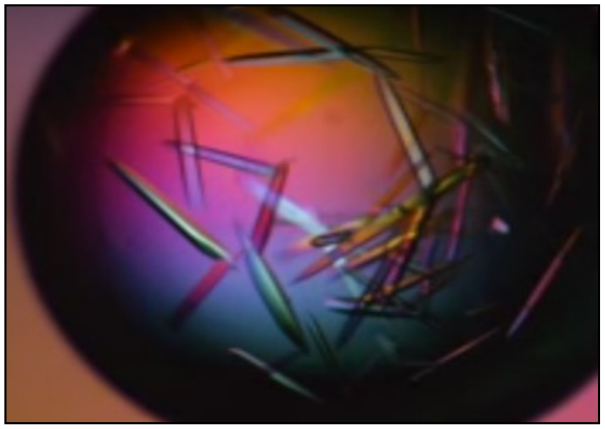
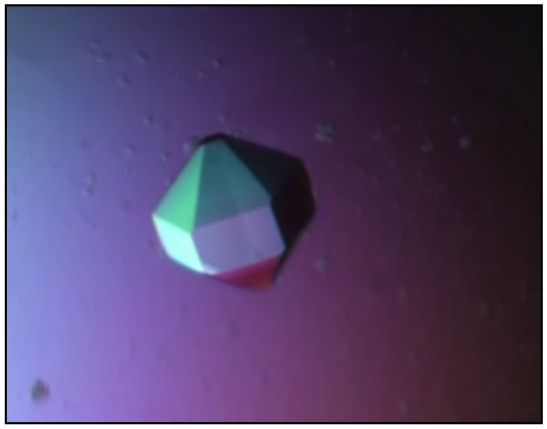
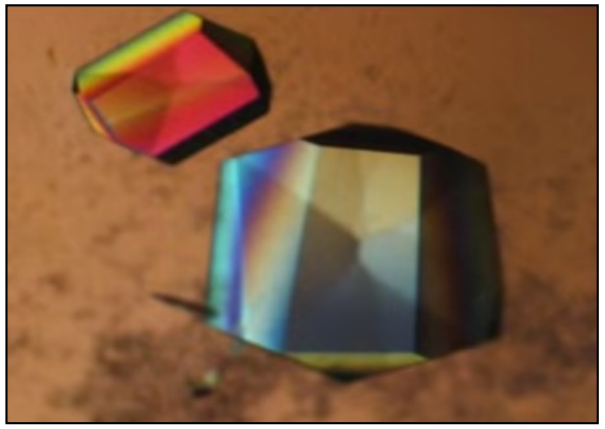
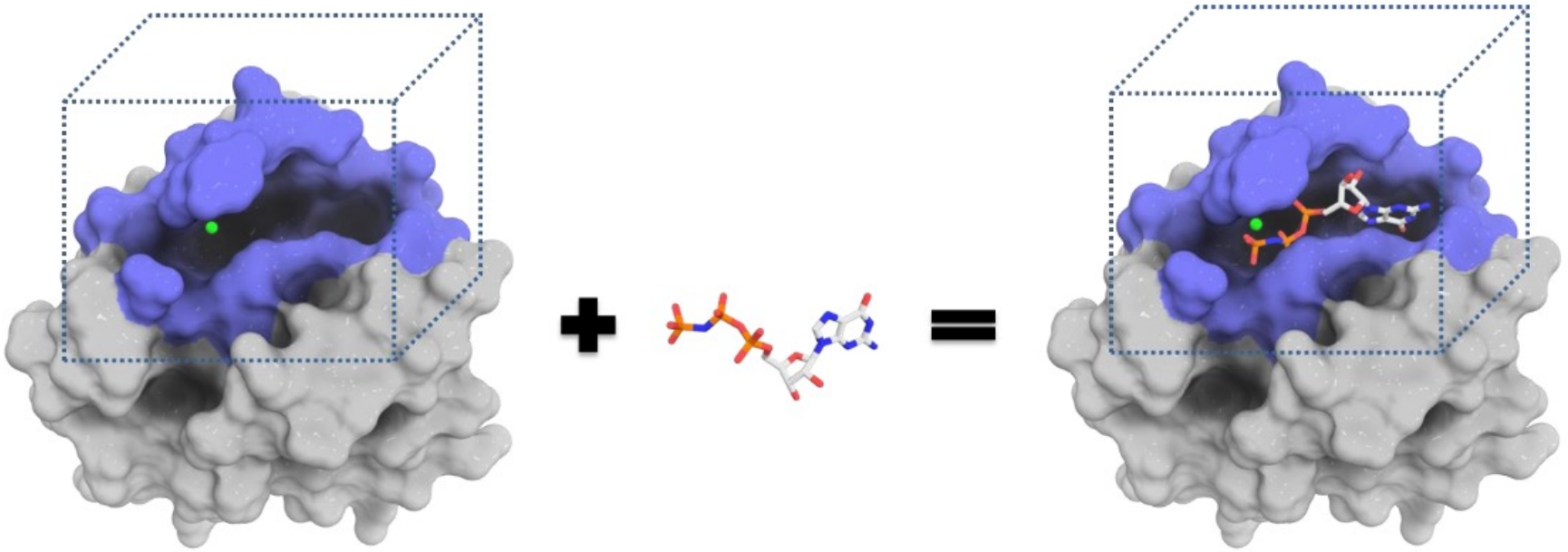
Silvia Onesti
silvia.onesti@elettra.eu

Macromolecular Crystallography

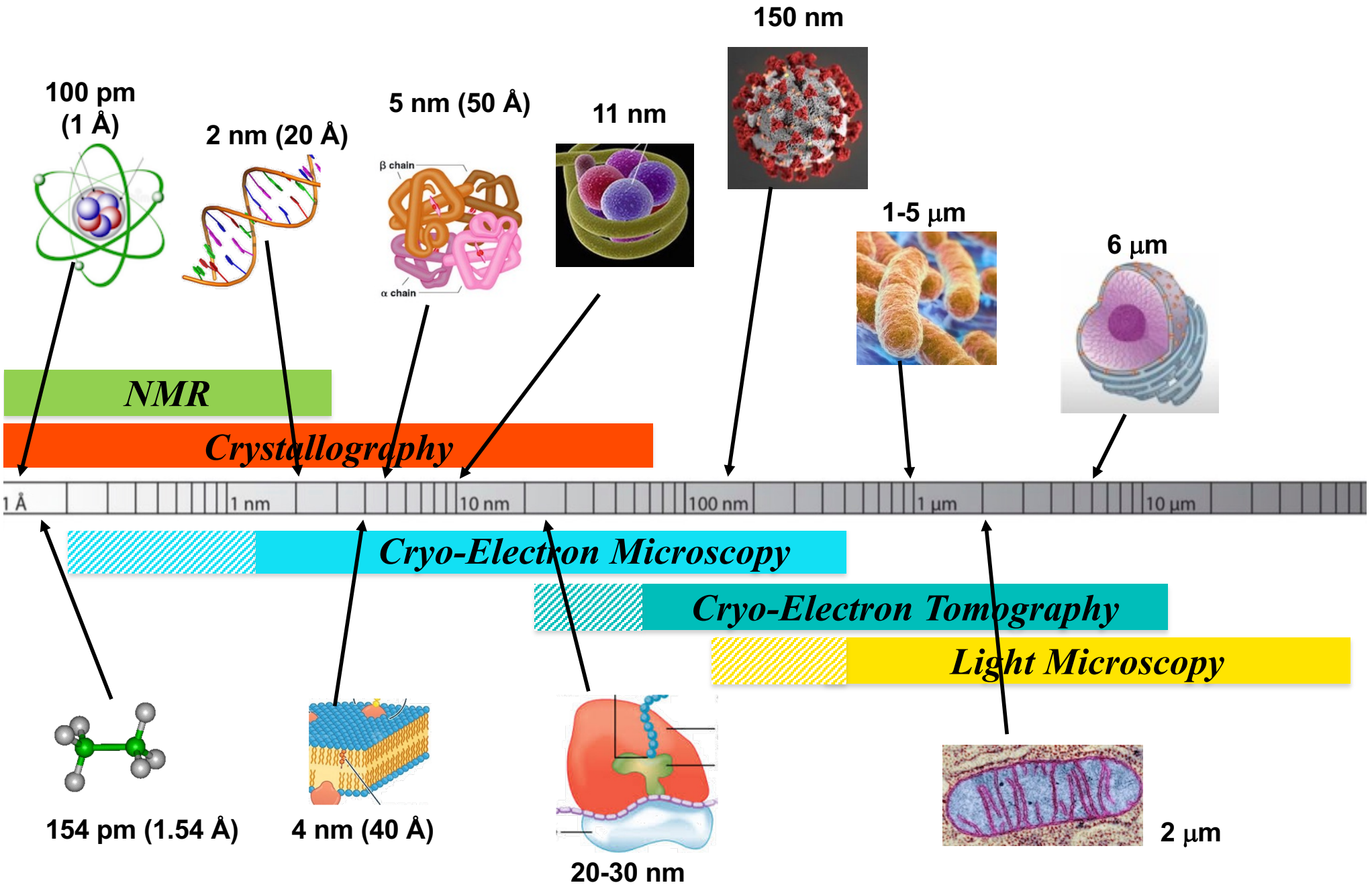




Why Structural Biology?

- Understanding basic biological processes at the molecular/chemical level
- Provide a framework for rational structure-based drug design:
  - virtual screening
  - optimisation of lead compounds
  - fragment-based drug design



The scale of life



Structural Biology: a historical prospective

- 1934 first diffraction pattern from a protein
- 1953 double helical structure of DNA by fiber diffraction
- 1960s atomic structures of myoglobin, haemoglobin, lysozyme (an enzyme) by macromolecular crystallography (MX)

Then exponential growth of the number of protein structures...

late 1980s

- 2-D NMR used to determine structures of small proteins
- structure determination of membrane proteins by MX
- structure determination of large complexes by MX

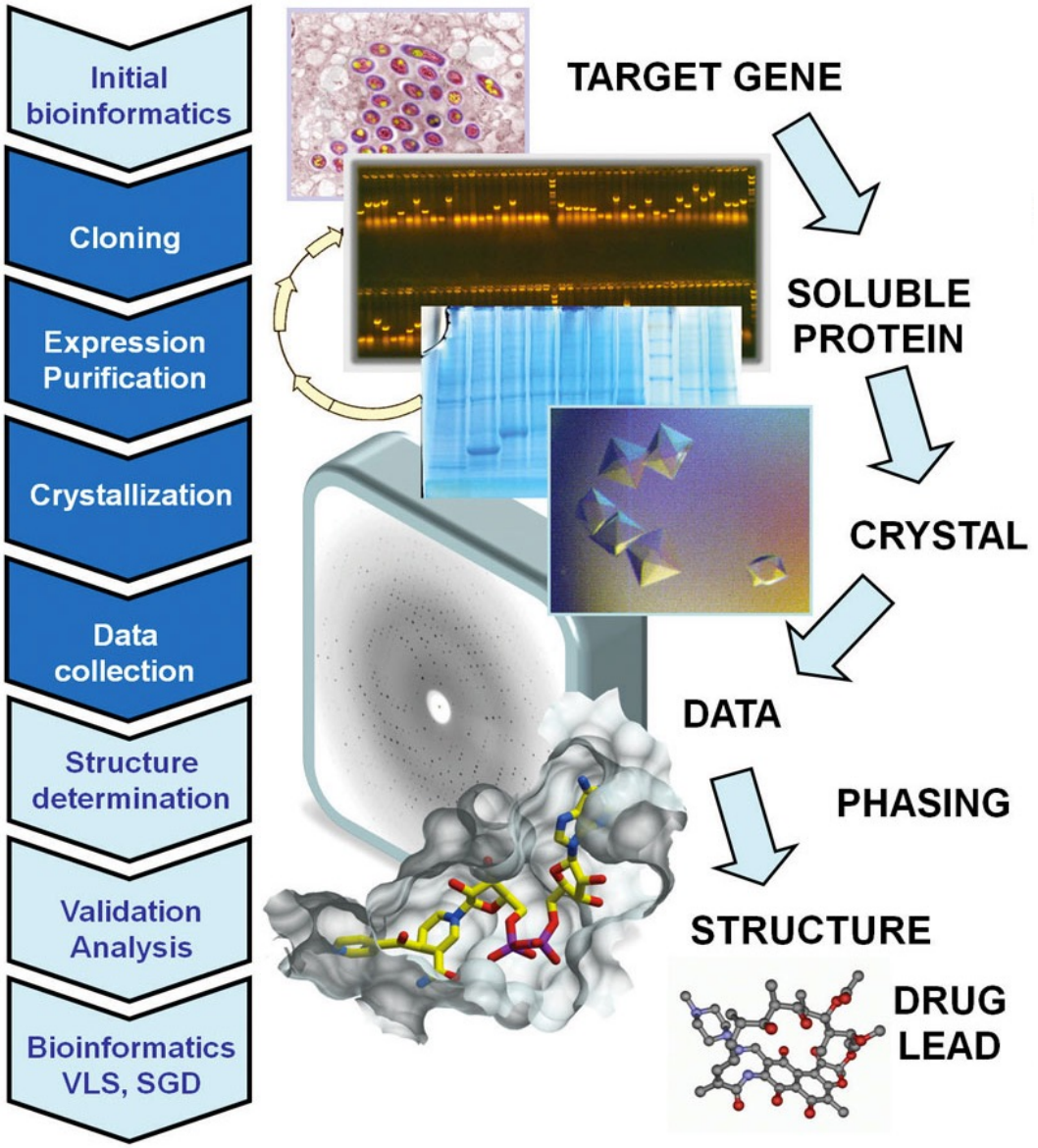
late 1990s

- single particle cryo electron-microscopy used to determine low resolution images of large complexes (eg. ribosome)
- structures of large assemblies at atomic resolution by MX (the nucleosome, the ribosome, large viruses, RNAPs, etc..)

