The Abdus Salam
International Centre for Theoretical Physics

2139-16

School on Synchrotron and Free-Electron-Laser Sources and their Multidisciplinary Applications

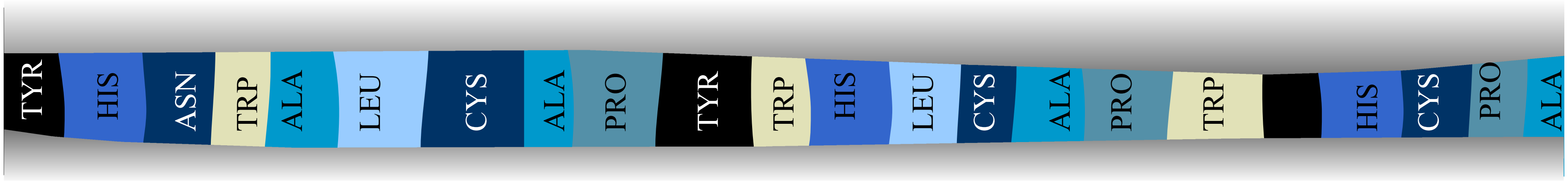
26 April - 7 May, 2010

Protein crystallography

Maurizio Polentarutti
ELETTRA - Trieste

Protein crystallography

[and applications]



Maurizio Polentarutti

ELETTRA, XRD1 beam line

Overview!

Protein

- What a protein is, why we are interested in

crystallography

- Why we want to use [X-ray] crystallography (a little of theory), crystals (and related “sciences”), the experimental set-up(s) and practical aspects of data collection (how the diffraction images should look and some examples from *real* life), data analysis.
- The missing data and the rest of the story: the phase problem and a list of possible solutions (MR, SAD, MAD, DM, IR,...), the model building.
- Additional concepts (if more confusion is needed)

… and applications

- Examples from real life: rational drug design, green chemistry, …

We should talk about math, biology, organic and inorganic chemistry, genes, physics, and more!

References I

- Protein Crystallography, T.L. Blundell & L.N. Johnson (1976), Academic Press
- Crystallographic Methods and Protocols, Methods in Molecular Biology Vol 56, Edited by C. Jones, B. Mulloy, M. R. Sandersson (1996), Humana Press
- Macromolecular Crystallography, Methods in Enzymology, Vols 276 & 277, edited by C.W. Carter and R.M. Sweet (1997) Academic Press
- Macromolecular Crystallography, Methods in Enzymology, Vols 368 & 374, edited by C.W. Carter and R.M. Sweet (2003) Elsevier Press
- Practical Protein Crystallography,2nd Ed., D. E. McRee (1999), Academic Press
- Macromolecular Crystallization – Edited by A. McPherson Methods, Vol.34, n° 3, (2004) Elsevier Press
- Crystallography Made Crystal Clear, 2nd Ed., G. Rhodes (1993) Academic Press

References II

- Direct Phasing in Crystallography, C. Giacovazzo (1998) IUCr, Oxford Science Publications
- The Principles of Protein X-ray Crystallography, 2nd Ed., J. Drenth (2002), Springer Verlag
- Introduction to Macromolecular Crystallography, A. McPherson (2002), Wiley & Sons
- Outline of Crystallography for Biologists, D. Blow (2002), Oxford University Press
- Protein Crystalization Techniques, Ed. T. Bergfors (1999) I.U.L
- Preparation & Analysis of Protein Crystals, A. McPherson (1989) Krieger Publishing
- Methods & Results in Crystallization of Membrane Proteins, Ed. S. Iwata (2002) IUL
- Membrane Protein Purification & Crystallization, A Practical Guide, G. Von Jagow, H. Schagger, C. Hunte (2002) Academic Press
- International Tables for Crystallography Volume F: Crystallography of biological macromolecules, Ed. M. G. Rossmann (2002) Springer

