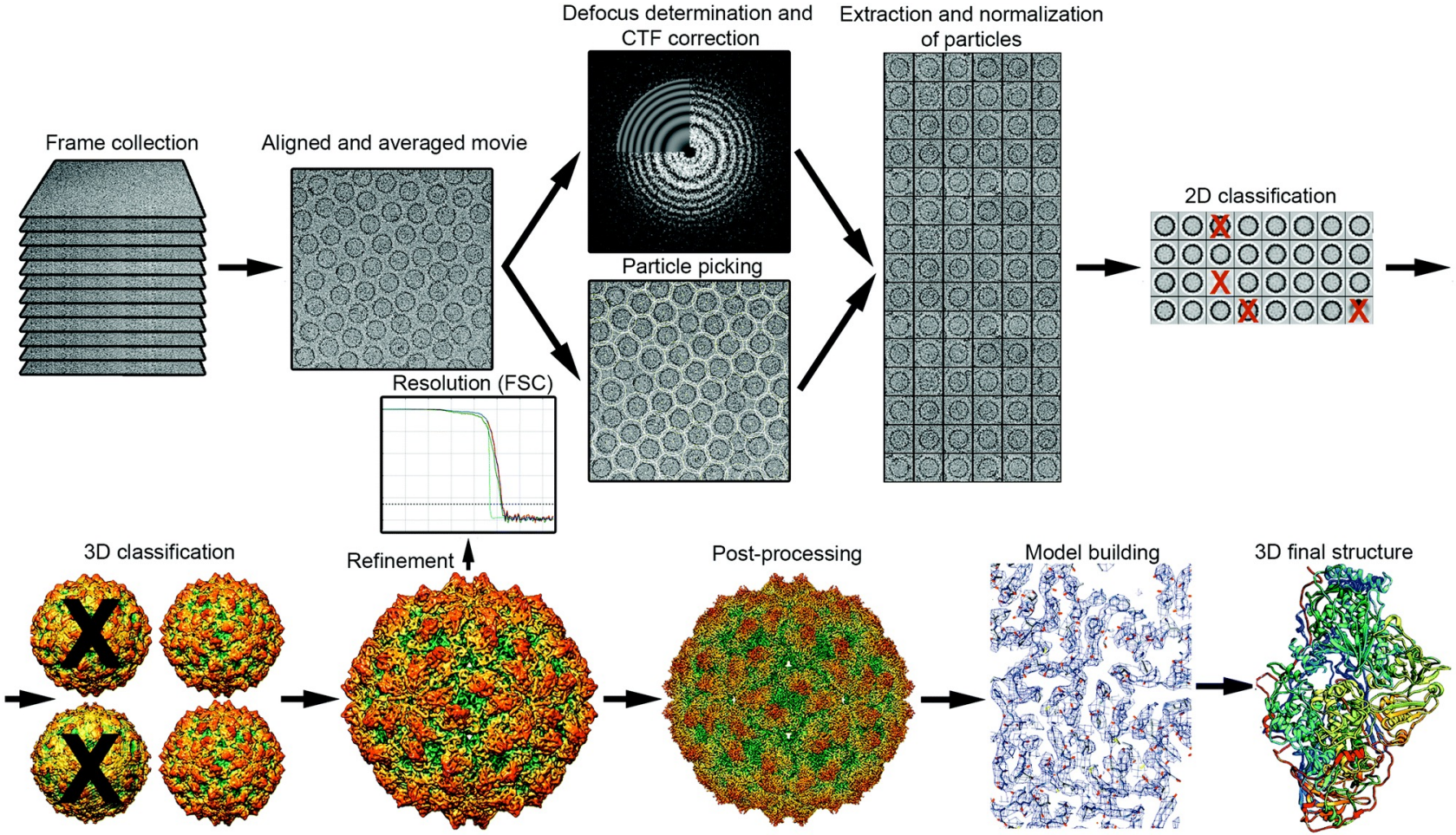
http://people.csail.mit.edu/gdp/cryoem.html

Beam Induced Movement



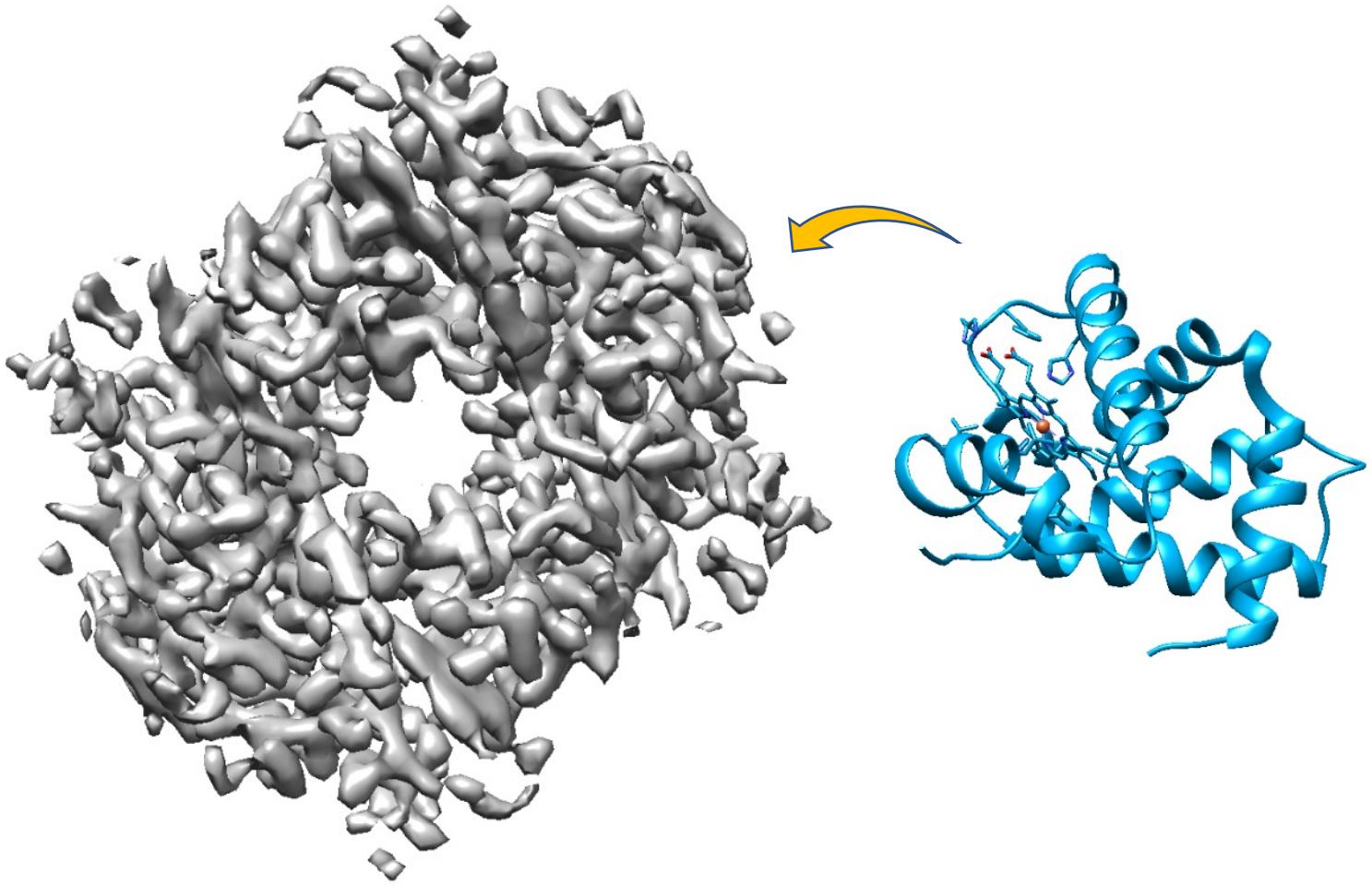
Bai et al. Trends in Biochemical Sciences, Elsevier, 2014, 40, 49-57