2010s

- EM revolution – structures of single molecules to atomic resolution

The crystallographic pipeline

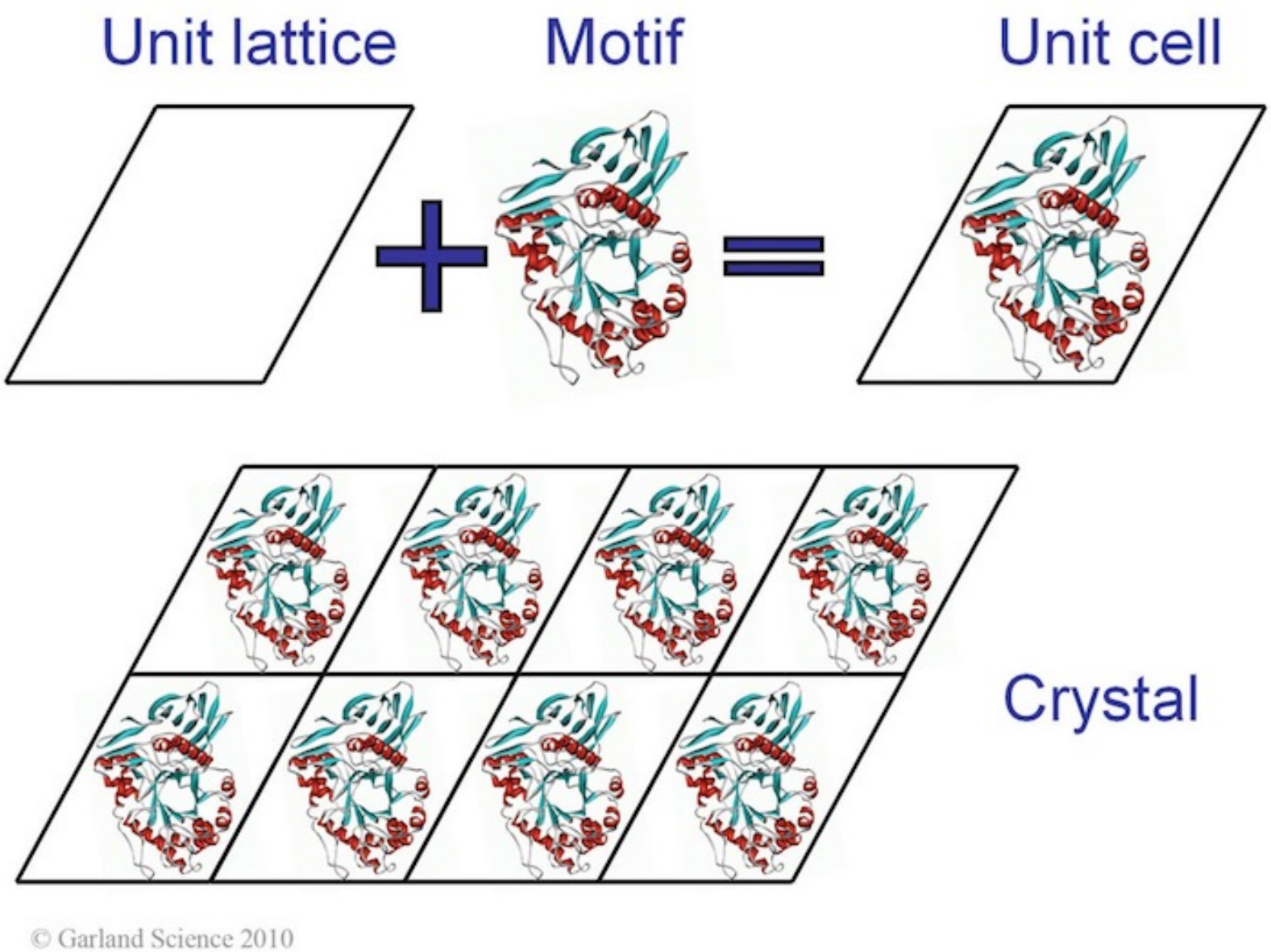


From Gene to Crystal
- lab!!!

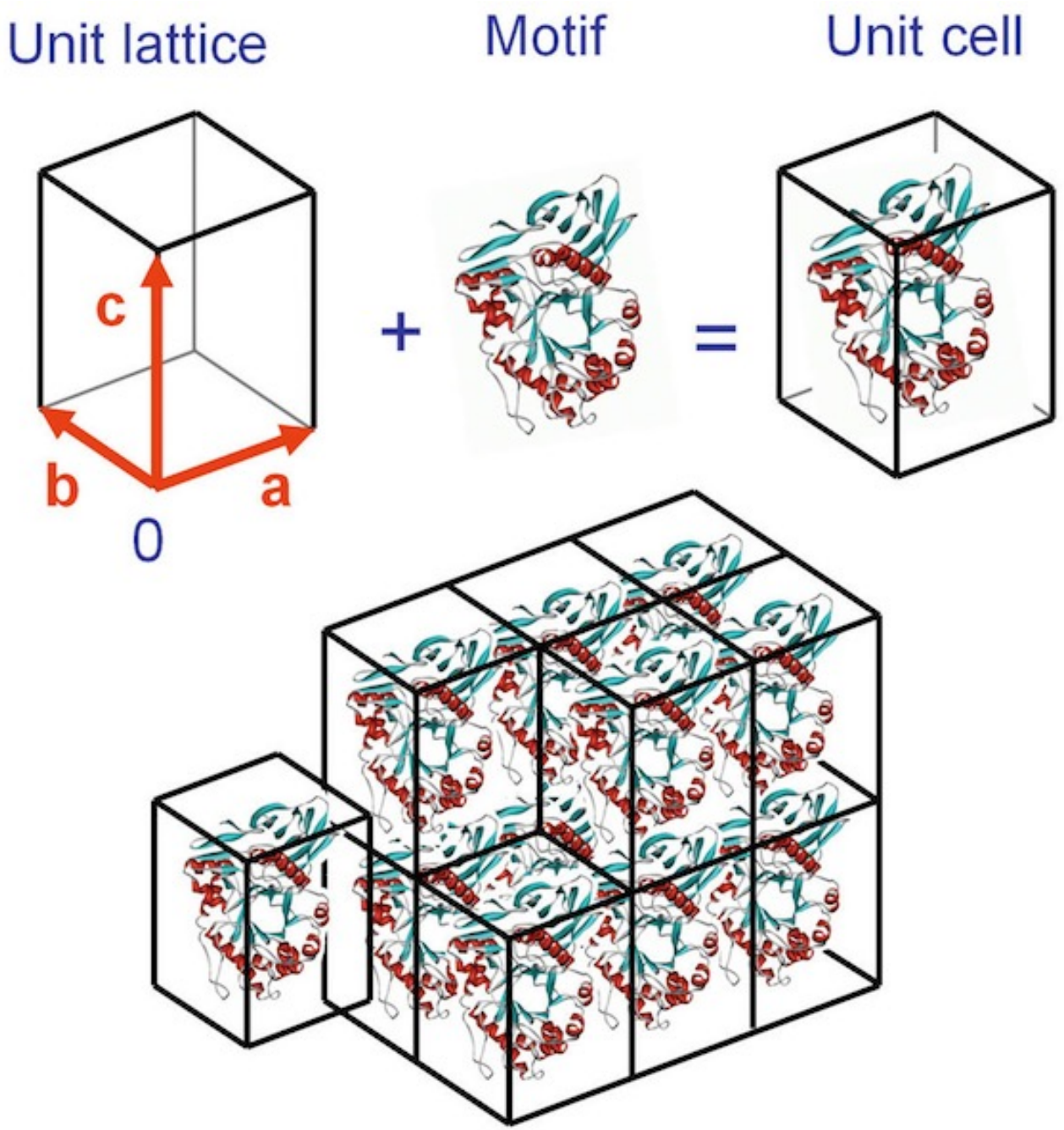
< bottleneck

From Crystal to Structure
- synchrotron
- computer

How does a protein crystal look like?



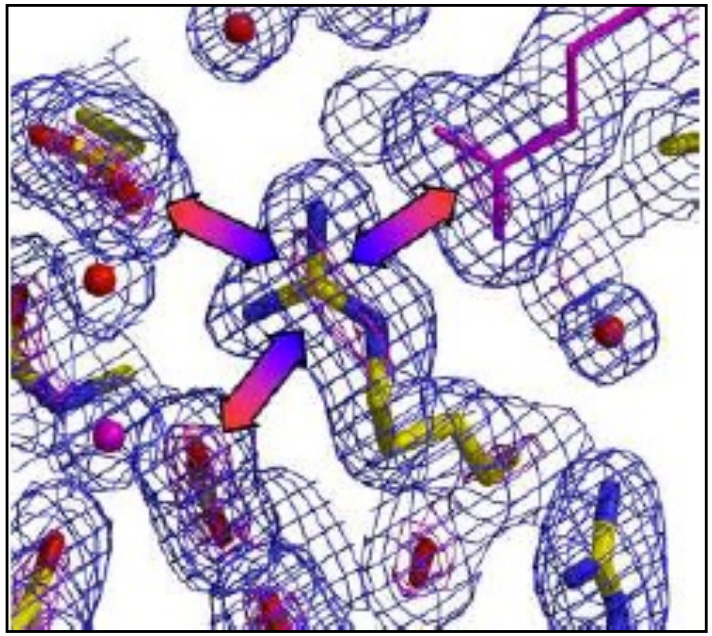
In 3D

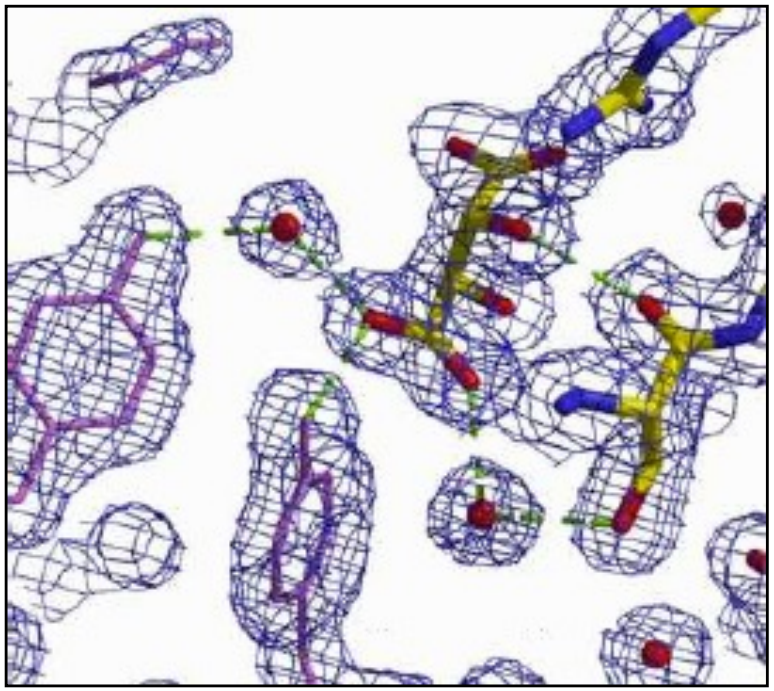


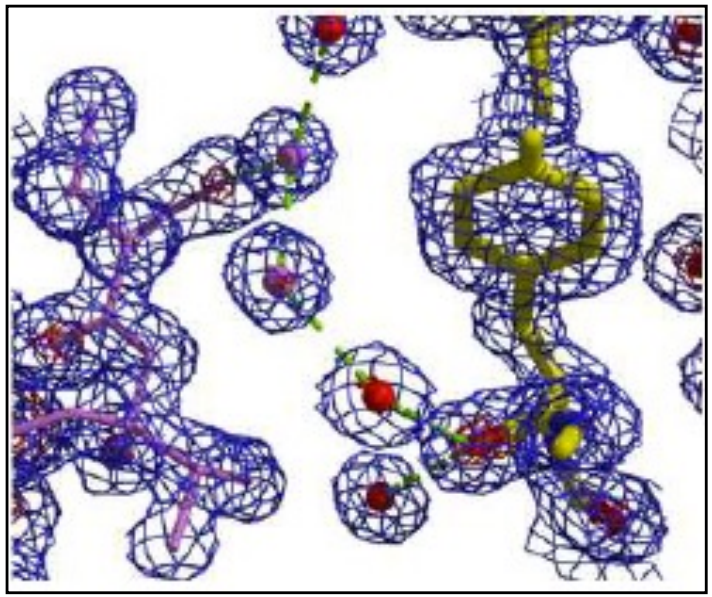
Crystal contacts

The fragile protein crystals are held together by very few & weak interactions:

- H-bonds
- VdW interactions
- salt bridges
- often mediated by H_2O





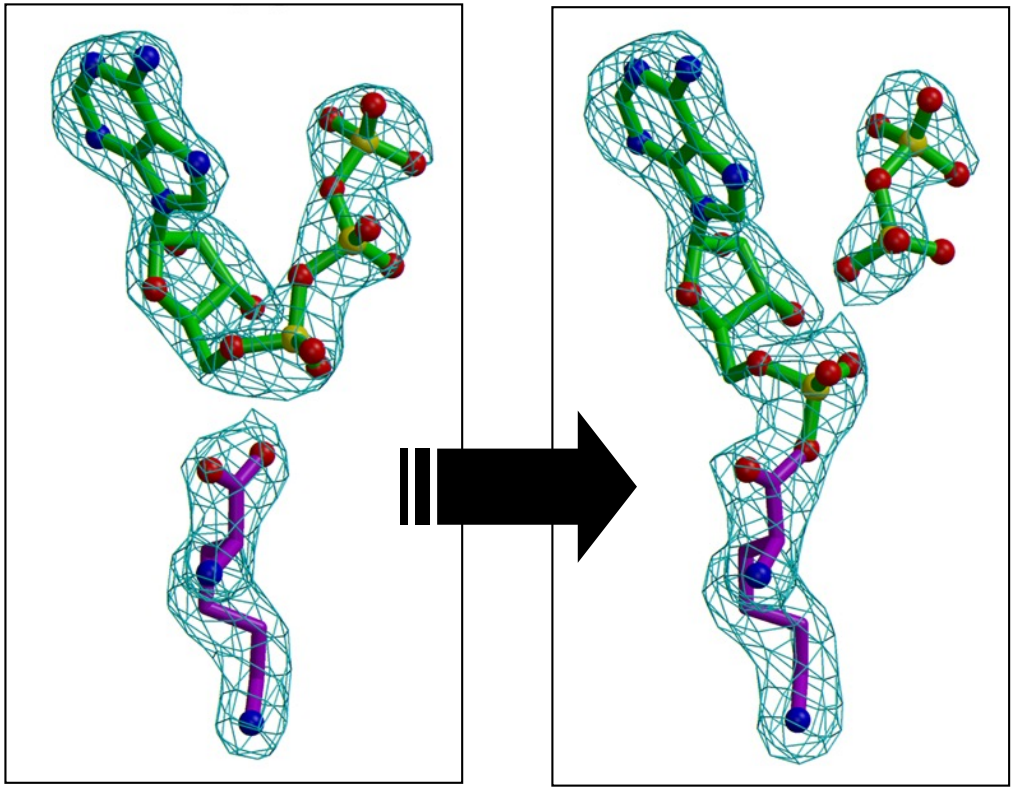


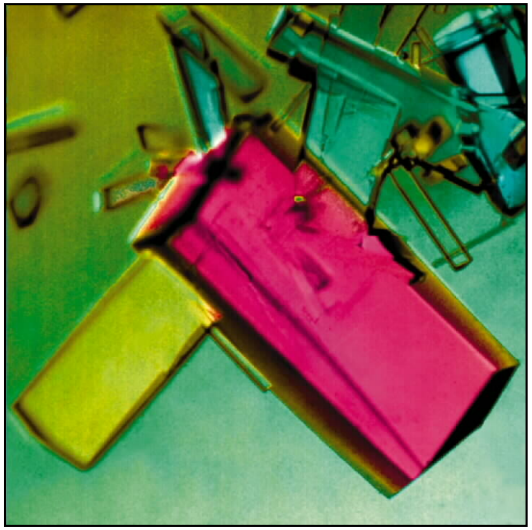
Does crystallisation affects the structure?