References III

Principles of Protein X-ray Crystallography, J. Drenth, Springer-Verlag

http://www.iucr.org/education/ IUCR site

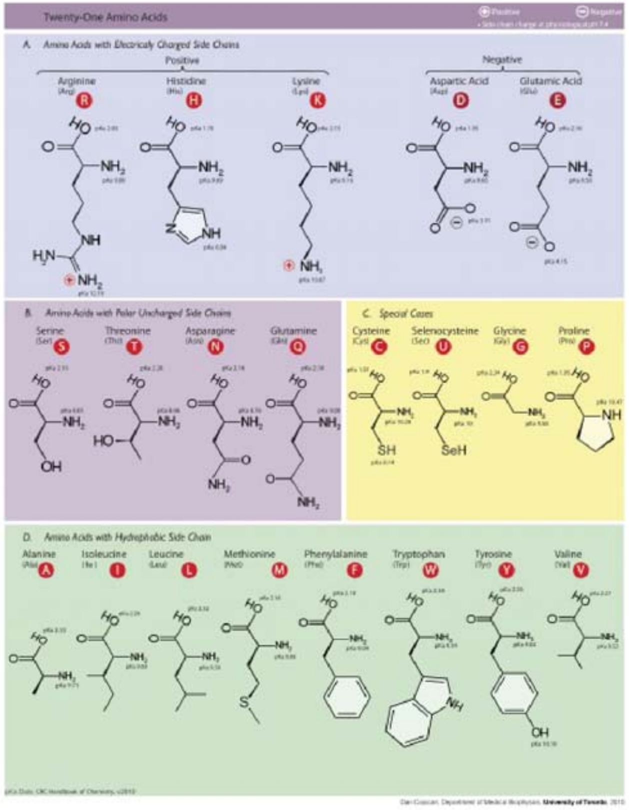
Fundamentals of Crystallography, C. Giacovazzo et al., IUCR books, Oxford University Press

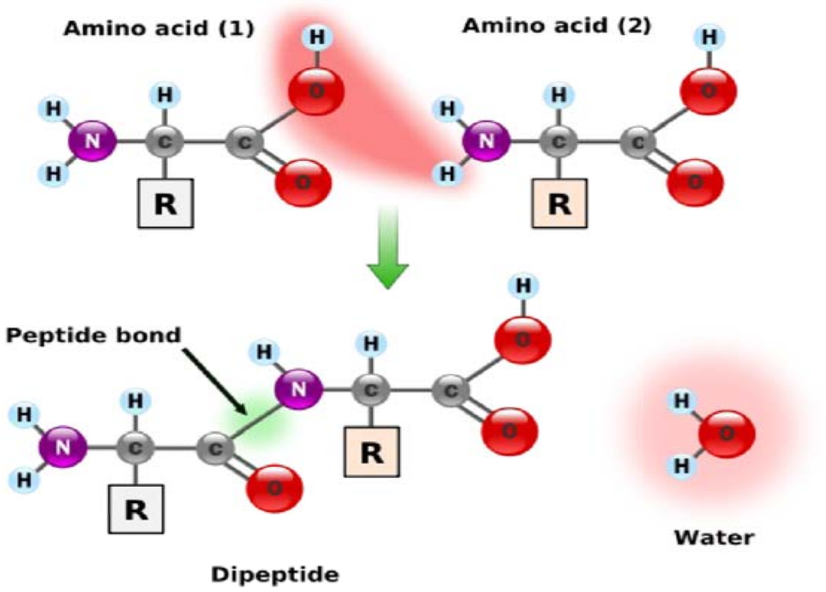
INTERNATIONAL TABLES for CRYSTALLOGRAPHY (vol.F)

The protein data bank (PDB) http://www.pdb.org/pdb/home/home.do

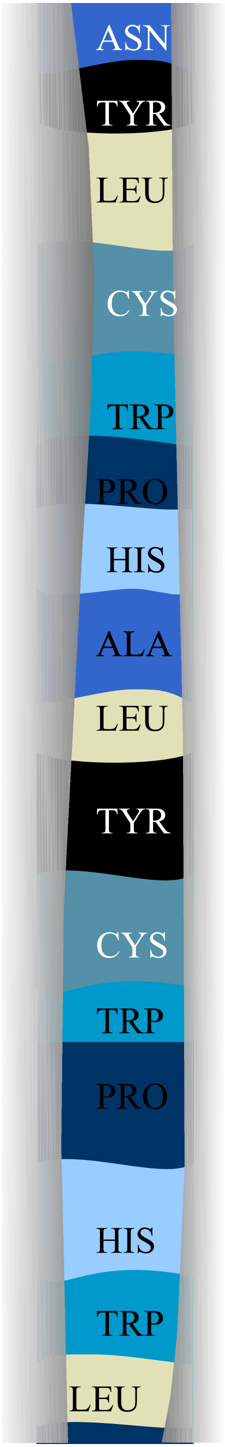
Protein?