Single particle processing workflow

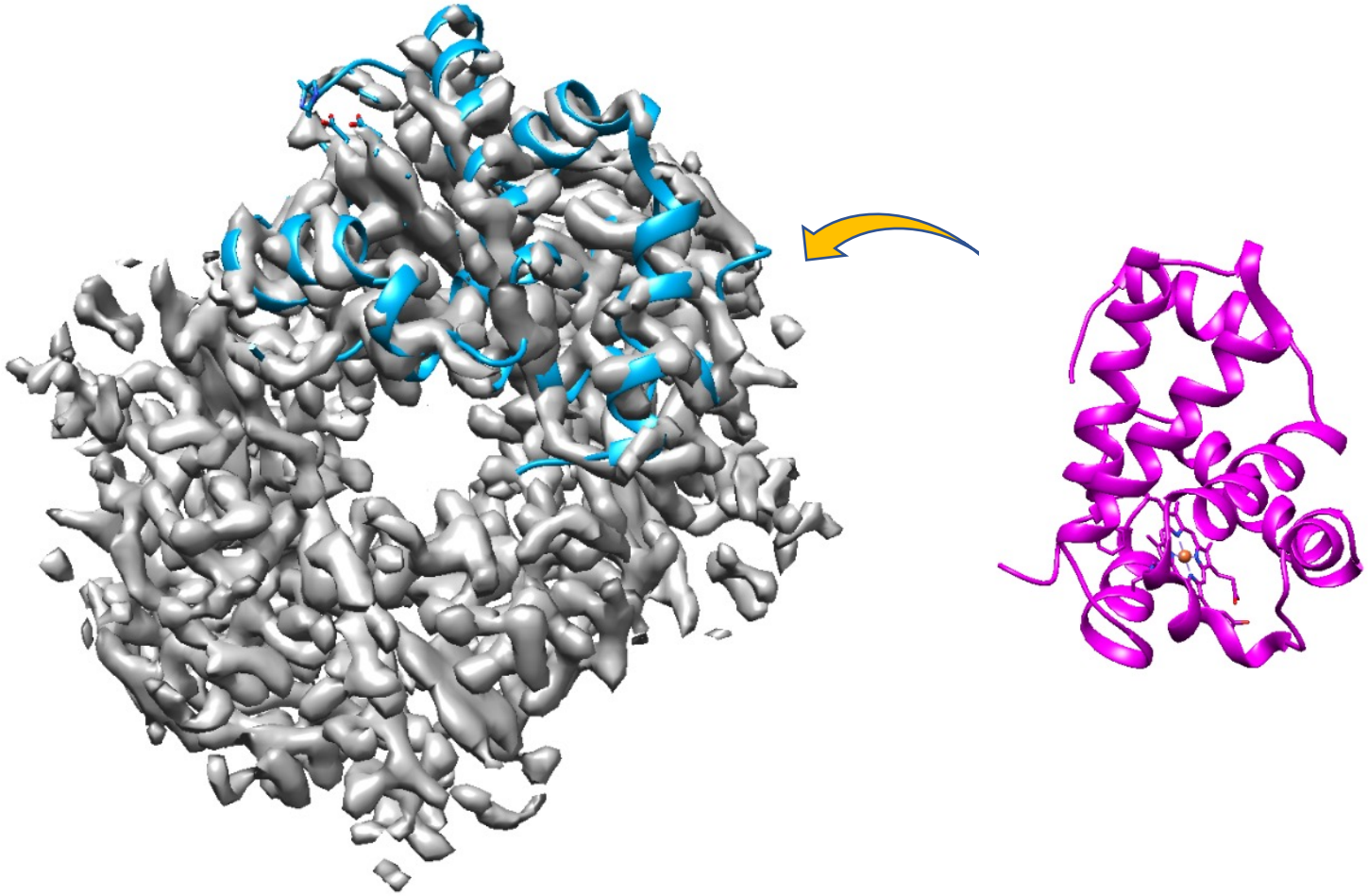


Mark V. de Ruiter, Robin Klem, Daniel Luque, Jeroen J. L. M. Cornelissen and José R. Castón, (2019) Structural nanotechnology: three-dimensional cryo-EM and its use in the development of nanoplatforms for in vitro catalysis, Nanoscale 11, 4130-4146

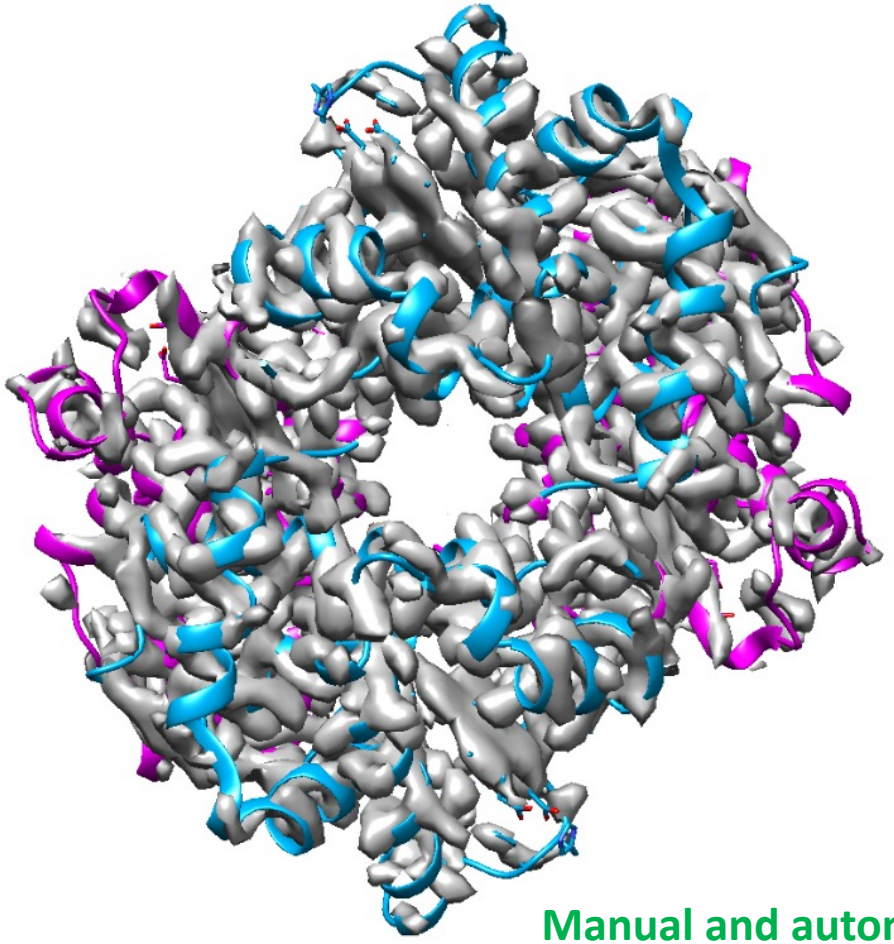
Rigid fitting of the model(s) in the map