- The forces that hold molecules in a lattice are very much weaker than those that hold protein structures together, so gross conformational changes triggered by crystallisation are unlikely (but they do occur). However the crystal lattice may favour a conformation that is not dominant in solution.

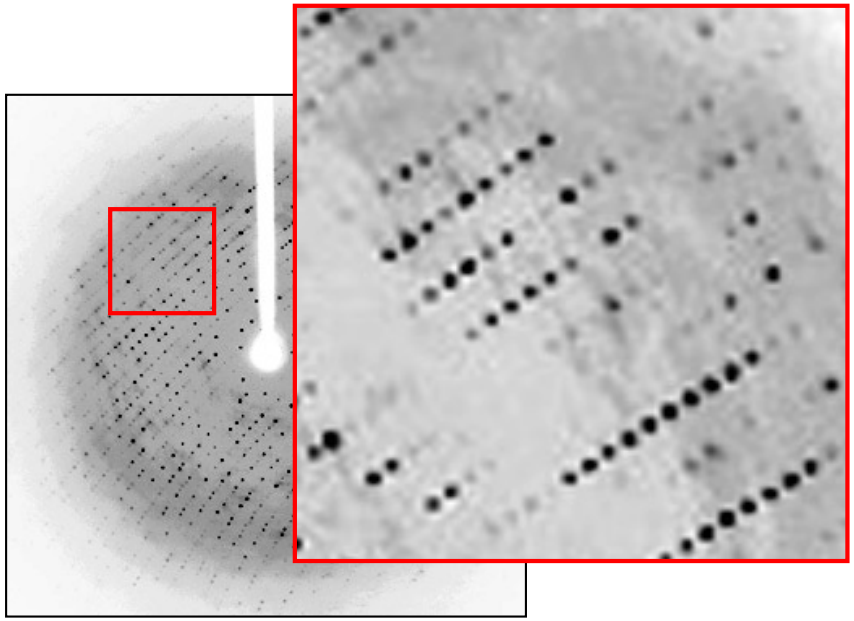
- Proteins crystallised in different crystal forms are often identical or almost identical.

- Some enzymes retain their biological activity within the crystal (a strong indication of a native-like structure).



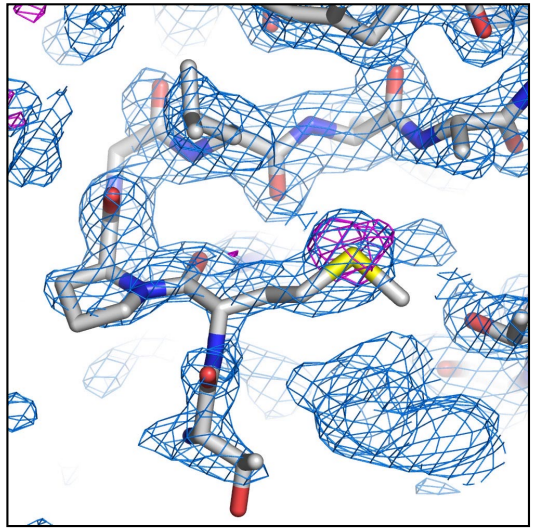


Crystal

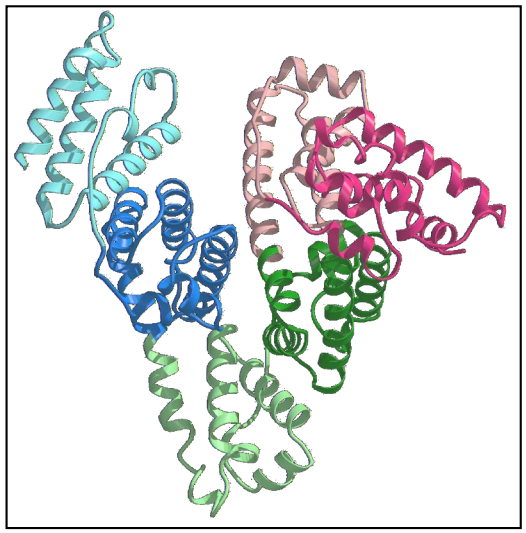


X-ray Diffraction pattern

Fourier Transform

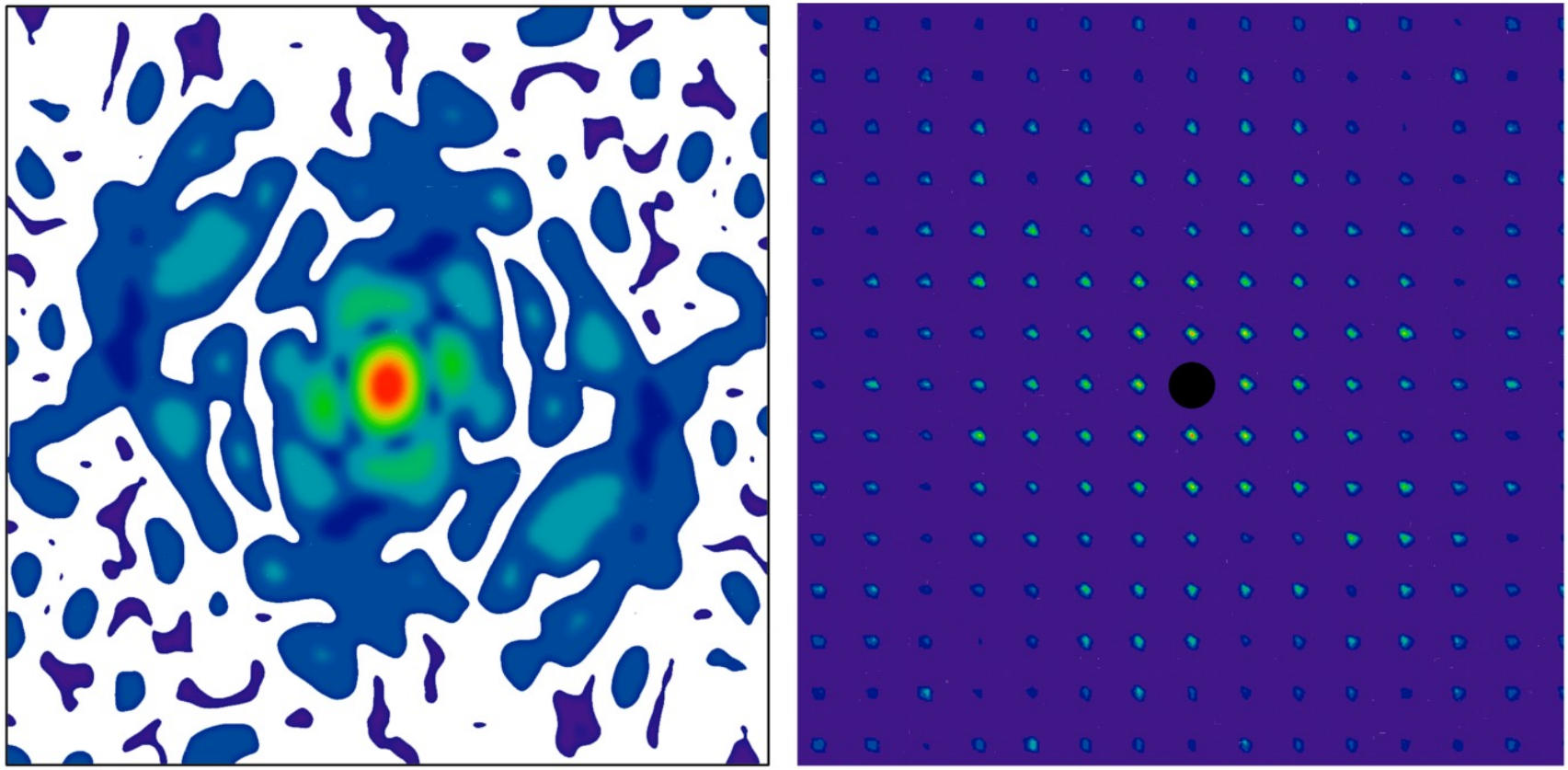


Electron density map



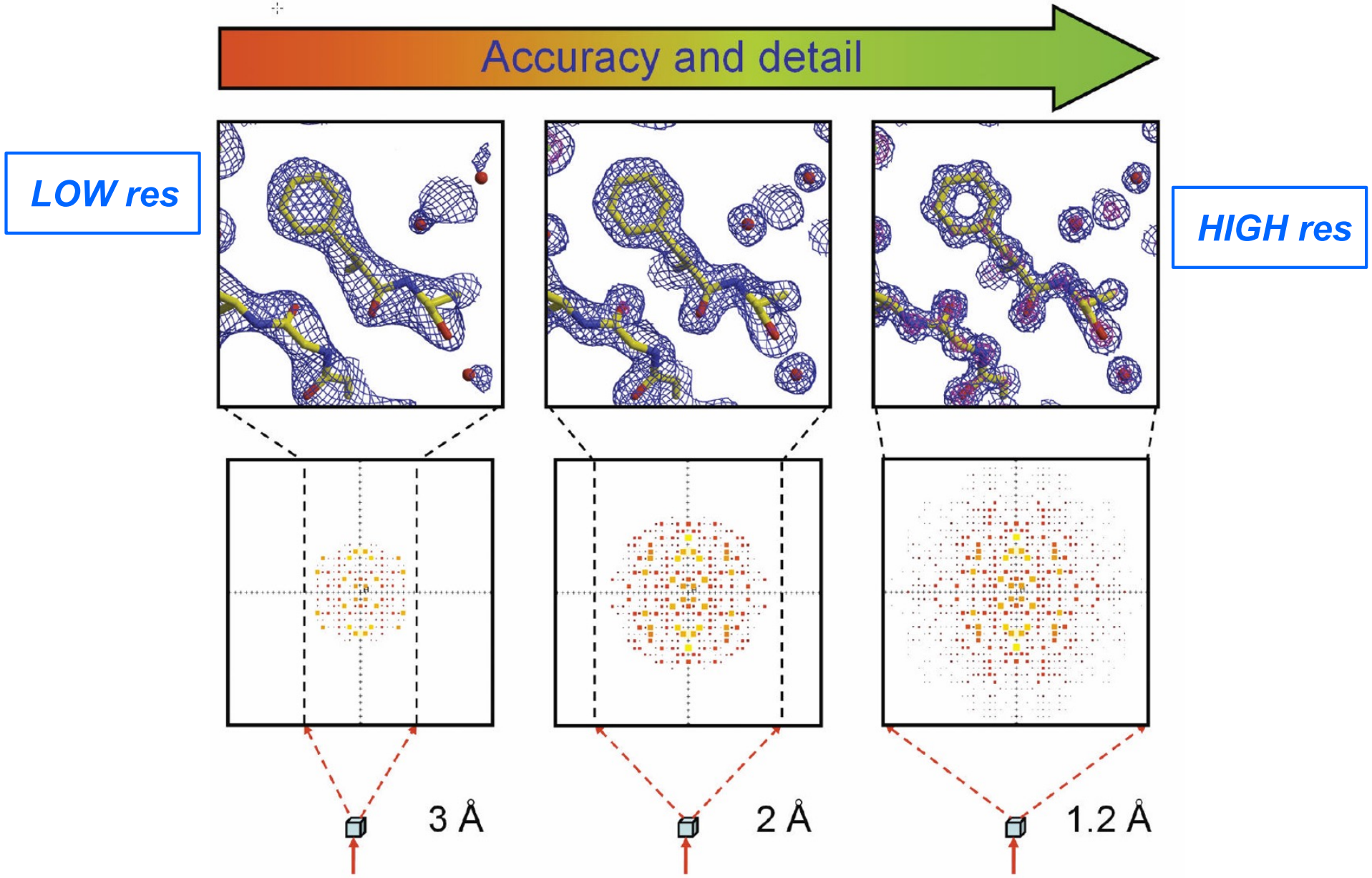
Protein structure

Diffraction from a protein crystal



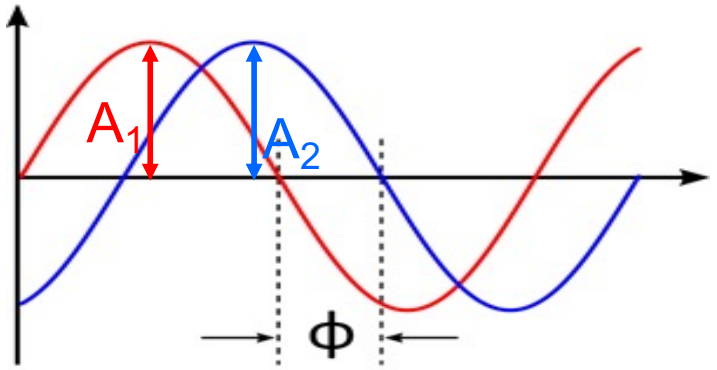
The fringe function (i.e. the sampling – **a***, **b***, **c***) depends on the lattice (**a**, **b**, **c**) and but the intensities depends on the underlying molecular pattern, and thus on the molecule.

Resolution



The phase problem

We can measure the intensity and thus get the amplitude ($I=A^2$), but we loose the information about the relative phase of the wave.

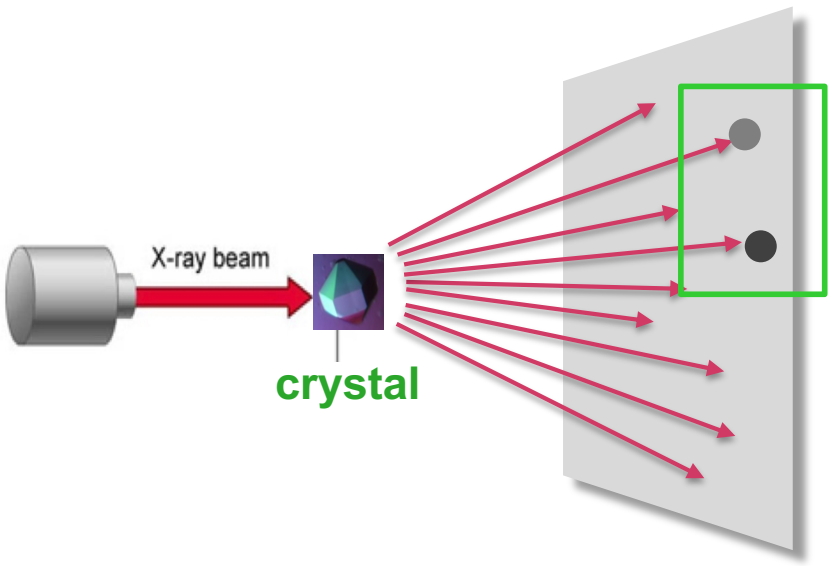


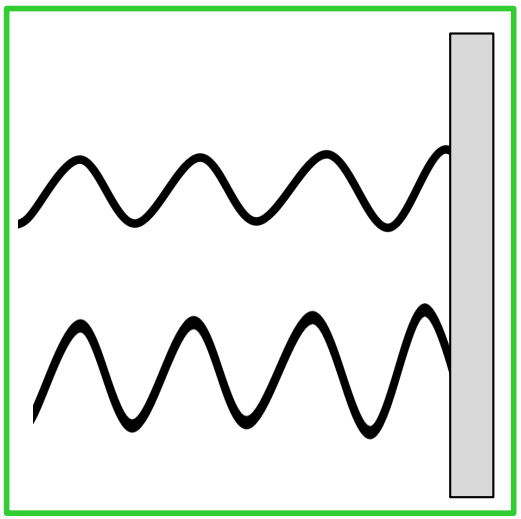
Same Amplitude $A_1=A_2$

Different phases Φ

$\Phi_1=0$

$\Phi_2=-90°$





$A_1=50$

$\Phi_1=90°$??