Proteins are organic compounds made of amino acids arranged in a linear chain and folded into a globular form. The **amino acids** in a polymer are joined together by the peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. The sequence of amino acids in a protein is defined by the sequence of a gene, which is encoded in the genetic code. The genetic code specifies **20** standard amino acids.





Note: most part of the atoms are: C, H, N and O, plus someone special



Why are we interested in proteins?

Proteins are essential parts of organisms and participate in virtually **every process within cells**. Many proteins are **enzymes** that catalyze biochemical reactions and are vital to metabolism. Proteins also have structural or mechanical functions, such as actin and myosin in muscle and the proteins in the cytoskeleton, which form a system of scaffolding that maintains cell shape. Other proteins are important in cell signaling, immune responses, cell adhesion, cell cycle,...

We want to:

- 1.interact with them because they have a central role in every metabolic/catabolic processes in our body [medicine]
- 2. use them (engineering) in industry, exploiting their high efficiency and selectivity [sensors]
- 3. know how they effectively do the job (working principles) and how they interact each other

In any case we want to know in great detail (atomic) their three-dimensional structure.

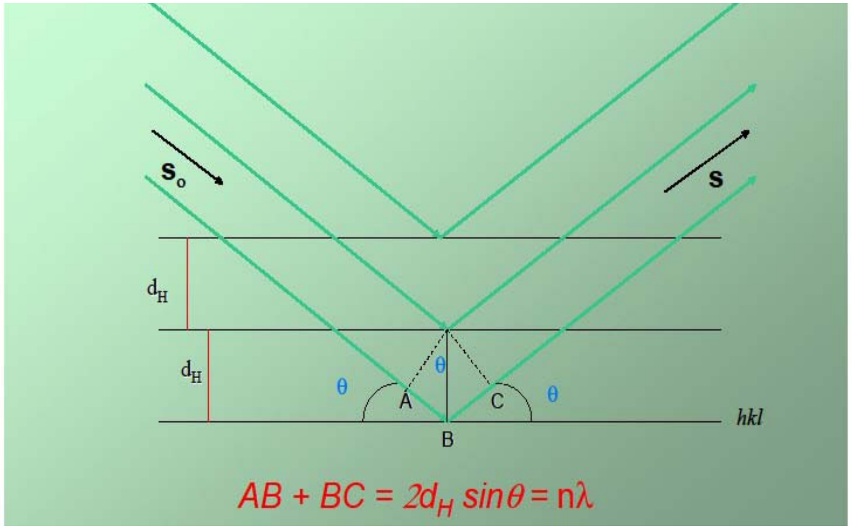
X-ray crystallography (I)

Is an experimental technique which exploits the **diffraction** of X-rays by a crystal to reconstruct the **electron density (ρ)** in the unit cell of the crystal:

- We deal with light elements: let's consider the interaction ruled by elastic scattering

- The incoming radiation is reflected by planes of the crystal (sample)

- Interference from the scattered waves determine at which angles the scattered beam is non zero (Bragg's low)