Rigid fitting of the model(s) in the map

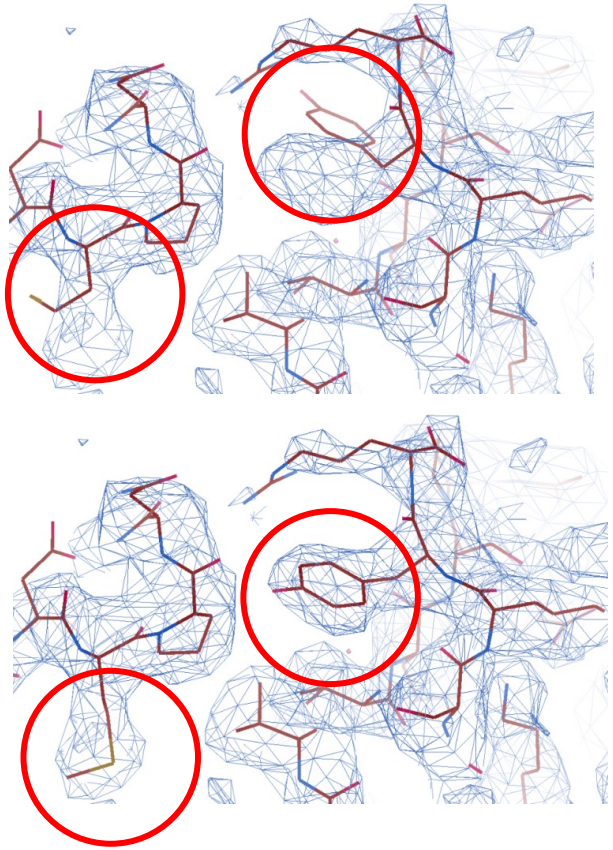


Refinement of models in the map



Manual and automatic

Get the final model(s)



Coffee Break

Structure validation

- Relevance of a (crystal) structure
- Resolution of the diffraction data (X-ray): accuracy of the structure
- R-value and free R-value (cross-validation): difference between observed and calculated scattering data