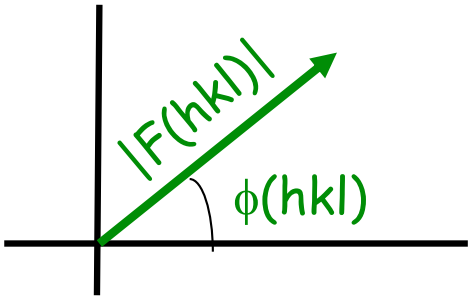
$A_2=80$

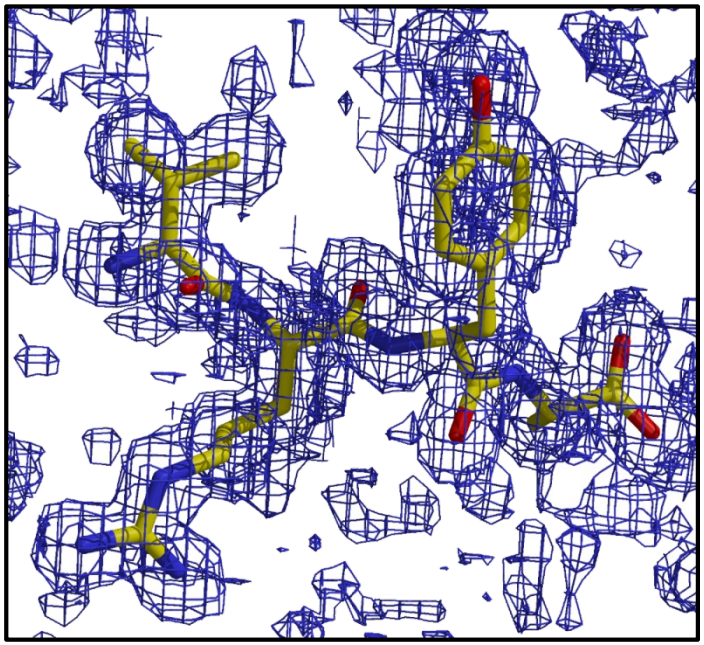
$\Phi_1=180°$??

close-up

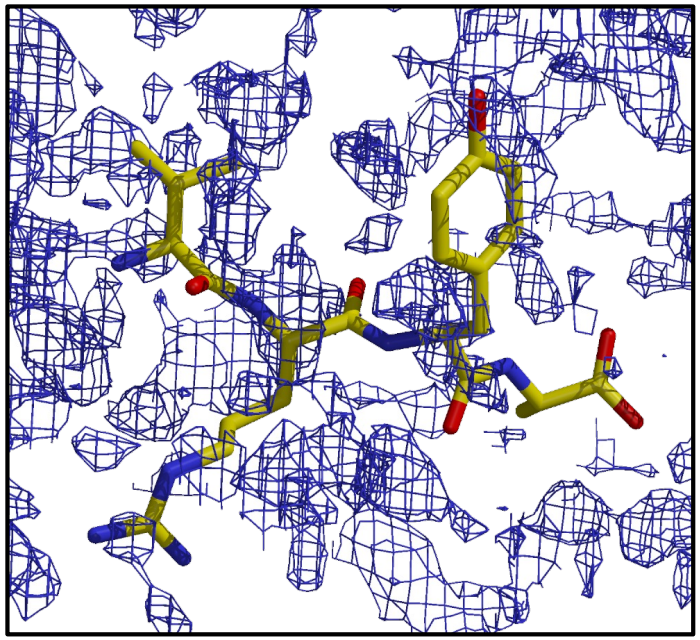
Phases are more important than amplitudes

$$\rho\,(xyz) = \sum_h \sum_k \sum_l \boldsymbol{F}(hkl)\,\exp[-2\pi i\,(hx + ky + lz)$$





Correct phases
Random amplitudes



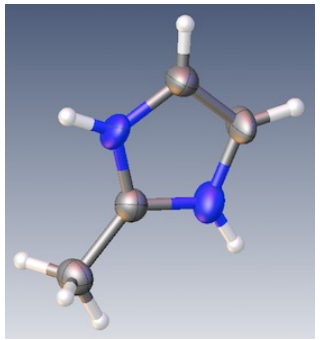
Correct amplitudes
Random phase

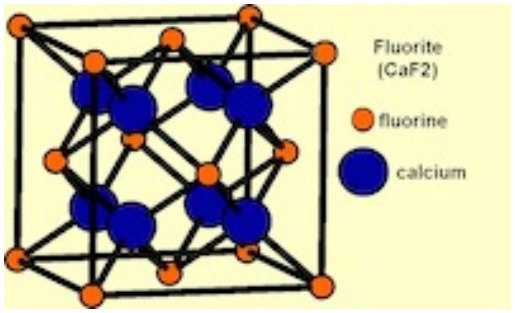
How to solve the phase problem?

The phase problem is common to all the X-ray crystallography methods.

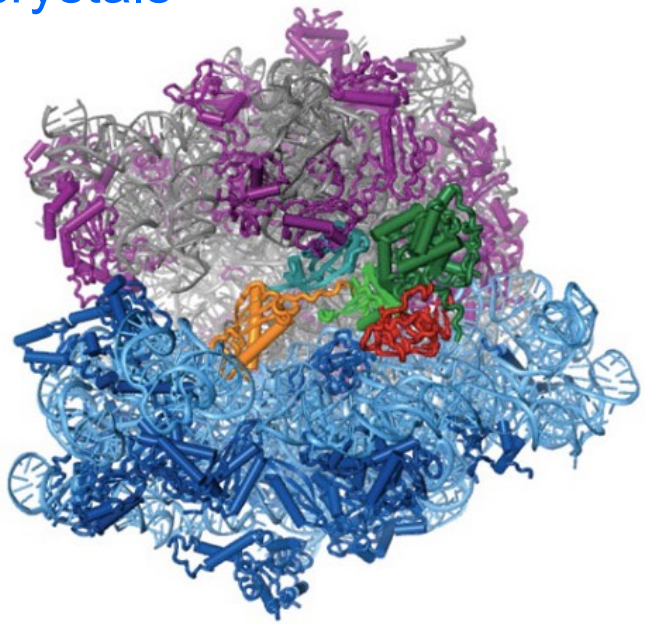
But the methods to solve it depends on the specific problem:

- it is much easier for small molecule crystals (mineralogy, chemistry, pharmacology, etc...)

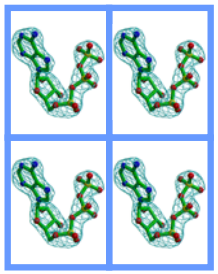




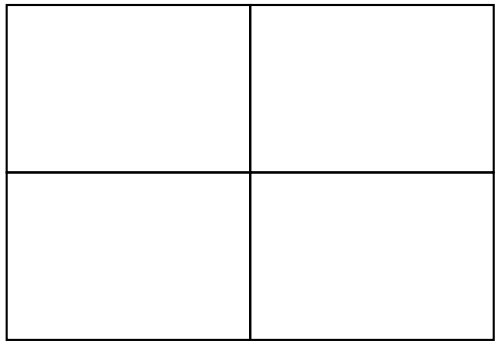
- it is much more difficult for protein/large macromolecular crystals



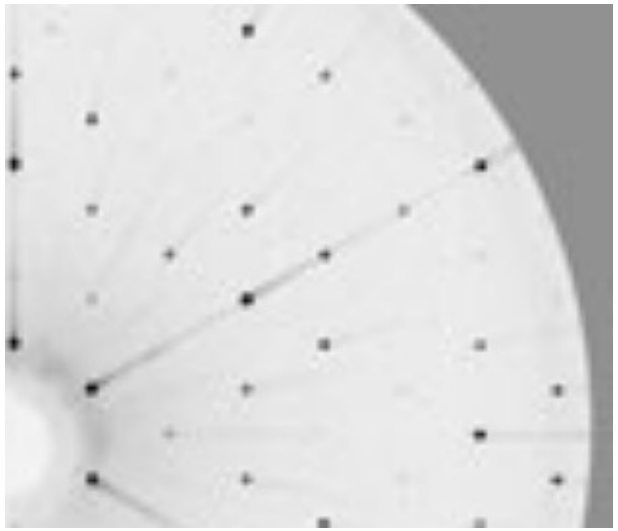
Small molecule crystal



small unit cell

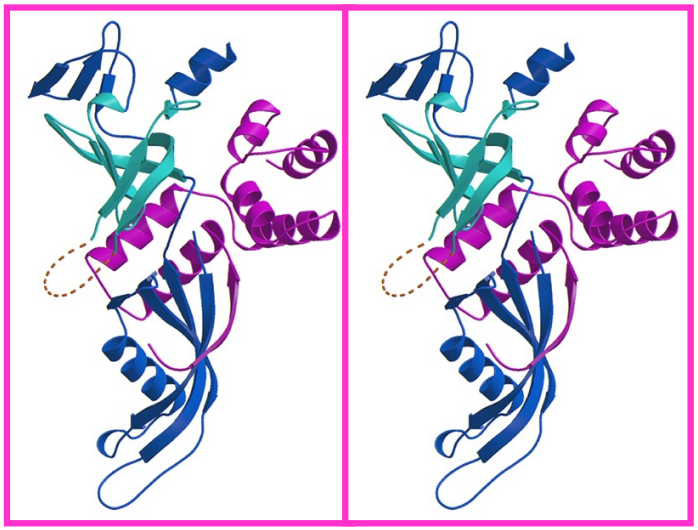


coarse fringe function

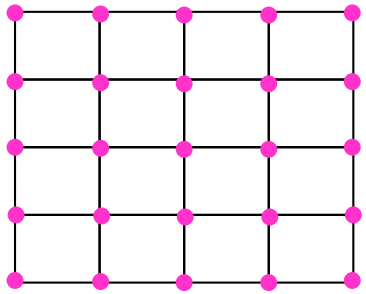


- a few, strong reflections
- diffract to high resolution

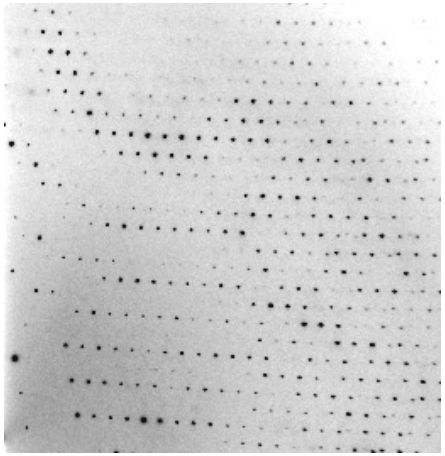
Protein crystal



large unit cell



fine fringe function



- many, weak reflections
- diffract to low resolution

How to solve macromolecular structures

MIR (multiple isomorphous replacement)

Older method (Cambridge, 60') – relies on binding "heavy" atoms to the crystal and compare he diffraction pattern to the native. Trial and error search for good heavy atoms, it may take longer to get it right

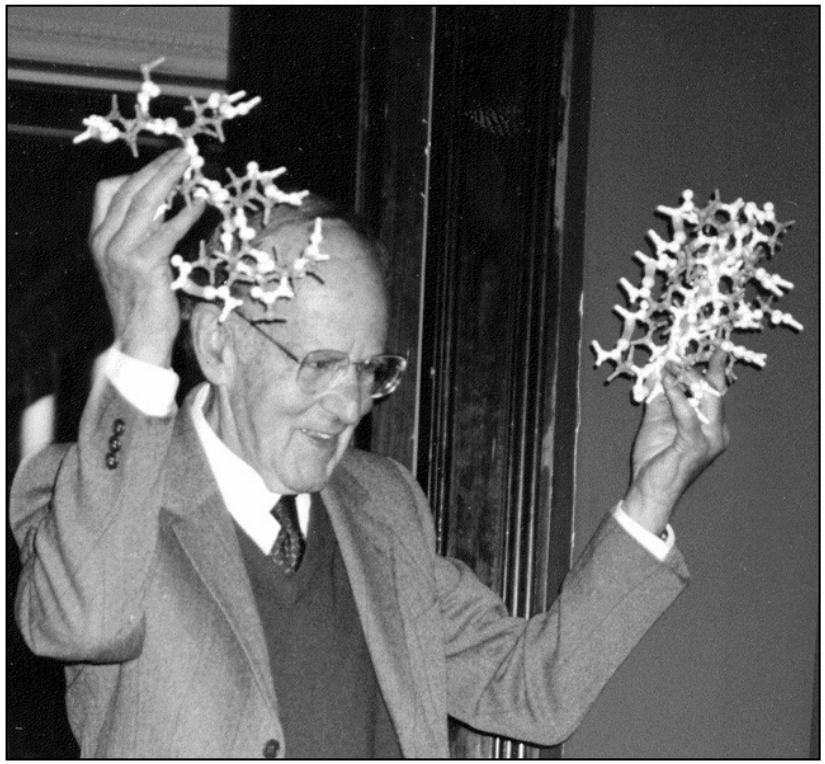
MR (molecular replacement)

Older method (Cambridge, 70'-80') – relies on the expected similarity between the protein and another whose structure is known. Cannot solve de novo structures. Requires high homology (30% sequence identity?)

MAD (multiwavelength anomalous dispersion)

Relies on the absorption of specific wavelengths due to electronic transitions within the atom core. Similar to MIR but generally far quicker and more accurate. Requires high specification synchrotron radiation.

Max Perutz (1914-2002), the inventor of MIR



In 1953, Perutz showed that the diffracted X-rays from protein crystals could be phased by comparing the patterns from crystals of the protein with and without heavy atoms attached.

In 1959, he employed this method to determine the molecular structure of hemoglobin. This work resulted in his sharing with John Kendrew the 1962 Nobel Prize for Chemistry.