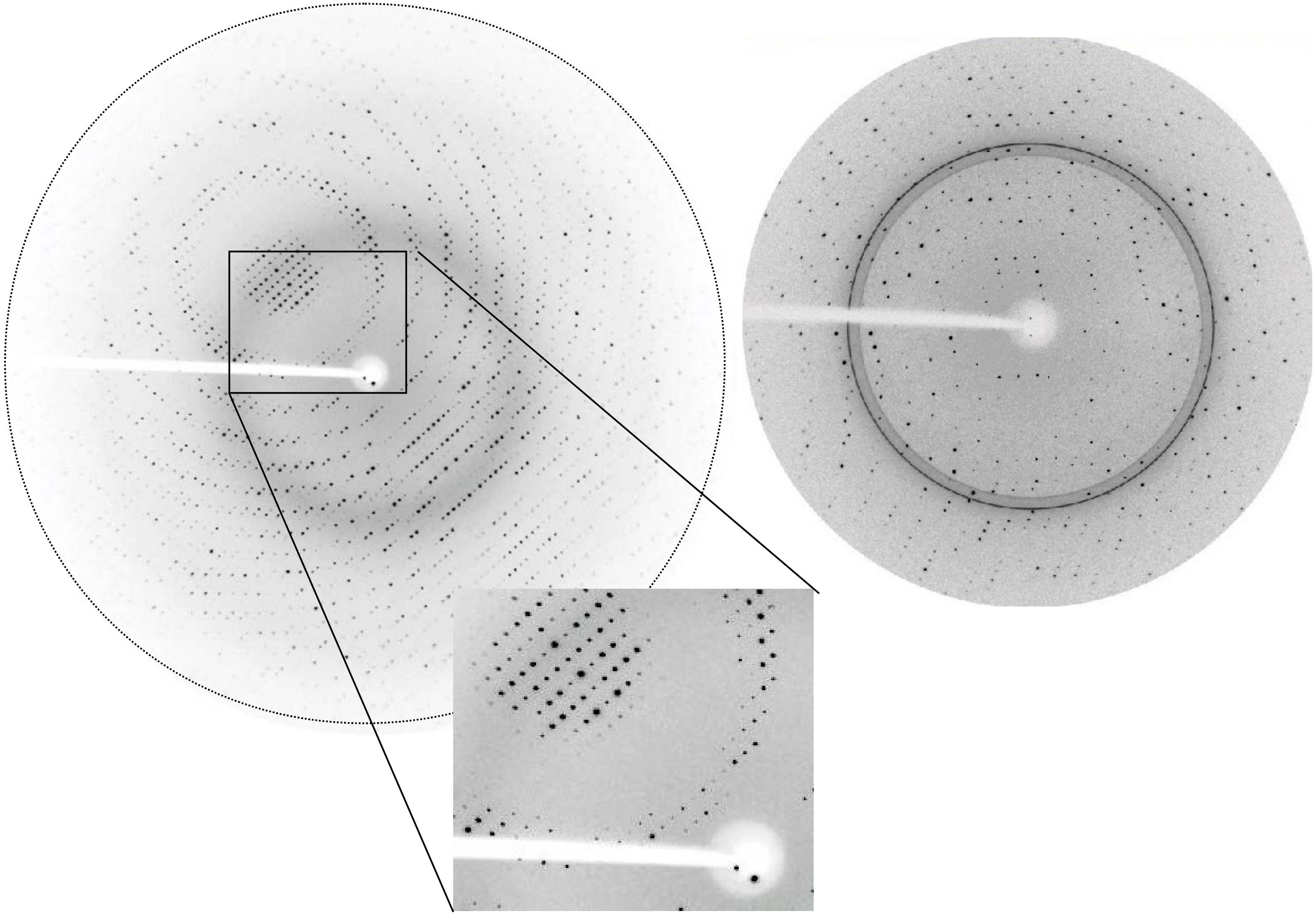


$AB + BC = 2d_H \sin\theta = n\lambda$

$$2d_H\sin(\theta)=n\lambda$$

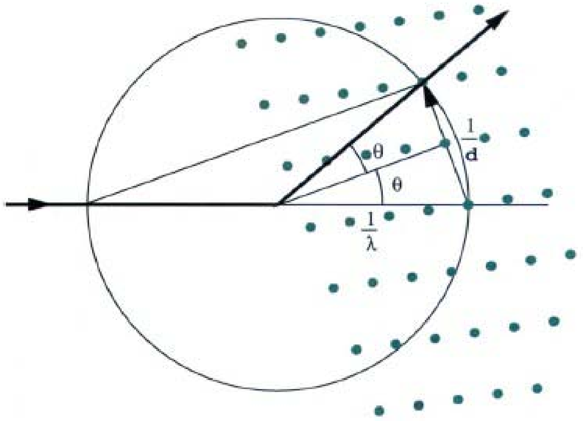
Examples from real life → →

X-ray crystallography (I)