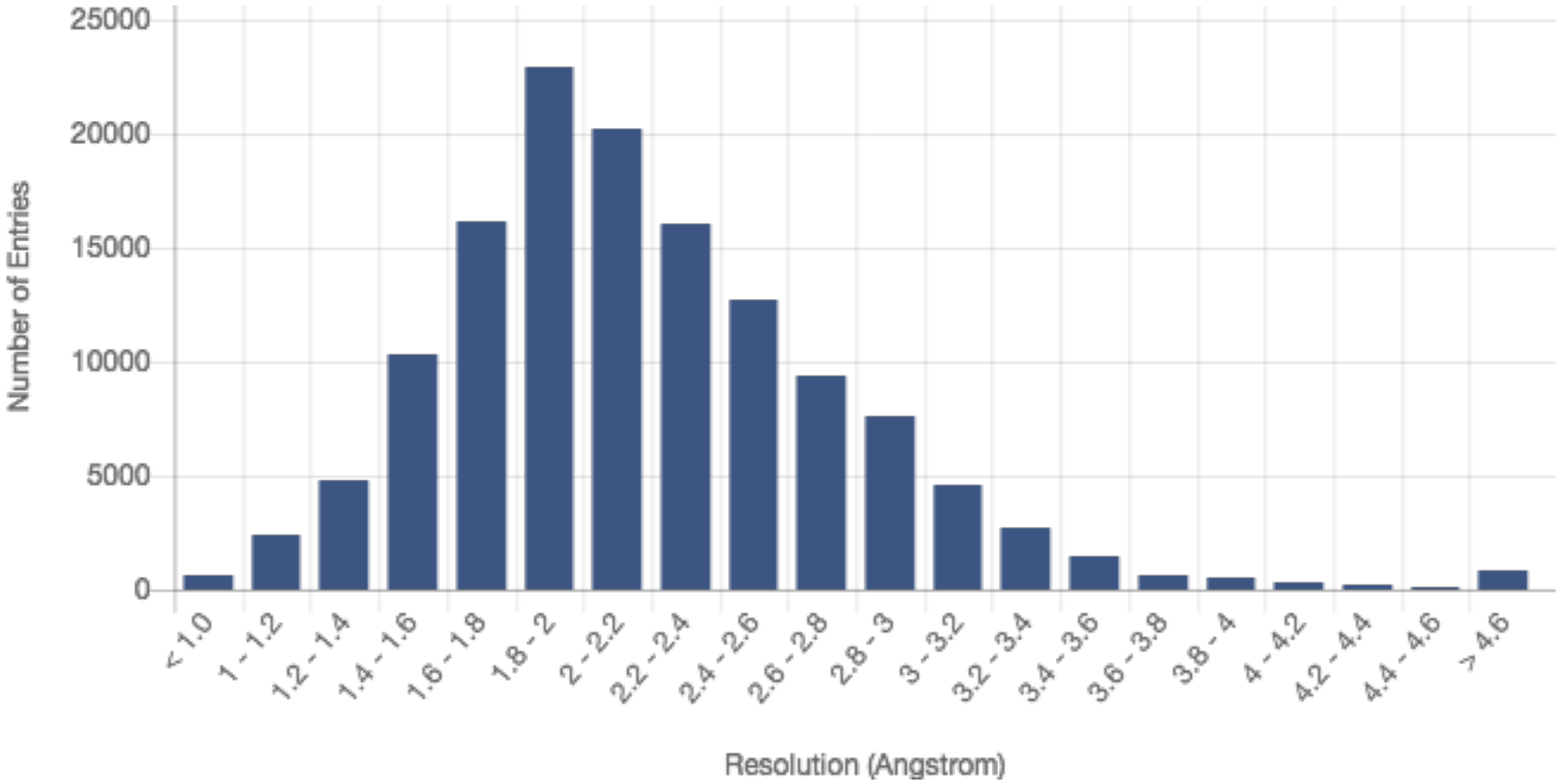
resolution

$$n\lambda = 2d \sin\theta$$

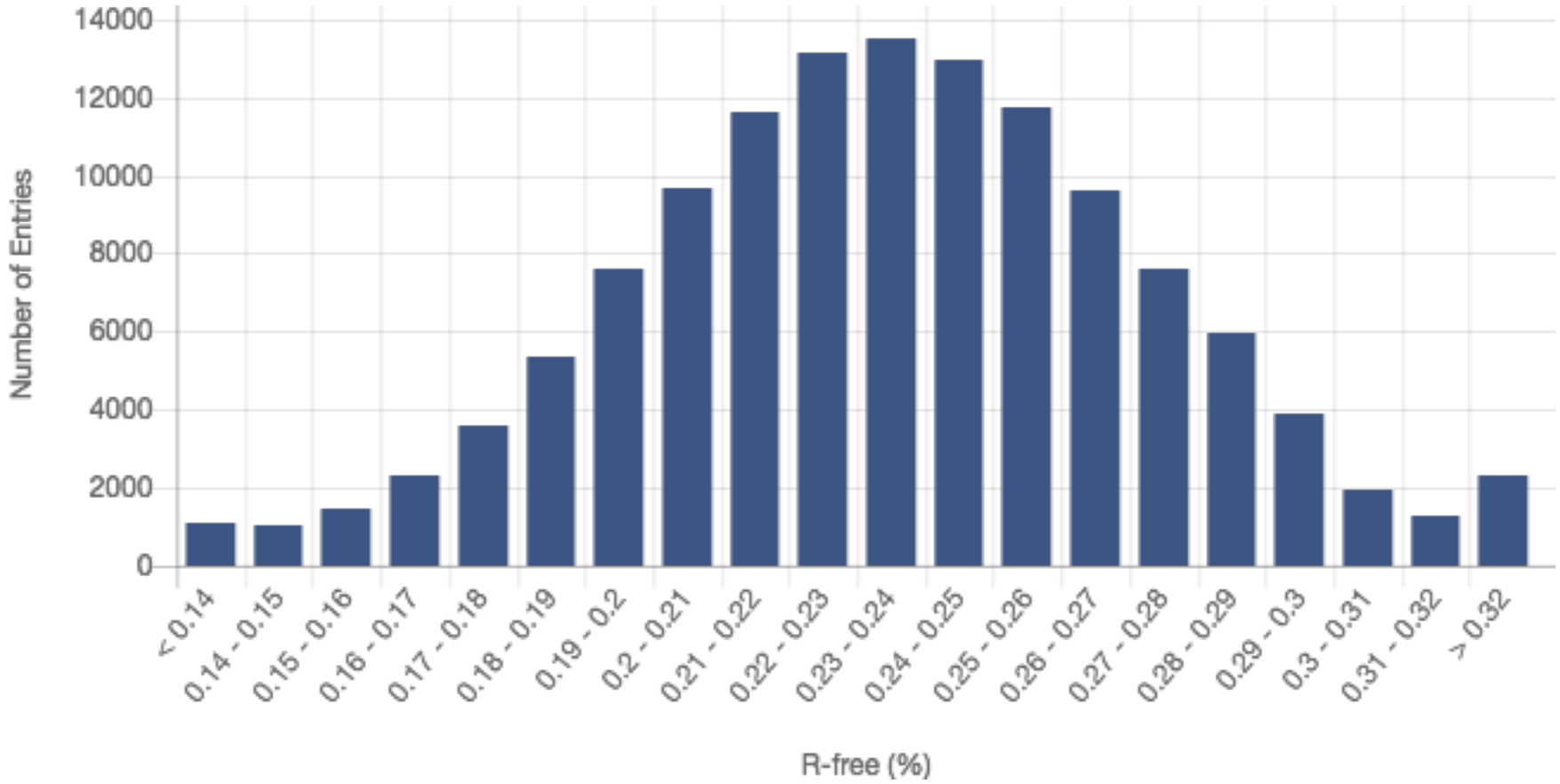
wavelength

scattering angle

PDB statistics: resolution