Multiple Isomorphous Replacement (MIR)

Introduce "heavy" atoms (H) at a relatively small number of sites on the protein molecules

Collect data from crystals of the protein alone (P), and crystals of the protein:heavy-atom complex (PH)

Compare the two data-sets and solve the H structure (i.e. determine *xyz* for each of the heavy atoms bound to the protein) as if it were a small molecule structure

Calculate the heavy atom contribution $\mathbf{F}_H$ (in modulus and phase) to the protein:heavy-atom scattering, for each *hkl*.

Combine all those information to put constraints on the protein phases

One heavy atom derivative is not enough to solve the structure – at least two are needed.

Error treatment in MIR

The real breakthrough in using MIR for the determination of protein structures came when people learned how to deal with errors.

Blow D.M. & Crick F.H.C. (1959) “The treatment of errors in the isomorphous replacement method”. *Acta Crystallogr.* **12**, 794-802

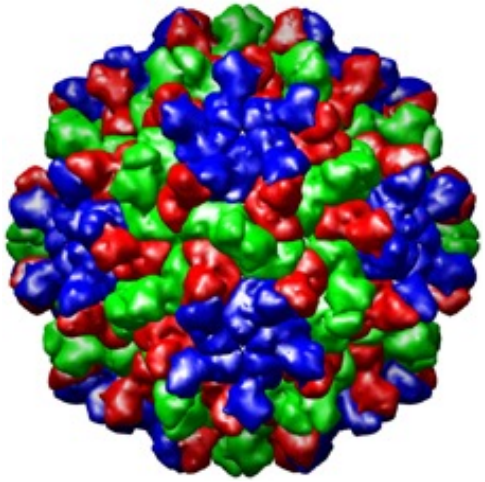
David Blow
(1931-2004)

Francis Crick
(1916-2004)

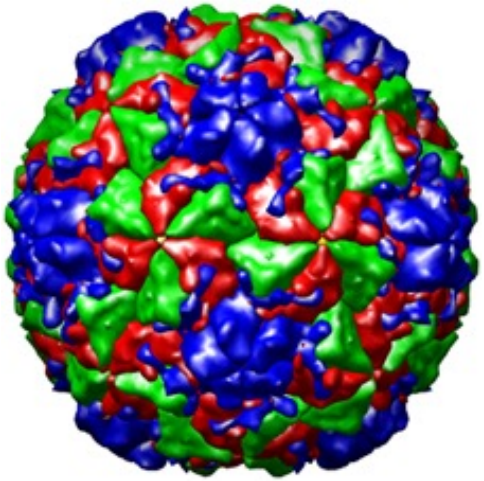
Error treatment is very complex – but it was an essential step in solving protein structures.

Molecular Replacement (MR)

The Molecular replacement method was mostly developed by Michael Rossmann. He used the structure of Tomato Bushy Stunt Virus, a plant virus, to determine the crystal structure of the Human Rhinovirus 14 (the common cold virus) in the early 80's.



Tomato Bushy Stunt Virus

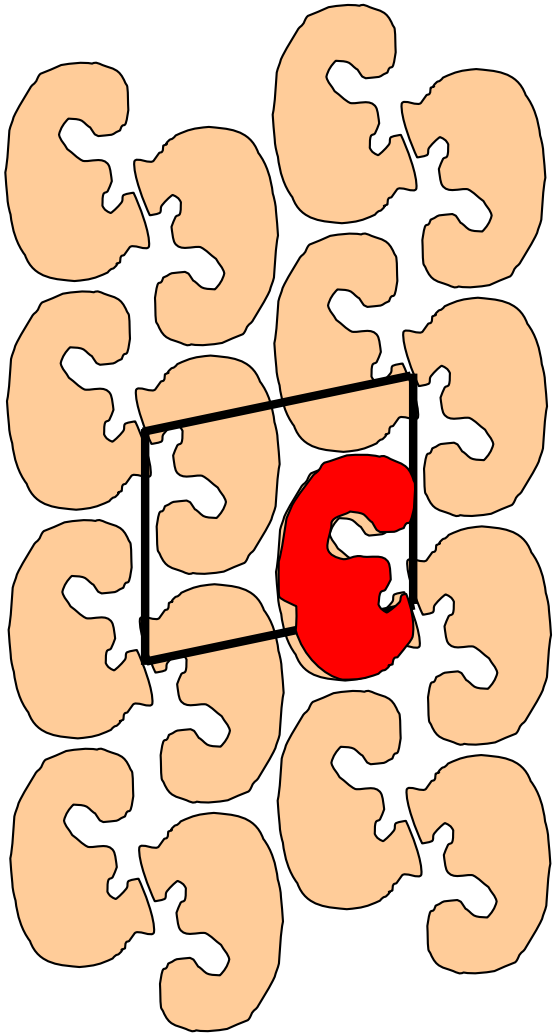


Human Rhinovirus 14

However the theoretical basis were developed much earlier:

Rossman, M. G. and Blow, D. M. (1962). *ActaCryst*. 15:24-31.

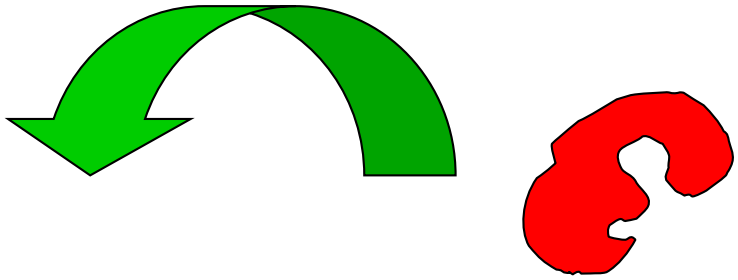
What is molecular replacement?



Unknown crystal structure

In molecular replacement a crystal structure is solved by using a known structure that is similar to the unknown structure as a search model. Initial phases are calculated from the correctly oriented and positioned model, as an approximation to the real phases.

Finding the movement corresponding to the green arrow constitutes the molecular replacement problem.



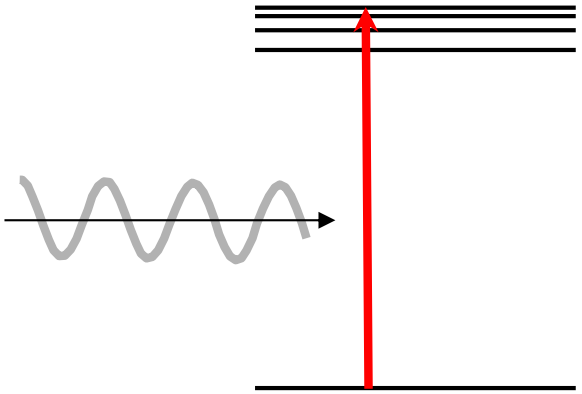
Search model

Anomalous scattering

- The diffraction pattern has a centre of symmetry
- One cannot distinguish a molecule from its mirror image

If the energy of photons in an incident X-ray beam is close to that of the binding energy of electrons, there will be small differences that will break the symmetry of the diffraction pattern, as a result of the "anomalous scattering".

Absorption edge: *wavelength needed to knock an electron out of the atom*



The differences in the intensities can be used to provide information on the phase of the diffracted X-ray beam and also to determine the 'absolute configuration' of the molecule (Bijvoet, 1951).

Multiwavelength anomalous dispersion (MAD)

MAD phasing relies on the fact that some atoms have an absorption edge close to the wavelength of X-rays used for crystallography (0.7-2 Å).

Key point: Close to the absorption edge, small changes in λ give rise to significant changes in the scattered wave (anomalous dispersion)

Using MAD, we can collect multiple "derivatives" from the same crystal just by changing the X-ray wavelength very slightly. Need tunable X-rays sources (synchrotrons!)

Intensity changes (i.e. signals) are smaller than for MIR, but all the data are collected from the same crystal, so there is no non-isomorphism. The results is that MAD phases are much more accurate that MIR phases.

Most of the times now we use SAD (single wavelength)

MAD phasing

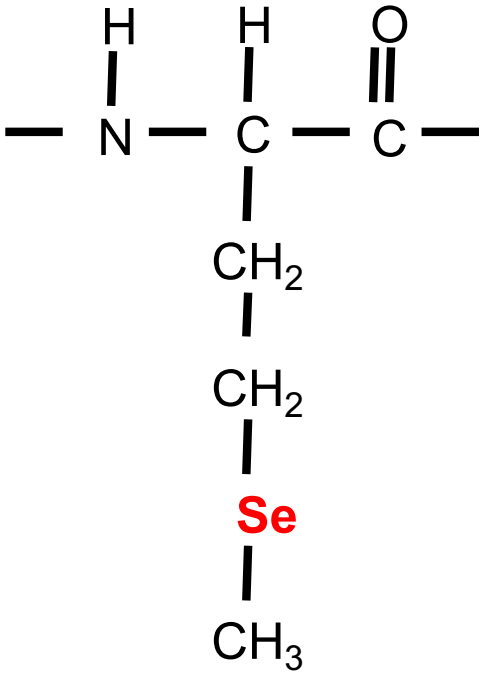
We need anomalous scatterers!

At wavelengths of around 1 Å there are no electronic transitions for the "light" atoms (C, N, H, O, S, P) of biological molecules.

Protein molecules can contain metals such as Zn, Fe and Cu which have accessible absorption edges.

The most common method is to replace Met residues by SeMet. Selenium has an absorption edge at 0.98 Å, by producing recombinant proteins in bacteria that are Met auxotroph and supplying Se-Met.

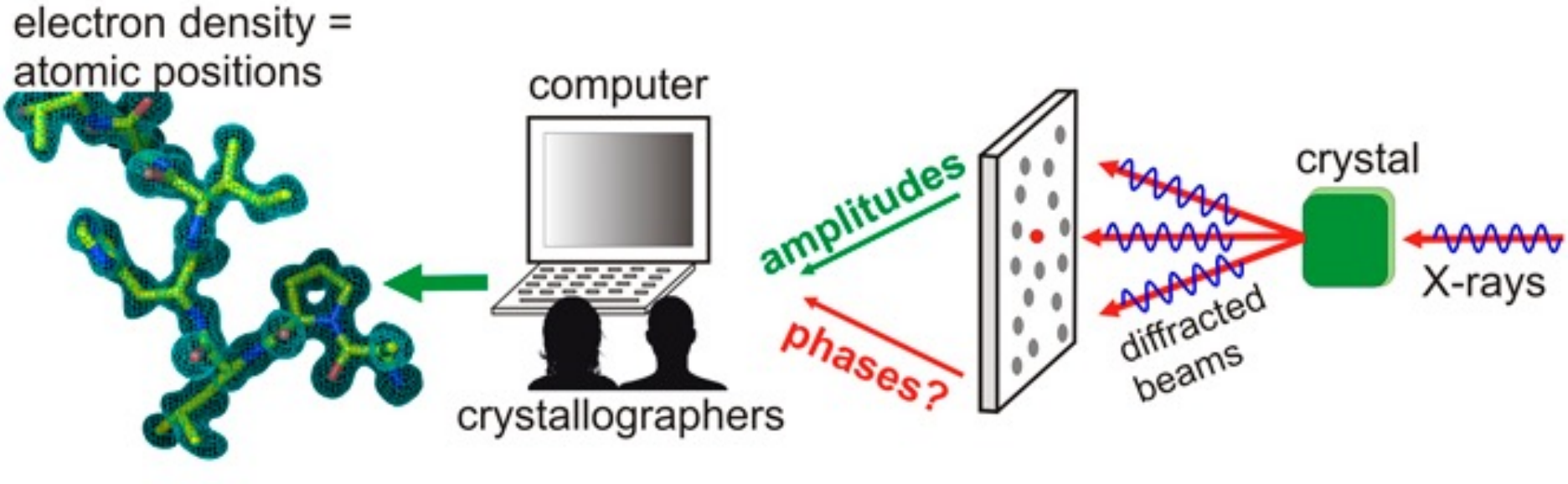
More difficult with eukaryotic expression systems, but not impossible.



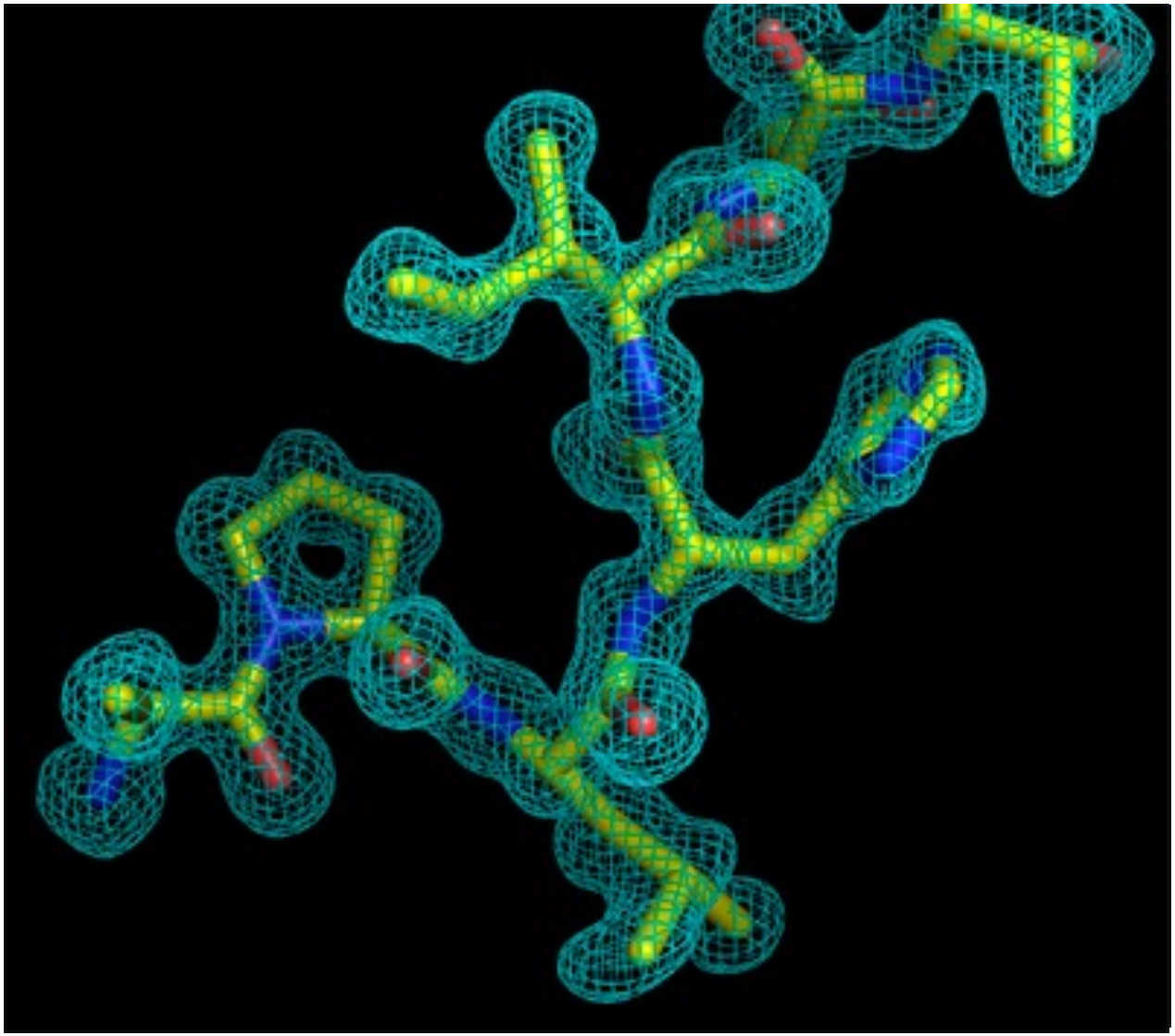
Phase problem solved?
Calculating an electron density map

We measure the **amplitudes** of the diffracted spots.