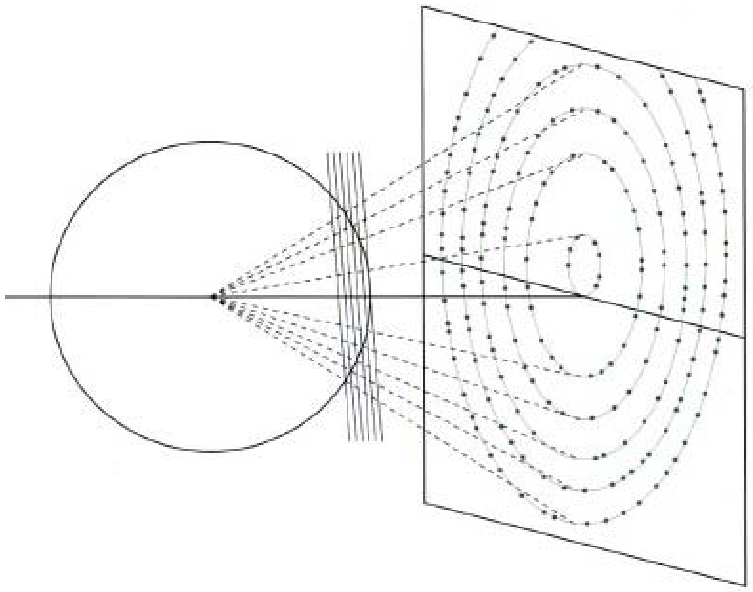


X-ray crystallography (I)

Understanding the diffraction pattern



The Ewald construction. When the reciprocal-lattice point crosses the surface of the sphere, the trigonometric condition $1/d = (2/\lambda)\sin(\theta)$ is fulfilled. This is the three-dimensional illustration of Bragg's law $\lambda = 2d\sin\theta$

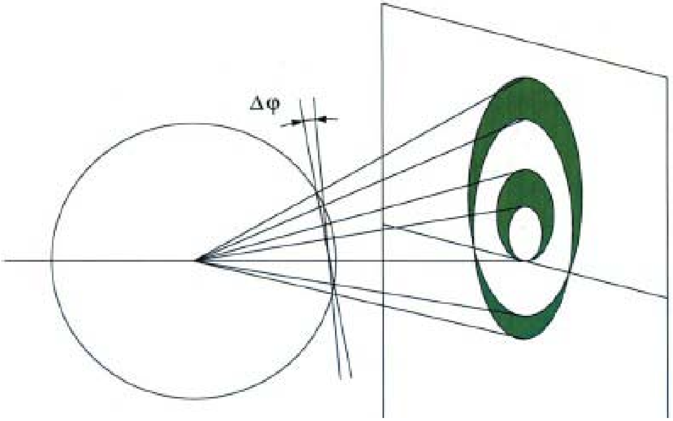


A still exposure with a stationary crystal contains only a small number of reflections arranged in a set of narrow ellipses.

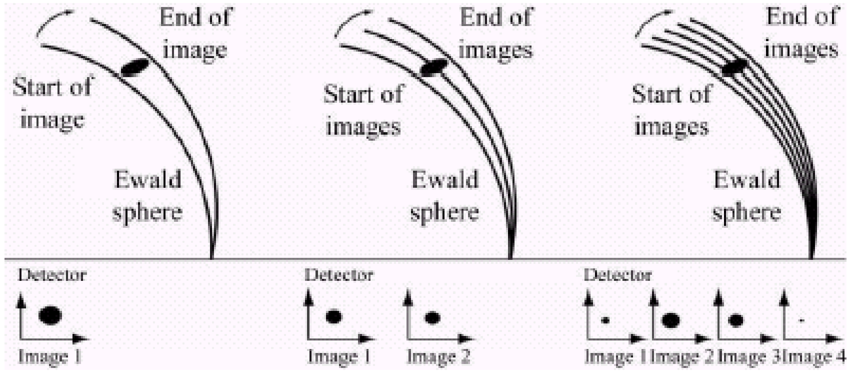
We can rotate the crystal in order to measure the intensity of the radiation coming from every (hkl) plane.

X-ray crystallography (I)

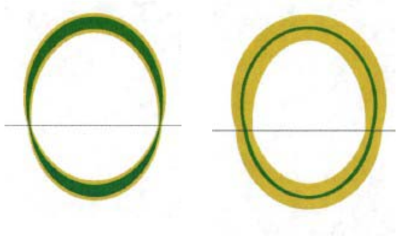
Understanding the diffraction pattern



When the crystal is rotated, reflections from the same plane in the reciprocal lattice form a **lune**, limited by two ellipses corresponding to the start and end positions.