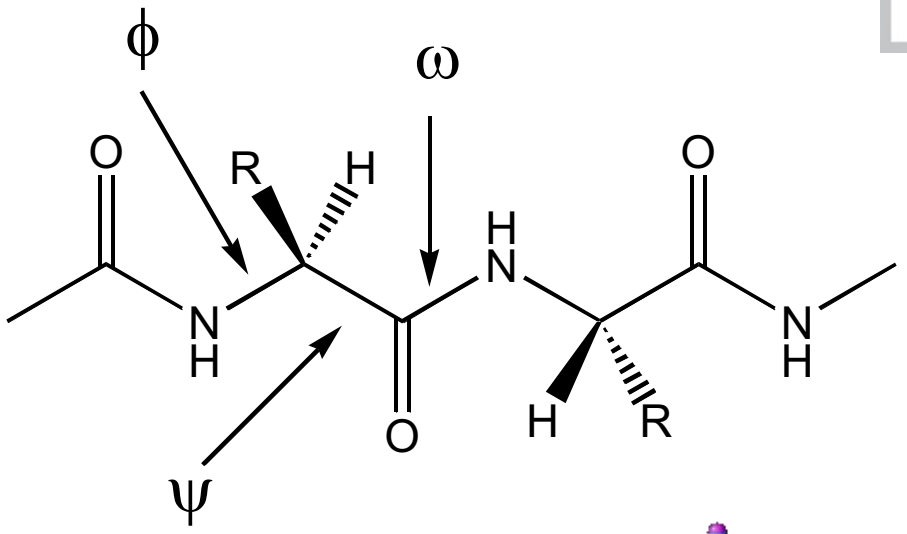


PDB statistics: R_{free} values



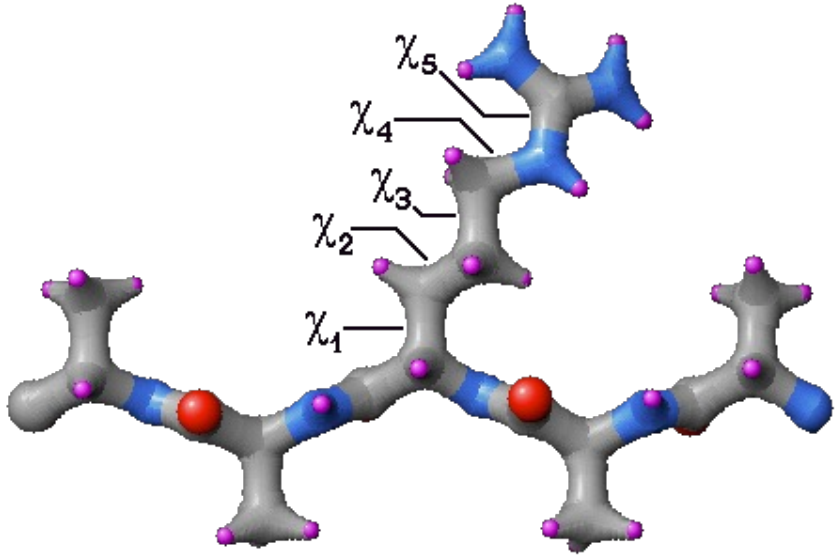
Structure validation

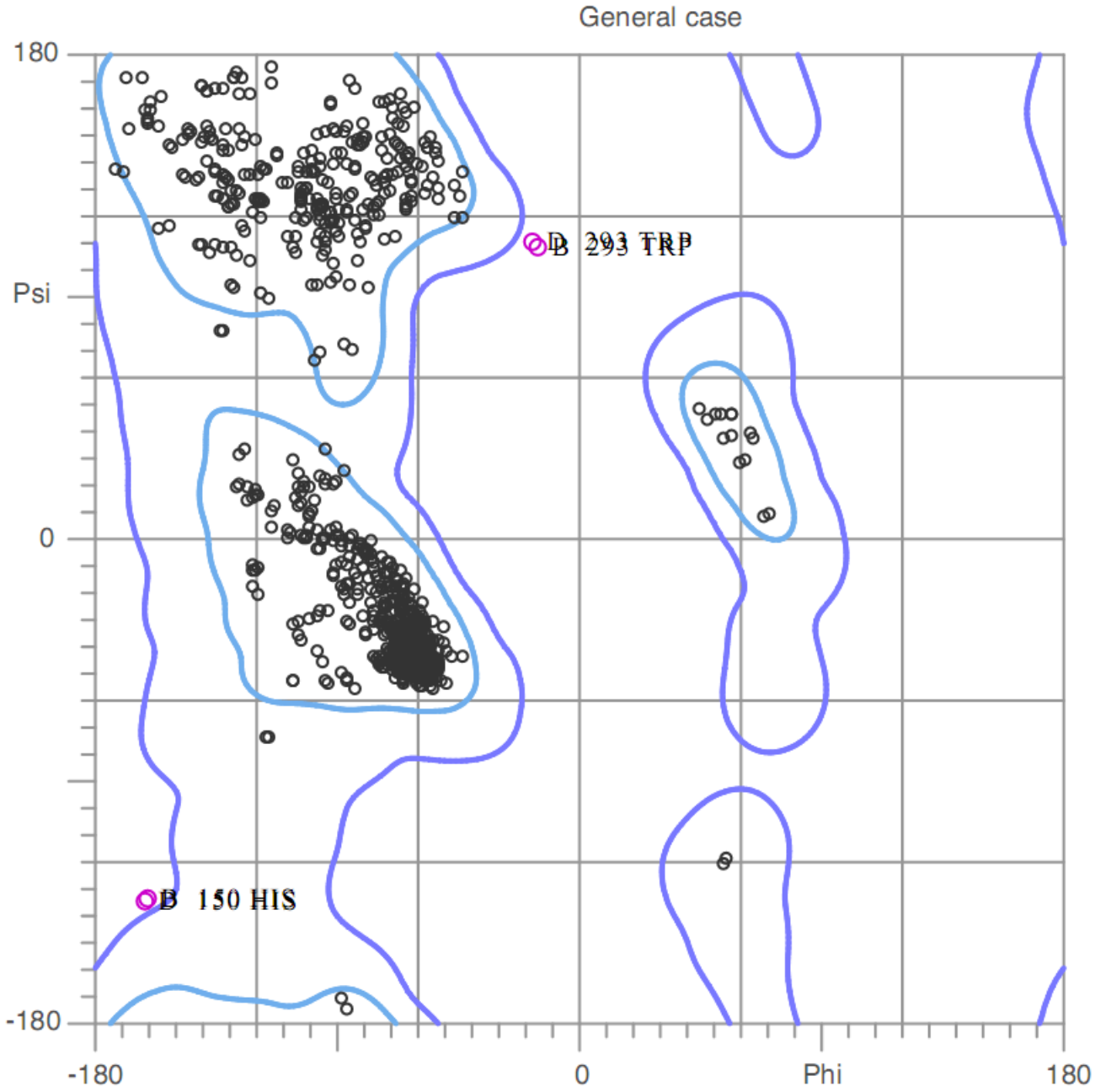