Once we have decent experimental estimates of the **phases** (from MIR, MR or MAD) we can calculate a Fourier Transform and obtain an electron density map, and from this get a 3D atomic model of our macromolecules.

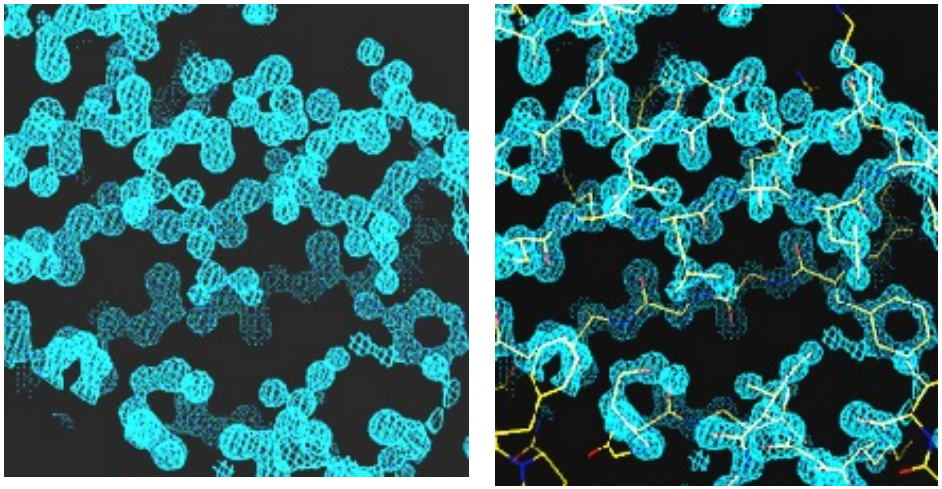


Phase problem solved?
Can you build a polypeptide chain?



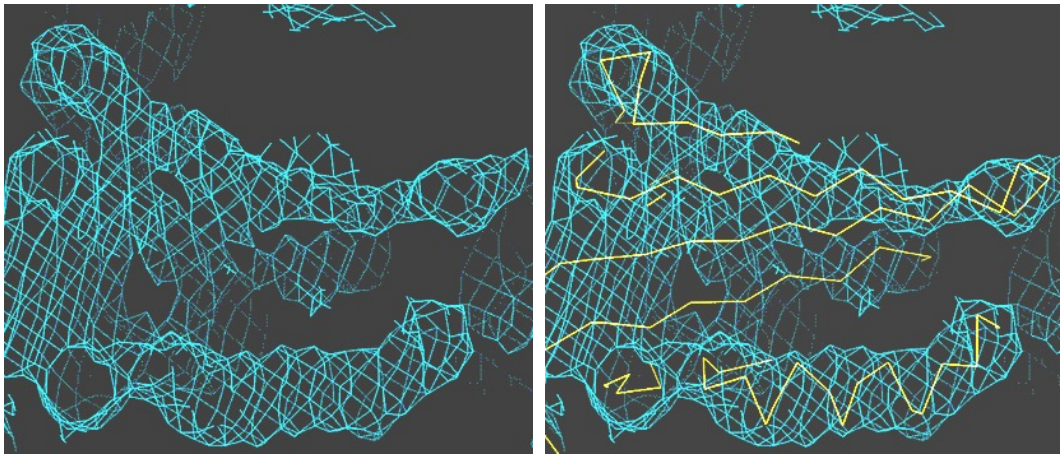
Maps and resolution

The task of model building is to interpret the electron density maps in light of chemical knowledge, basic stereochemistry, chemical sequence, etc...
The level of interpretation depends on the resolution of the map:



Here is a 1 Å map

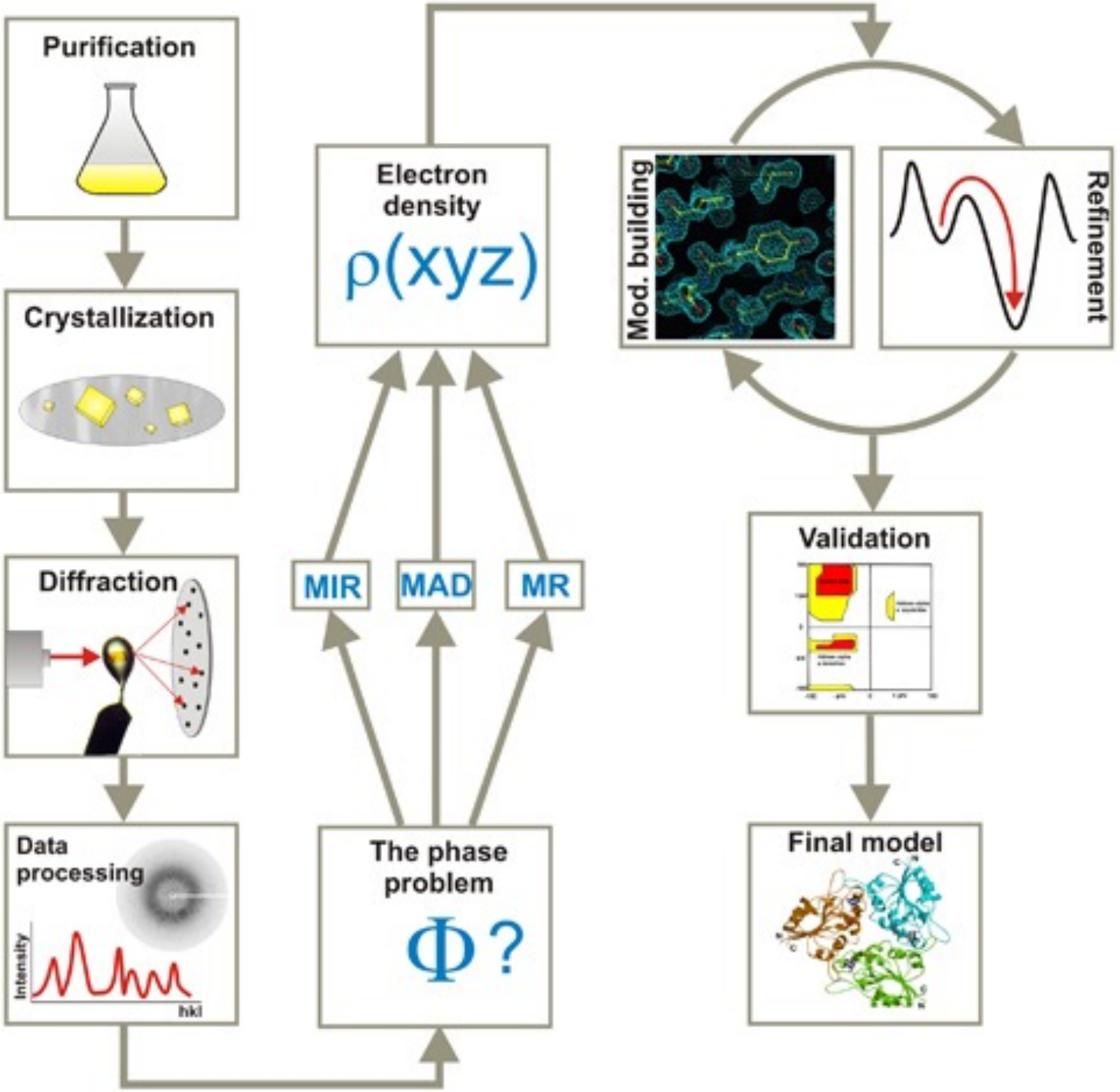
At very high resolution, individual atoms can be seen and fitted in the electron density blobs: the problem therefore is reduced to 'join-the-dots'



Here is a 6 Å map

At very low resolution only large features can be seen - for example helices look like rods and β-sheets can barely be detected.

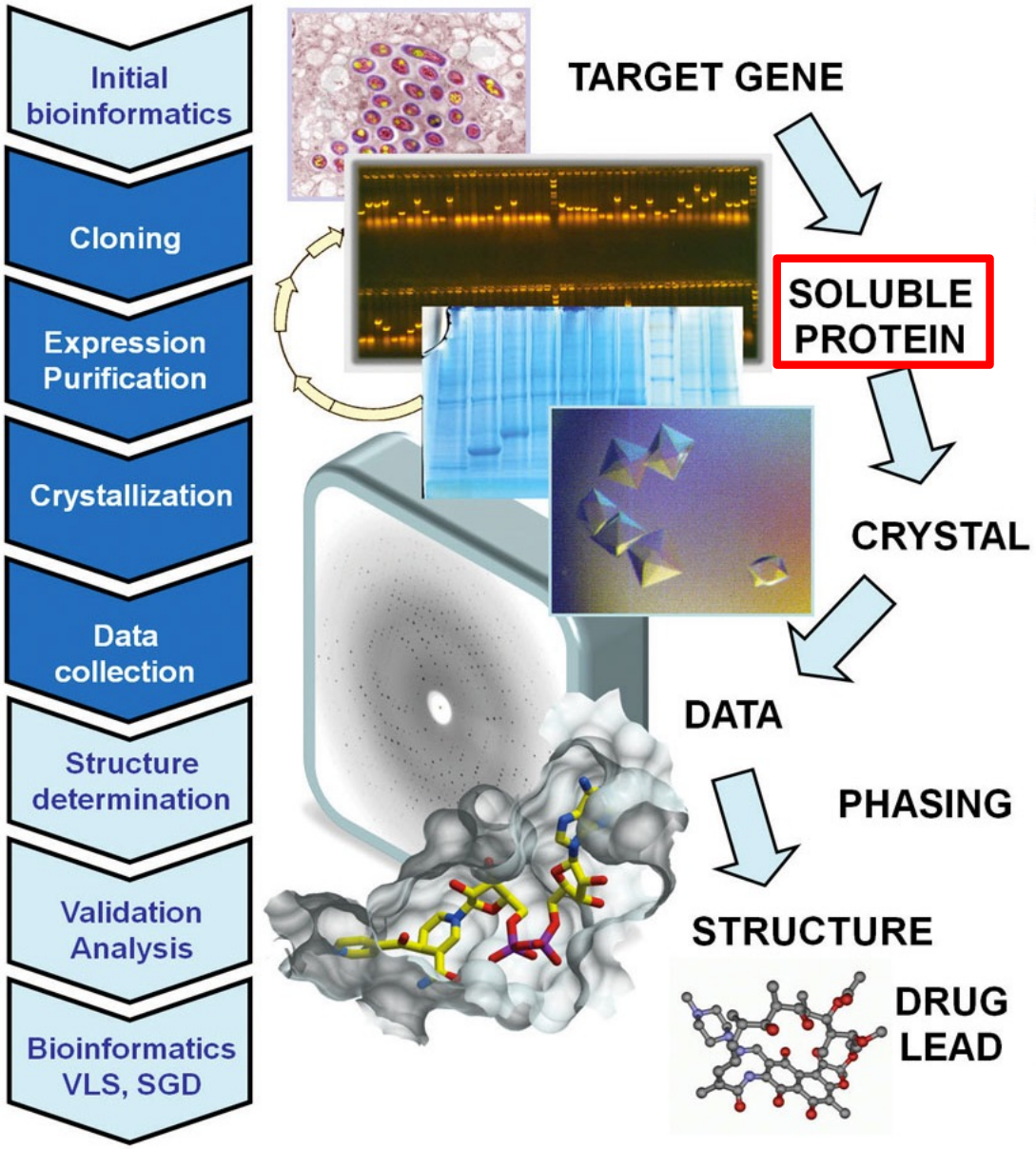
Solve a structure?



Practical aspects:

1. *Protein production*
2. *Crystallization*
3. *Data collection*

The crystallographic pipeline



< bottleneck #1

No protein, no party: a lot of your time as a structural biologist will be spent in cloning, protein expression and purification

Natural vs recombinant sources

Recombinant DNA technology makes it easier to produce large quantities of purified protein.

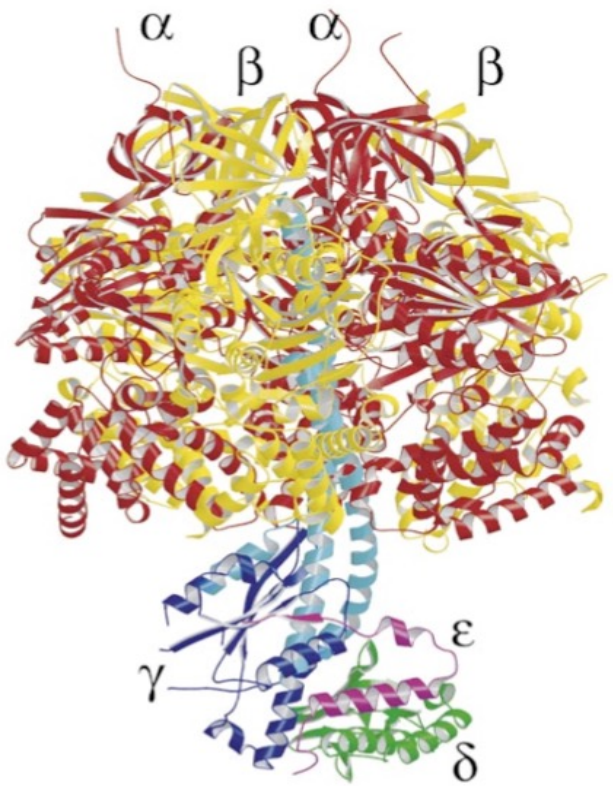
However “natural” sources still useful especially for big complexes, since these cannot be easily re-constituted by recombinant expression.

Examples:

F_oF_1 ATP synthase - purified from bovine heart muscle/yeast cells

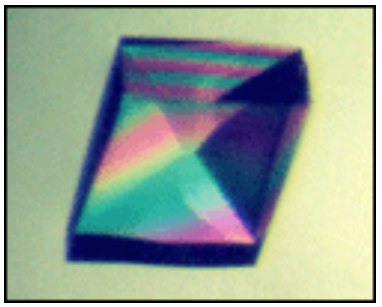
ribosomes - purified from archaea

RNAPII - purified from yeast cells



Easy/difficult to crystallise?