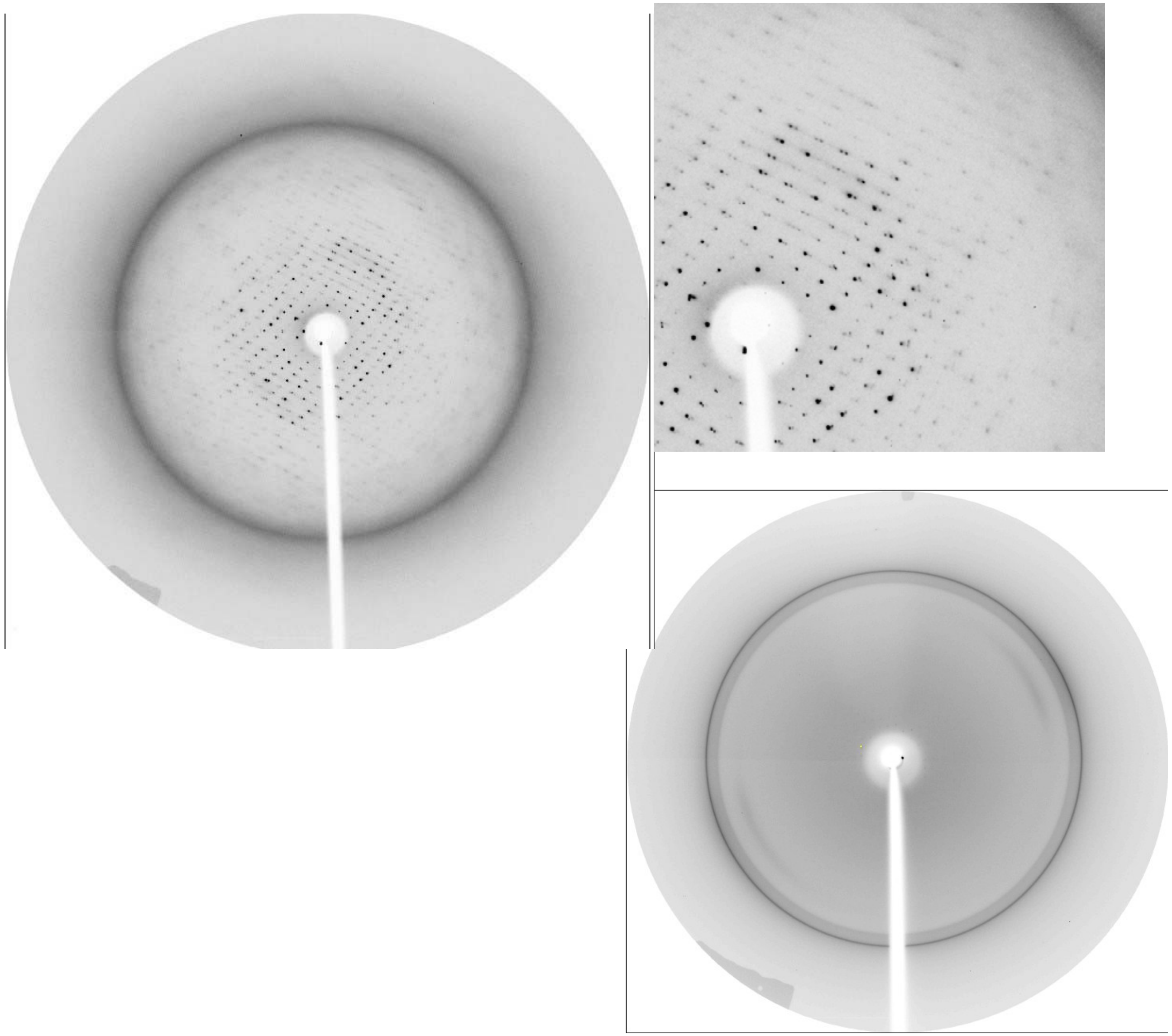
Real crystals are composed of small mosaic blocks slightly misoriented with respect to one another, which adds some divergence to the total rocking curve, that is to the amount of rotation during which an individual reflection diffracts.



Low and High mosaicity lunes, with partially recorded and fully recorded reflections.