- Geometry:
  bond lengths, bond angles

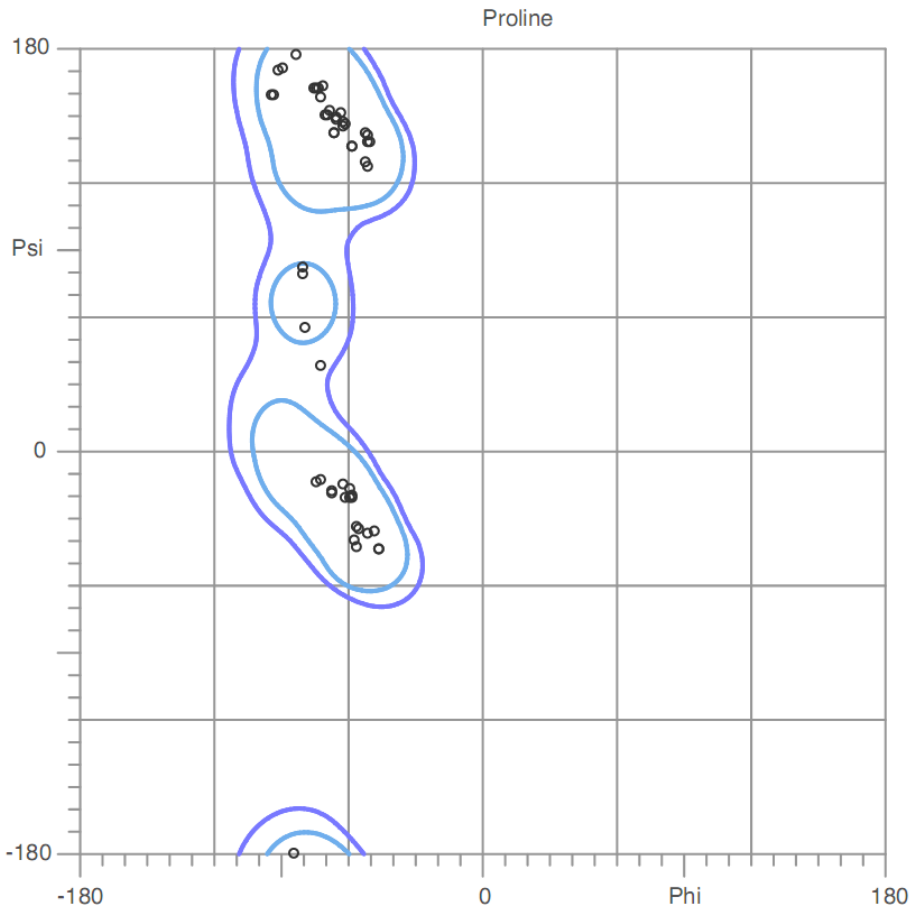


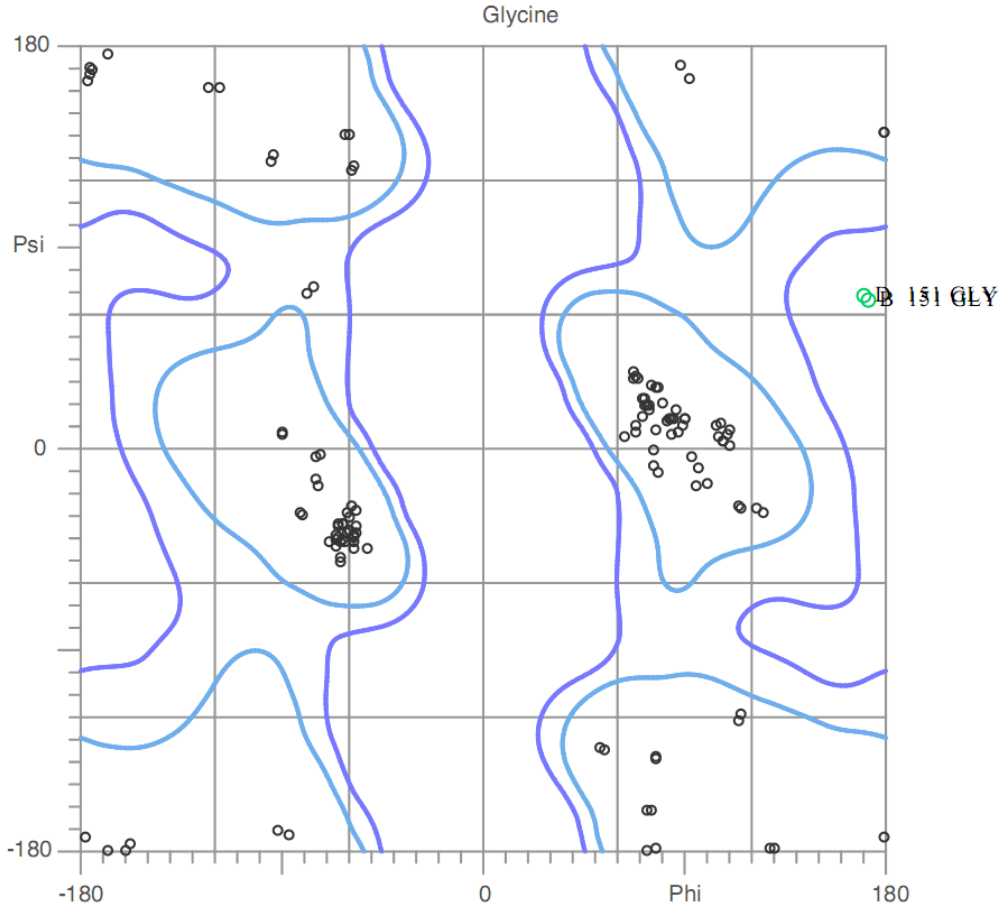
- Ramachandran-Plot:
  dihedral angles along the main chain