A lot of interesting proteins are difficult to crystallise.



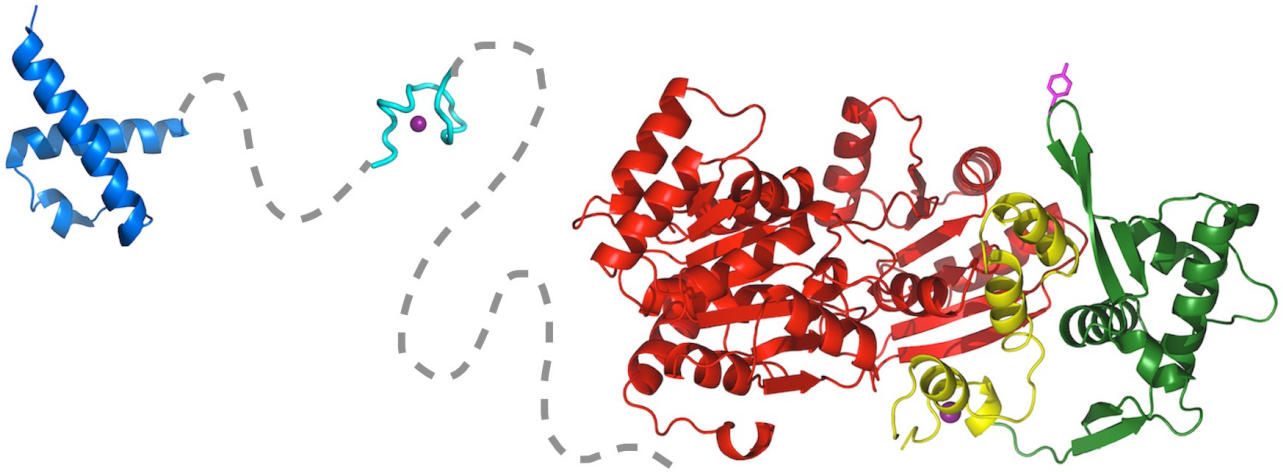
Best cases:

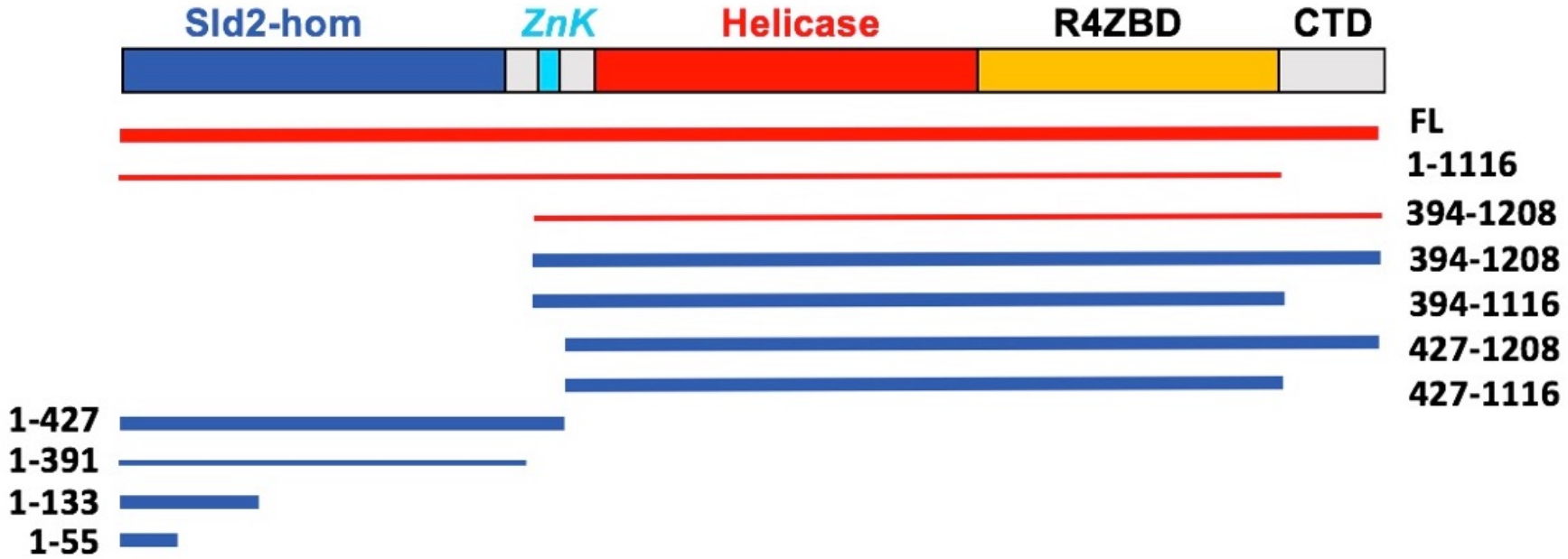
- single proteins
- rigid domain structure
- one dominant conformation

Worst cases:

- multi-domain proteins with flexible unstructured linkers
- proteins with flexible N- and/or C-termini
- proteins that are part of large macromolecular complexes
- presence of posttranslational modifications

"Optimise" proteins for crystallography

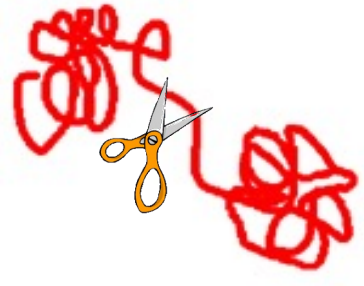




"Optimise" proteins for crystallography

Use bioinformatics (database searches, sequence alignments) to identify "core domains" that can be expressed in a soluble form.

Use limited proteolysis to identify compact domains.



beware:
you can
cut flexible
loops).

Co-express proteins that are part of the same complex.

Add ligands/inhibitors/cofactors/metals
to stabilise one conformation.

Avoid/encourage post-translational modifications (mutagenesis of target residues, mimic modifications by mutagenesis, change expression system/cell lines, so that the process does/does not occur).

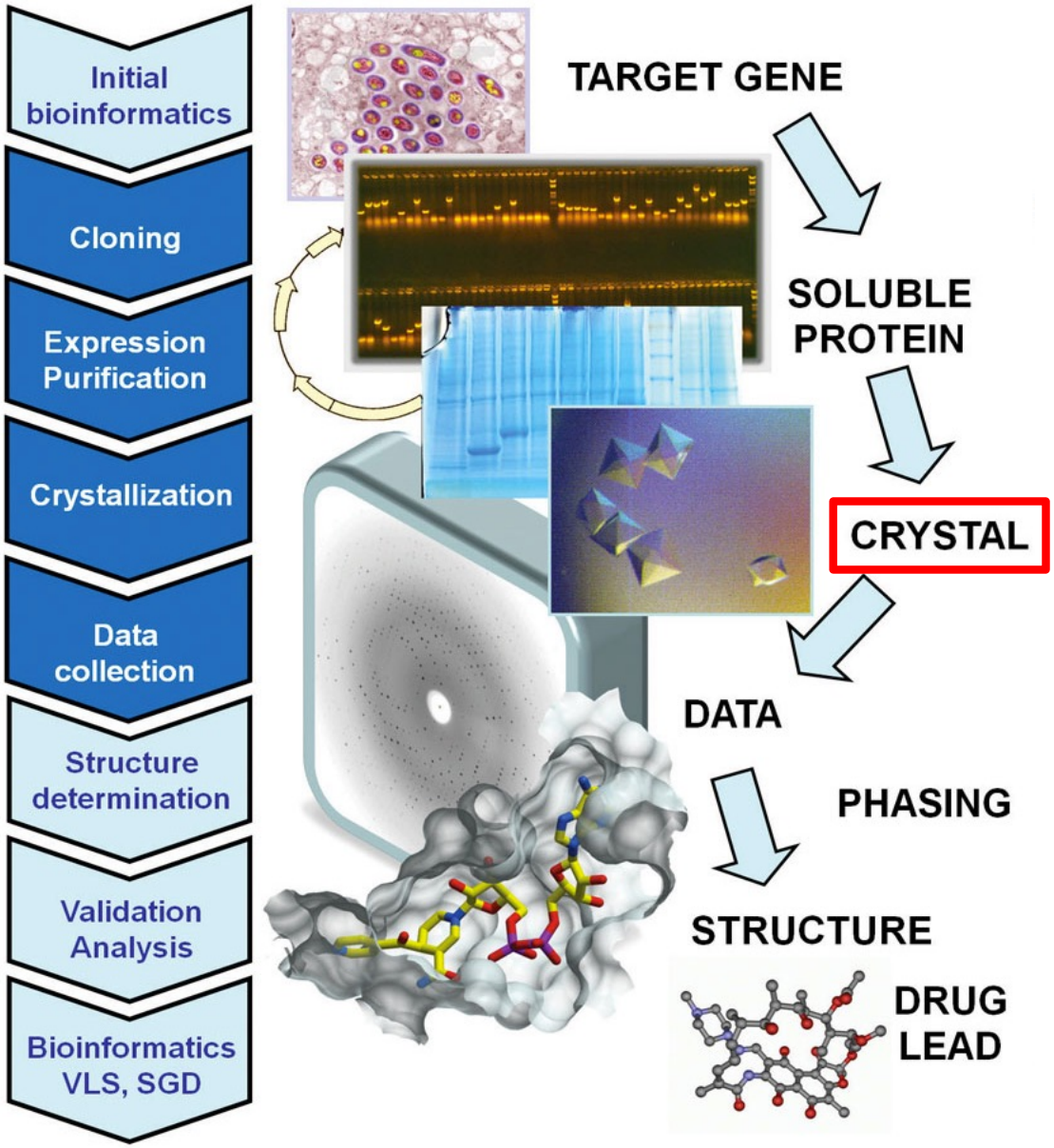
Practical aspects:

1. Protein production

2. Crystallization

3. Data collection

The crystallographic pipeline



< bottleneck #2

No crystal, no party!
crystallization is the real bottleneck!

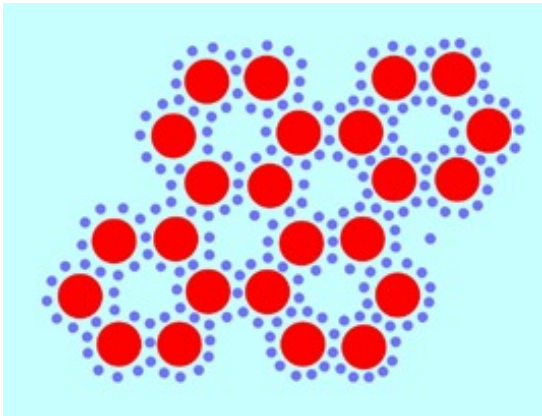
Crystallising proteins

Why proteins are difficult to crystallise:

- most proteins are labile and easily denatured
- large proteins often exist in multiple conformations
- complex behaviour -> polymorphism
- many proteins are difficult to obtain in large amounts
- proteins needs to be highly purified for crystallisation

Why protein crystals are difficult to handle:

- high solvent content (30-80%)
- mechanically fragile
- not well ordered -> resolution limits
- sensitive to radiation damage



Crystallisation of membrane proteins may presents additional problems such as homogeneity of the purified samples, choice of detergents, presence of micelles, tendency to form hydrophobic interactions which are less directional and ordered, etc..

High throughput crystallisation

Automated methods for crystallisation (and crystal visualization) are now routinely used by most labs. These development are driven by the needs of pharmaceutical companies and structural genomics projects.

Robotic crystallization carries out each step of the procedure **quickly, accurately, in smaller volumes**.

Manual crystallisation:
- slow and time consuming
- error prone and not always reproducible
- expensive in terms of amount of purified protein (drops: 1-2 μl)

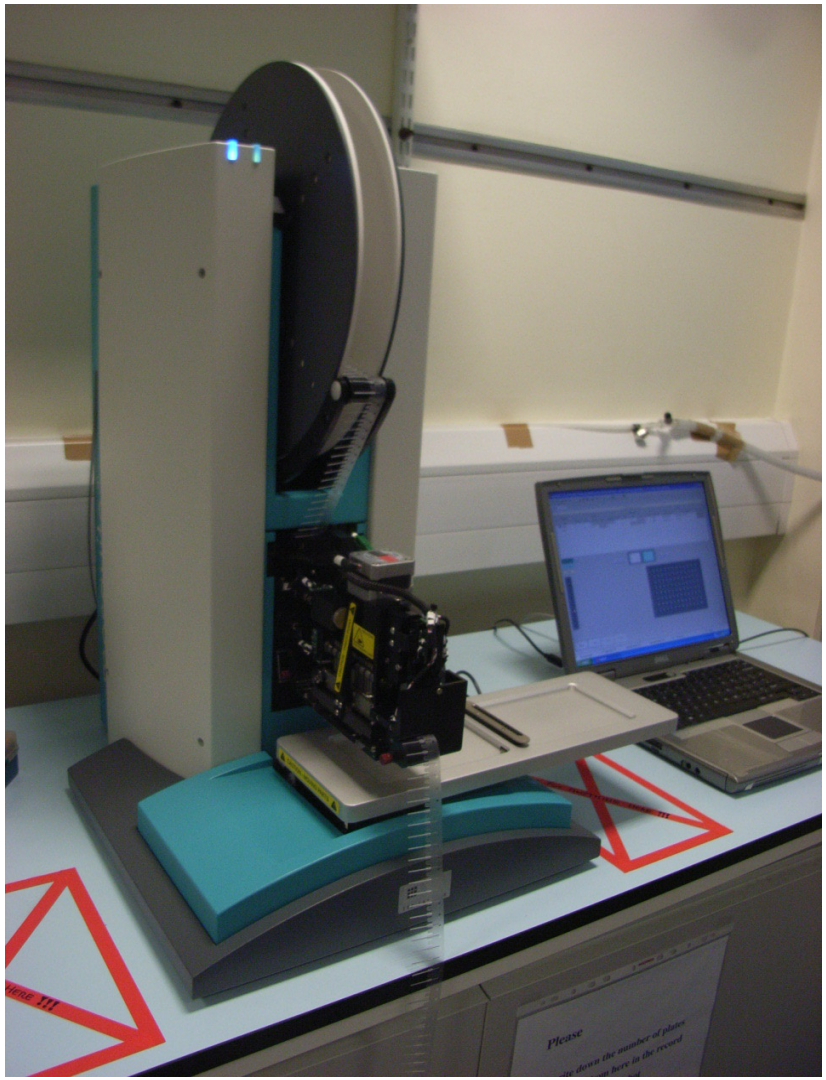
Robotic crystallisation:
- faster and more efficient
- more accurate and reproducible
- smaller sample sizes (down to 50-100 nl drops) cut down on expenditure of purified protein.