- Side chain torsion angles

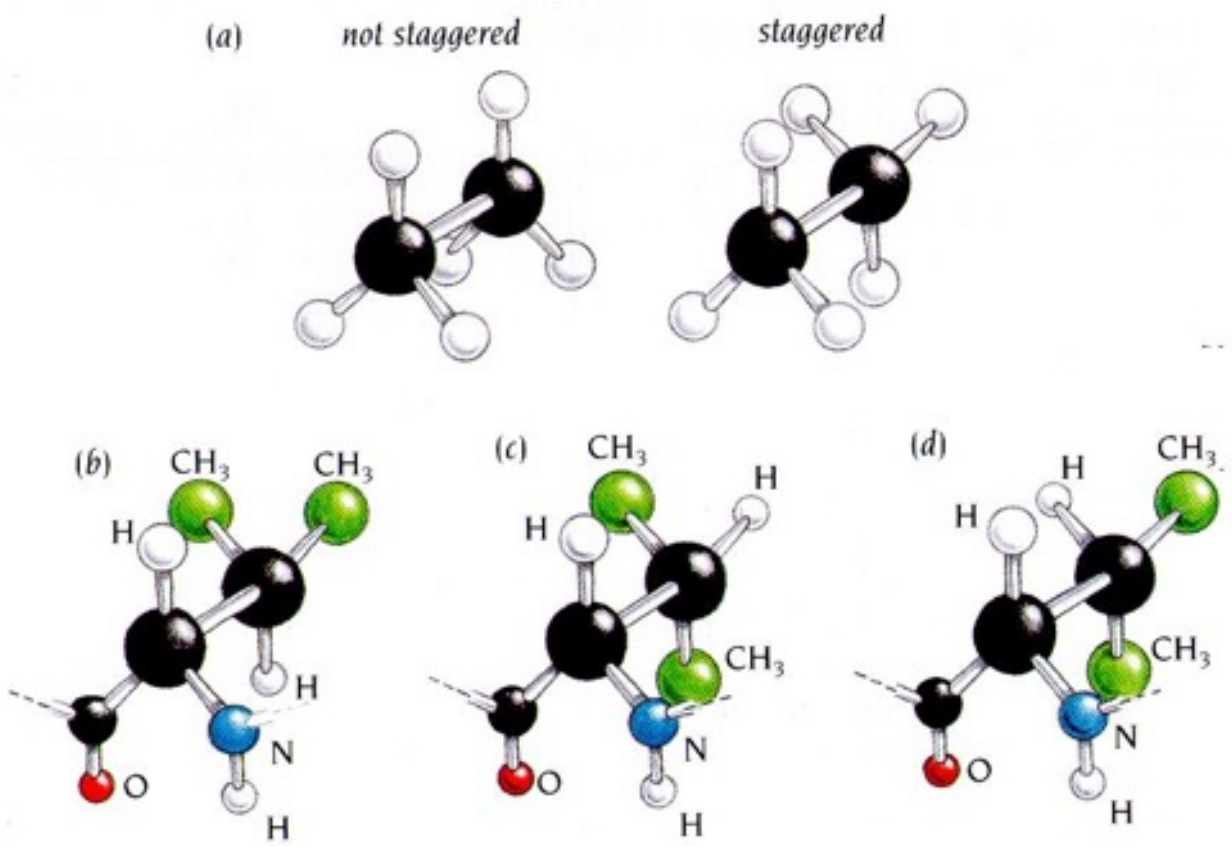


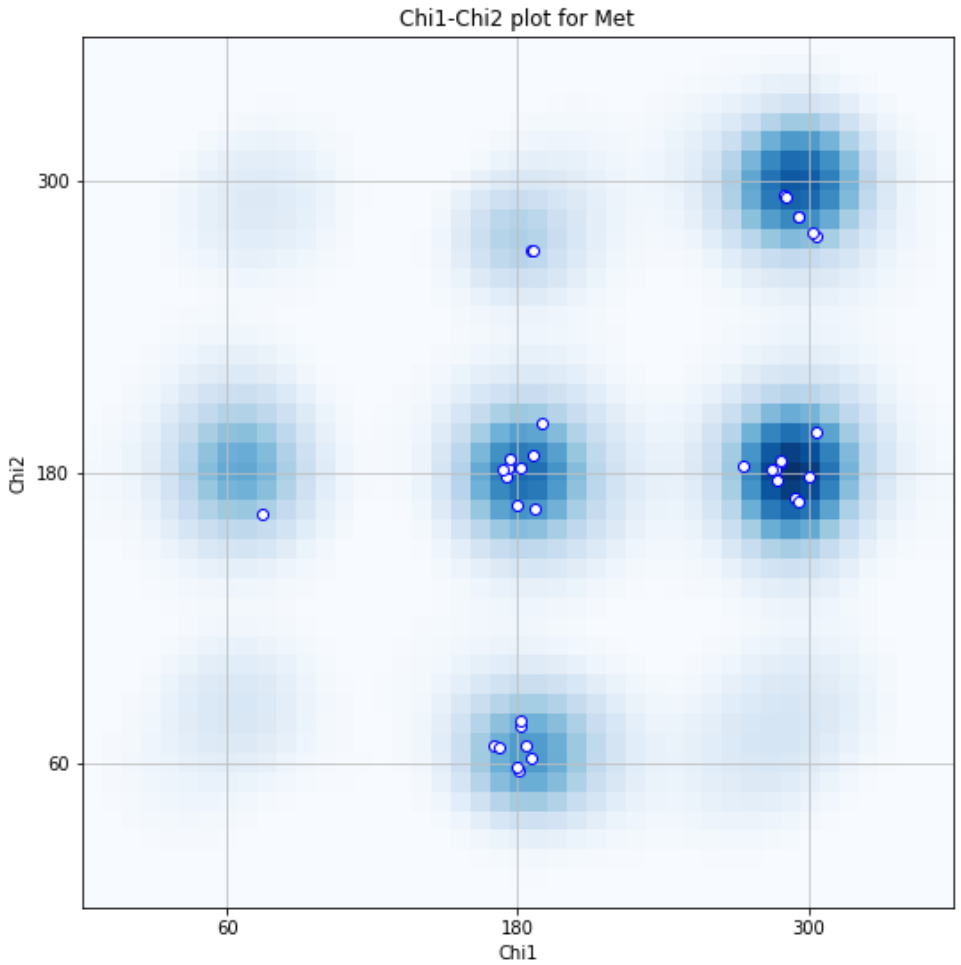

General case
180
Psi
0
-180
-180
0
Phi
180
D 293 TRP
B 150 HIS


Proline
180
Psi
0
-180
-180
0
Phi
180


Glycine
180
Psi
0
-180
-180
0
Phi
180
B 151 GLY

Sidechain conformations

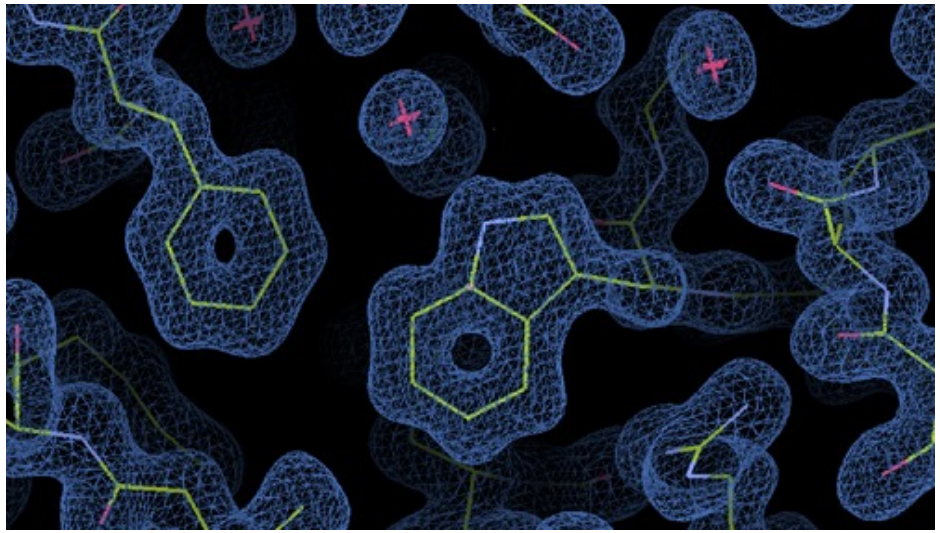




Structure validation

- Biologically active form:
  quaternary structure in the crystal
  EBI-Pisa server

- Electron density:
  primary result of an x-ray
  diffraction experiment

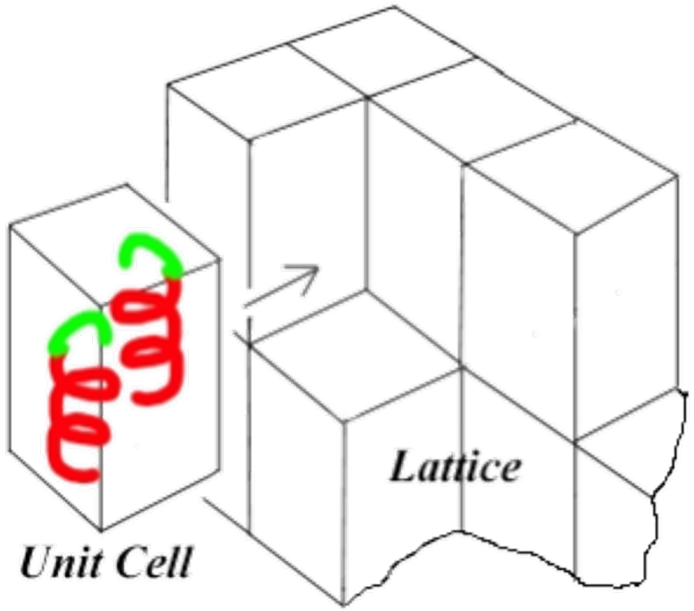


Quarternary structure in crystals

- The coordinates stored in the PDB **may or may not** correspond to the **biologically active oligomer** of a protein!
- *e.g.*: a PDB entry contains **4 chains**, but, in solution, the protein is a **monomer**.
- or *e.g.*: a PDB entry contains **1 chain**, but the biologically active oligomer is a **tetramer**.
- In the last example, the tetramer is generated by **crystal symmetry**.

Crystal symmetry

- Crystals consist of regularly arranged copies of the "unit cell" (translational symmetry).



Crystal symmetry

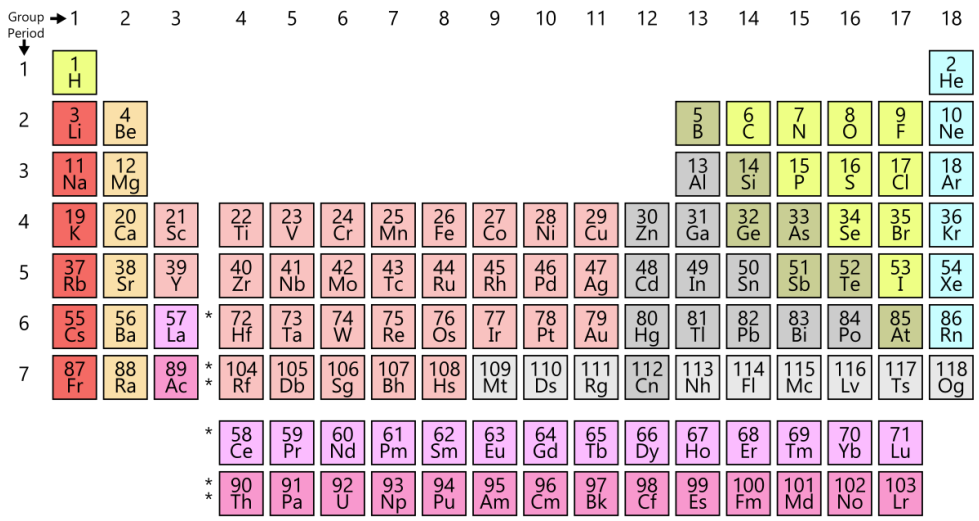
- **Asymmetric unit:** the portion of the unit cell that has to be considered in a crystallographic experiment
- The content of the entire unit cell results from the application of **symmetry operations** (rotation axes and screw axes).
- The whole crystal is ”generated" by copying and translating the unit cell in all three spatial directions.
- **For x-ray crystal structures, only the content of the asymmetric unit is stored in the PDB.**

EBI-Pisa server

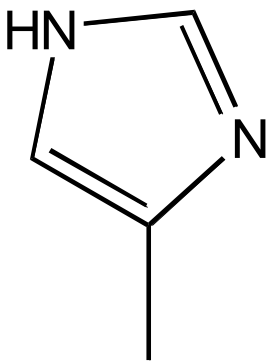
- http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html
- It analyses all **interactions** between molecules in the crystal (within the asymmetric unit and between symmetry-equivalent molecules).
- The interacting molecular surfaces are categorized according to their **size** and **chemical properties**.
- From this data, the program determines those interactions that are likely to be present also in solution.
- **Prediction of the biologically active oligomer (in solution)**

Structure validation – preparation for modeling calculations

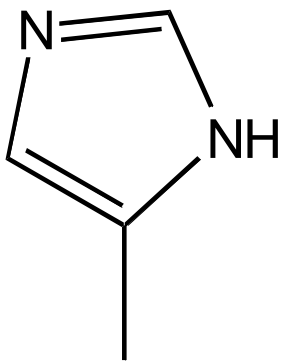
- Missing atoms, missing residues
- Protonation, addition of H-atoms
- Tautomers (His)
- Sidechain conformations (His, Asn, Gln)
- Disulfide bridges



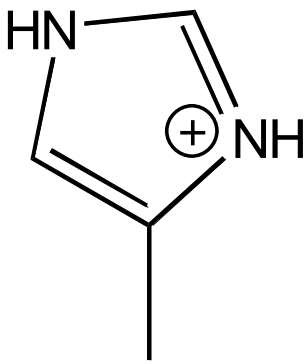
Protonation - Tautomers



His-E
HIE

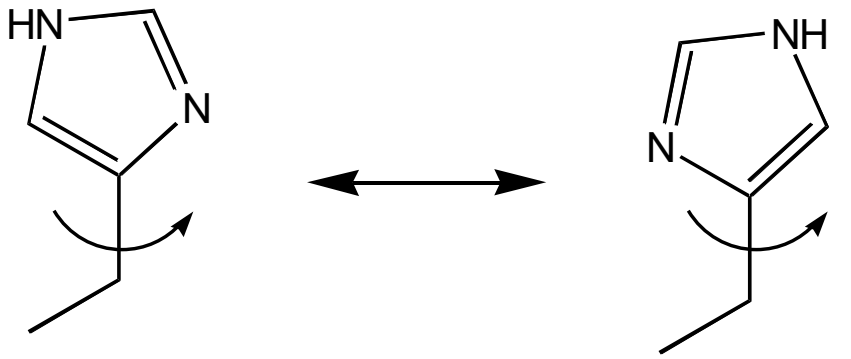


His-D
HID

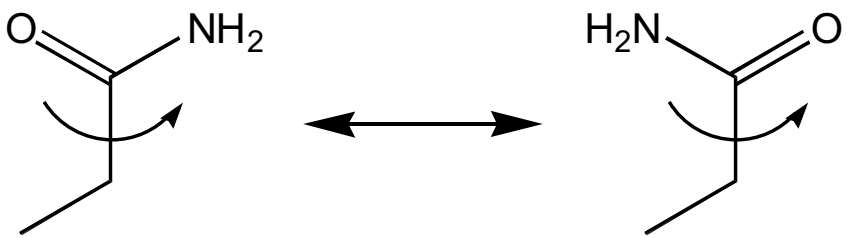


His-P
HIP

Conformations of certain sidechains – problems distinguishing C, N und O

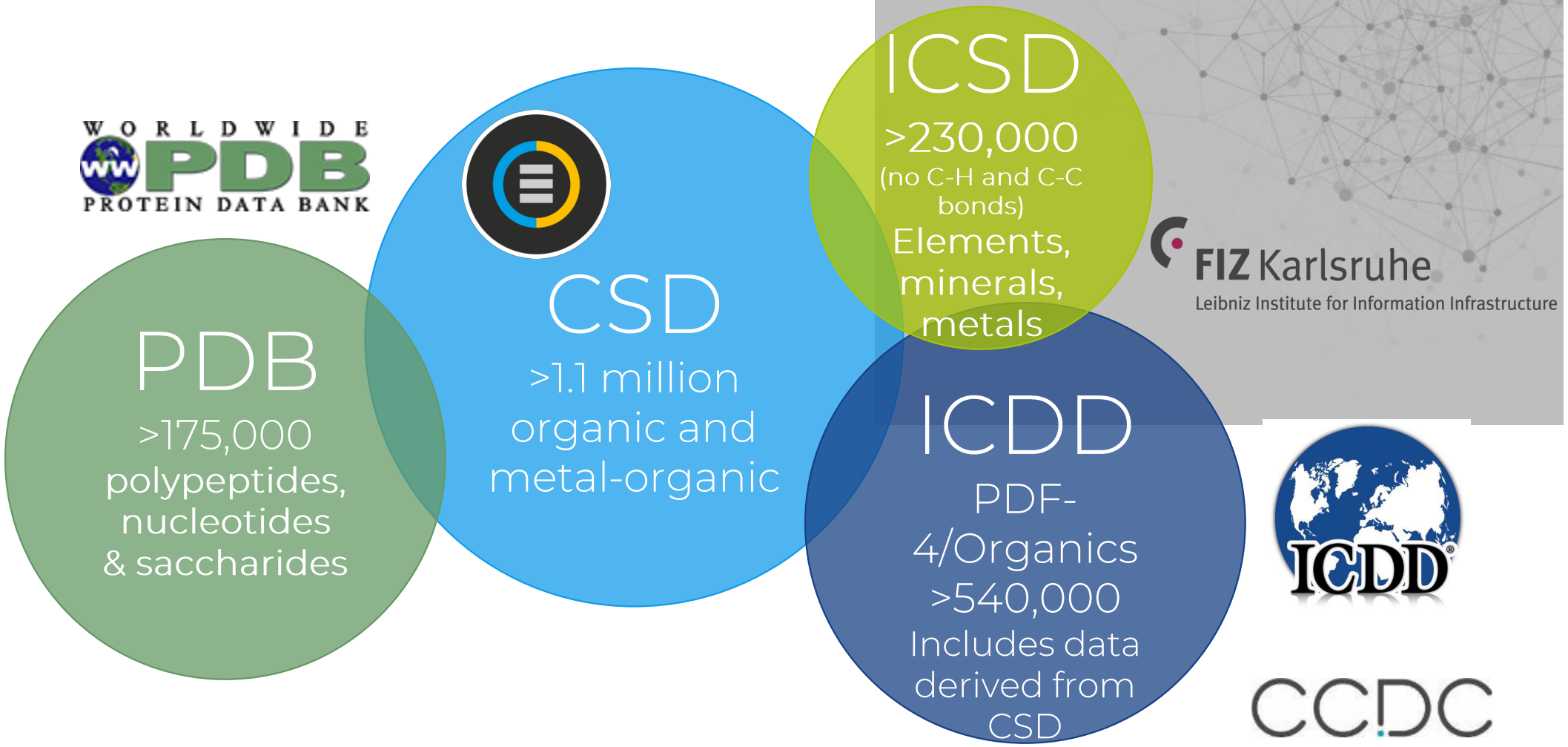


His



Gln oder Asn

Structural databases



https://www.ccdc.cam.ac.uk

Cambridge Structural Database (CSD)

The world repository of small molecule crystal structures.
The CSD records bibliographic, chemical and crystallographic information for organic and metal-organic compounds, whose 3D structures have been determined using X-ray or neutron diffraction.