Mosquito crystallisation robot

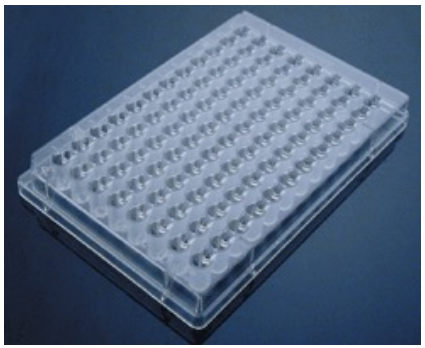
To set up the crystallisation drops containing nanolitre volumes of protein and well solution.

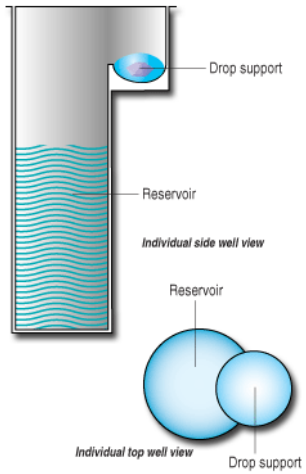
Employs disposable tips and pipettes 100-200 nl drops.

Takes only 2 min to set up 96 drops.



Use of specialised 96-well plates





Practical aspects:

1. Protein production

2. Crystallization

3. Data collection

X-ray sources

Laboratory source

Single wavelength (Copper K_α $\lambda = 1.5418Å$)

Low intensity → long exposure times, radiation damage

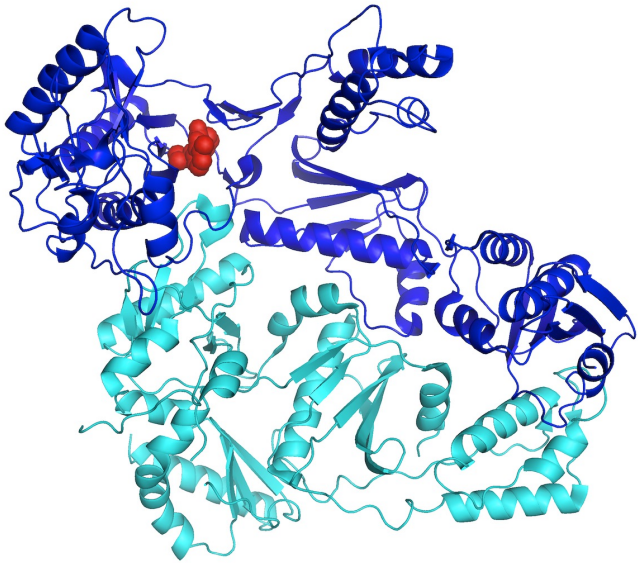
Synchrotron sources

Variable wavelength → allow phasing using MAD/SAD

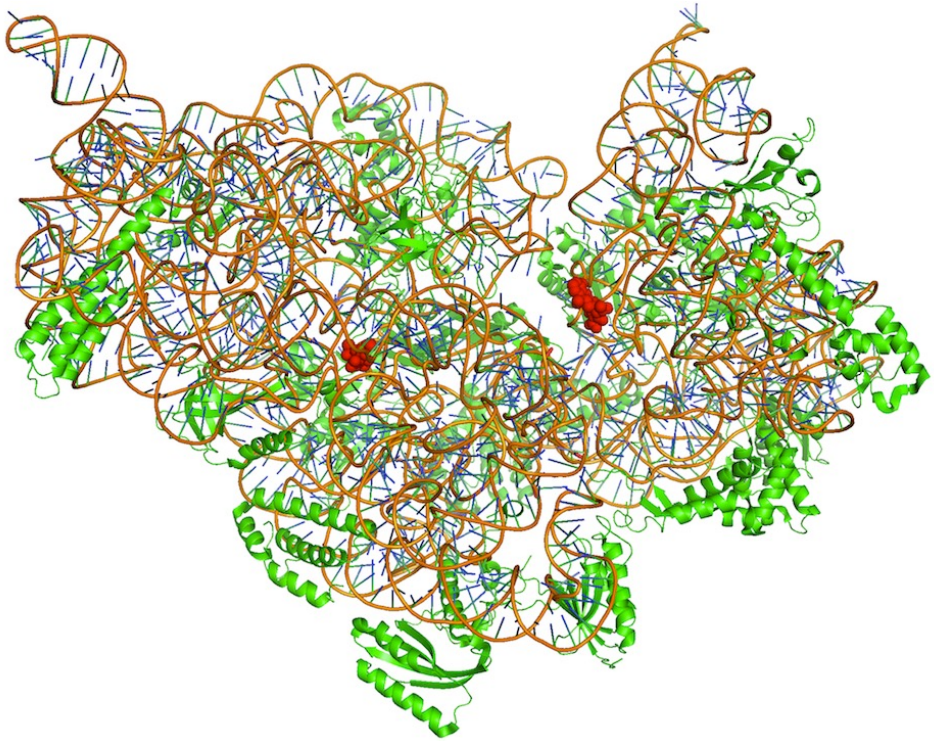
High intensity → short exposure times & less damage per photon

Synchrotron radiation & protein crystals

Most protein crystals are small and diffract poorly. The intense X-rays from a synchrotron has revolutionized structural biology and allowed to see structures of very large and complex proteins, including proteins of medical and pharmacological interest.



*The reverse transcriptase from the **HIV virus bound to Nevirapine**, a component of the anti-AIDS cocktail.*



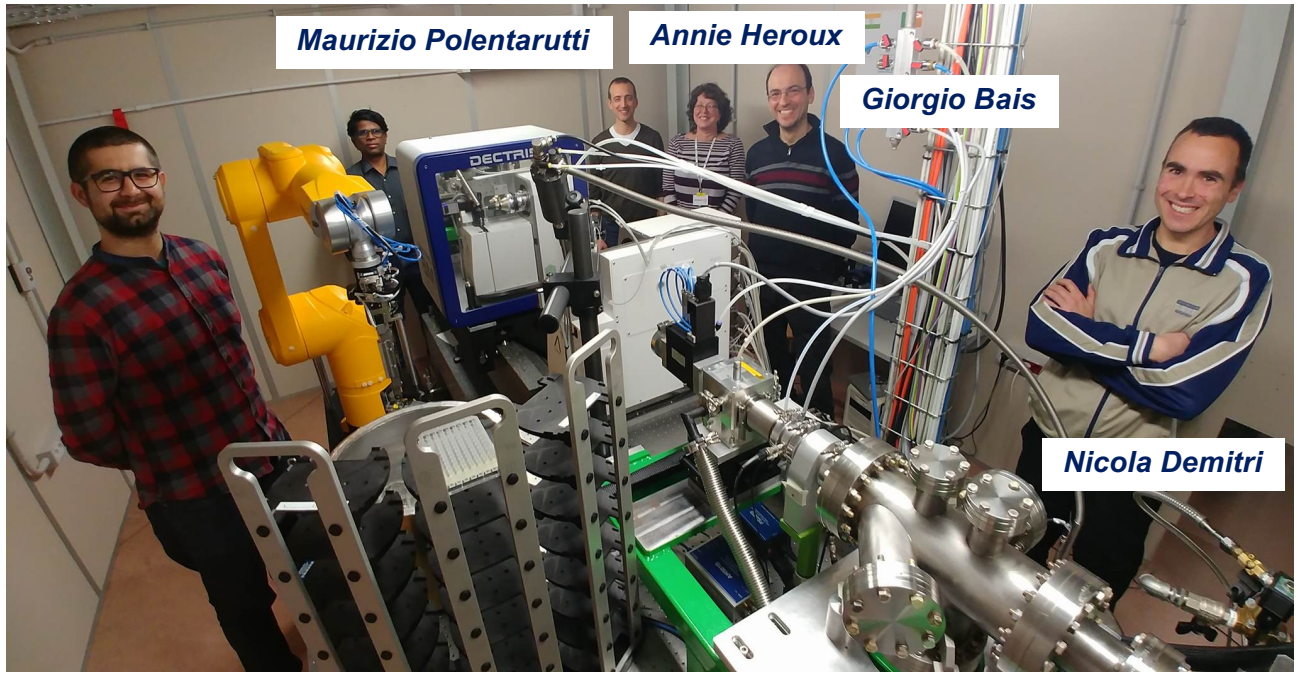
*The small subunit of a bacterial ribosome bound to **tetracicline.***

Synchrotron remote data collection

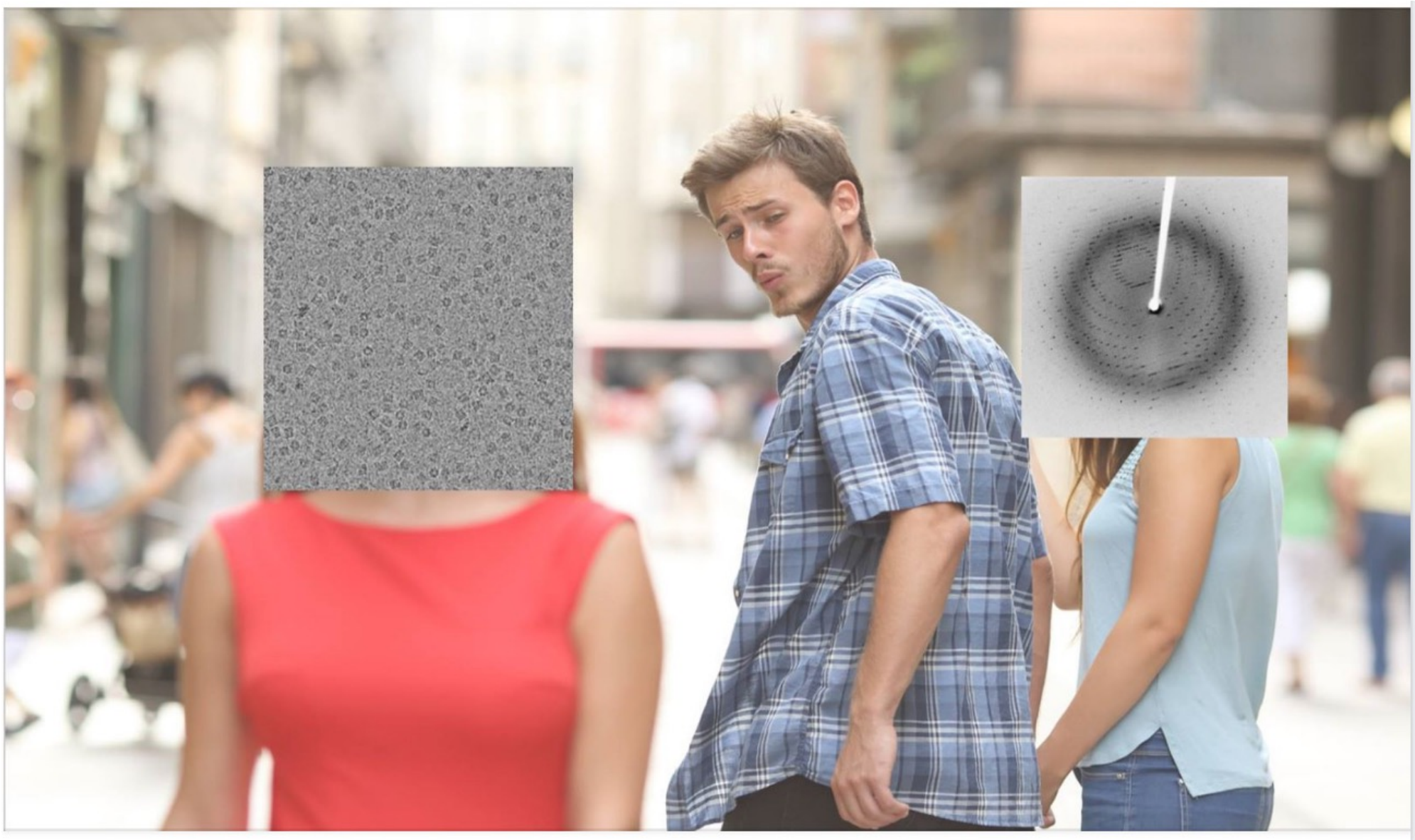
Most MX synchrotron beamlines are highly automated and offer remote data collection to both expert users and novices.

Well designed pipelines allow students, postdocs and PIs to send Dewars with frozen crystals and to collect data online, evaluating the quality of the diffraction, processing the data on-the-fly, and taking decisions as if they were at the synchrotron beamline.

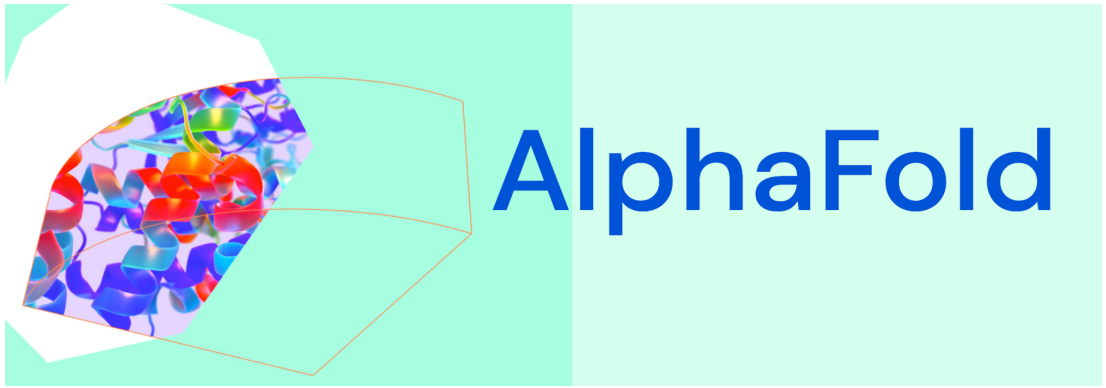
Here is the XRD2 beamline @Elettra



EM vs crystallography?



The role of Artificial Intelligence in Structural Biology?



AI predictions contributions to SB?

- Identify functional domains/domain boundaries for construct design where expression of the full-length protein is problematic.
- Provide models to obtain phases by molecular replacement in MX
- Provide models to fit medium/low resolution CryoEM/CryoET maps
- Provide models for computational biology/molecular dynamics
- Design more stable mutants (i.e to increase the half-life of an antigen for vaccine development, to obtain stable protein/crystals)

AI predictions: current limitations?

- AI works well with **single proteins or domains** – a lot of biology involves homomeric or multimeric complexes (AlphaFoldMultimer?)
- AI models do not include **ligands or co-factors** (often a “hole” where ligand/co-factor-metal should be) and **post-translational modifications**?
- AI does nor provide **proteins/nucleic acid complexes**
- AI networks require **continuous training** on existing protein structures: difficult to predict the structures of proteins with folds that are not well represented in the PDB (**“the dark proteome”**)
- **Proteins are not a static or rigid assemblies**: many exist in multiple conformations and they change shape according to the ligands, substrates, partners they bind. AI gives me ONE conf.
- **Drug design requires better precision** than it can be now achieved with AI