X-ray crystallography (I)

Understanding the diffraction pattern



X-ray crystallography (II)

The integrated intensity for each reflection (h,k,l)

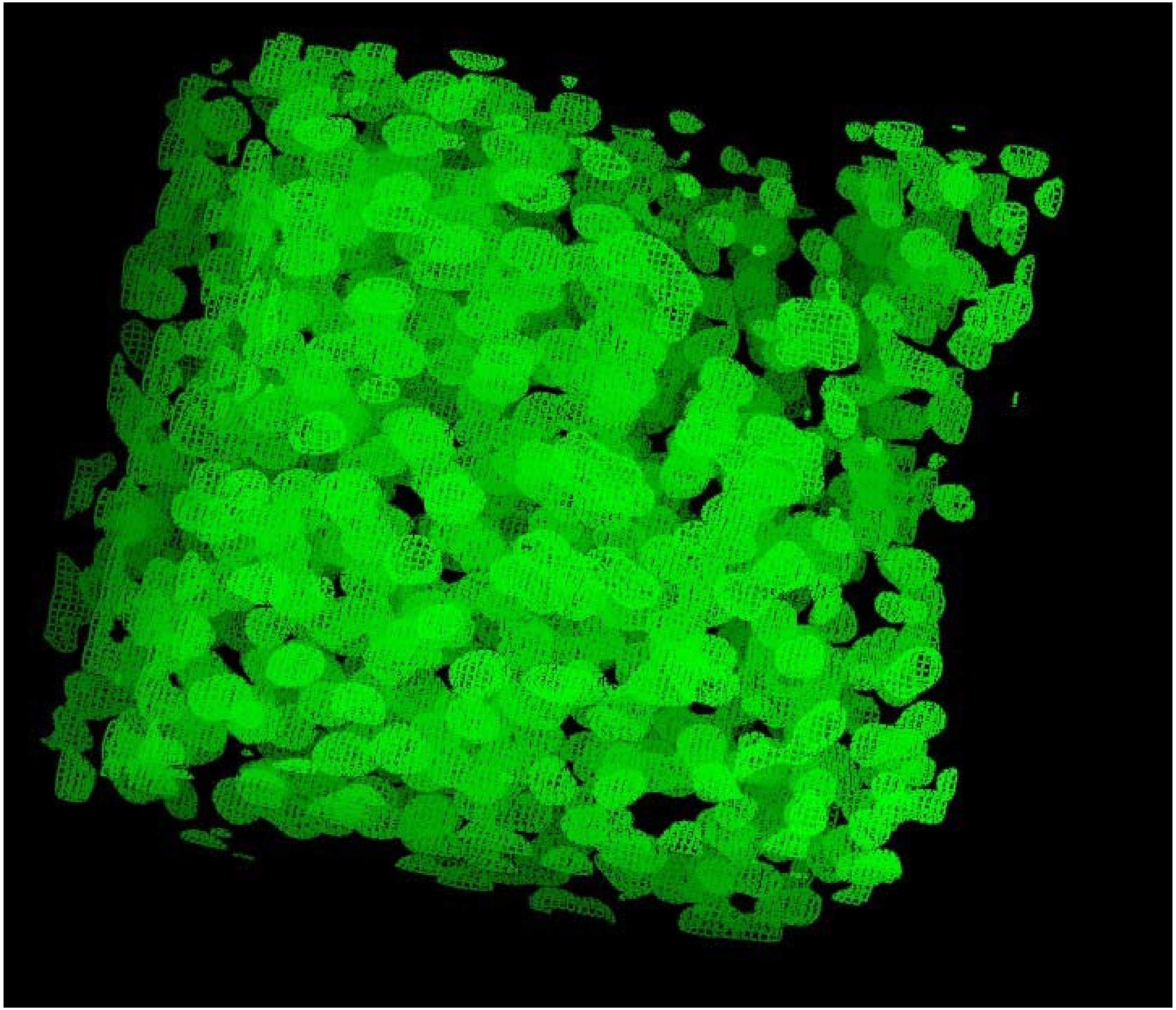
$$I(\mathrm{int}, h\,k\,l) = \frac{\lambda^3}{\omega \cdot V_{\mathrm{cell}}^2} \times \left(\frac{e^2}{mc^2}\right)^2 \times V_{\mathrm{cr}} \times I_0 \times L \times P \times A \times \left|F(hkl)\right|^2$$ where,

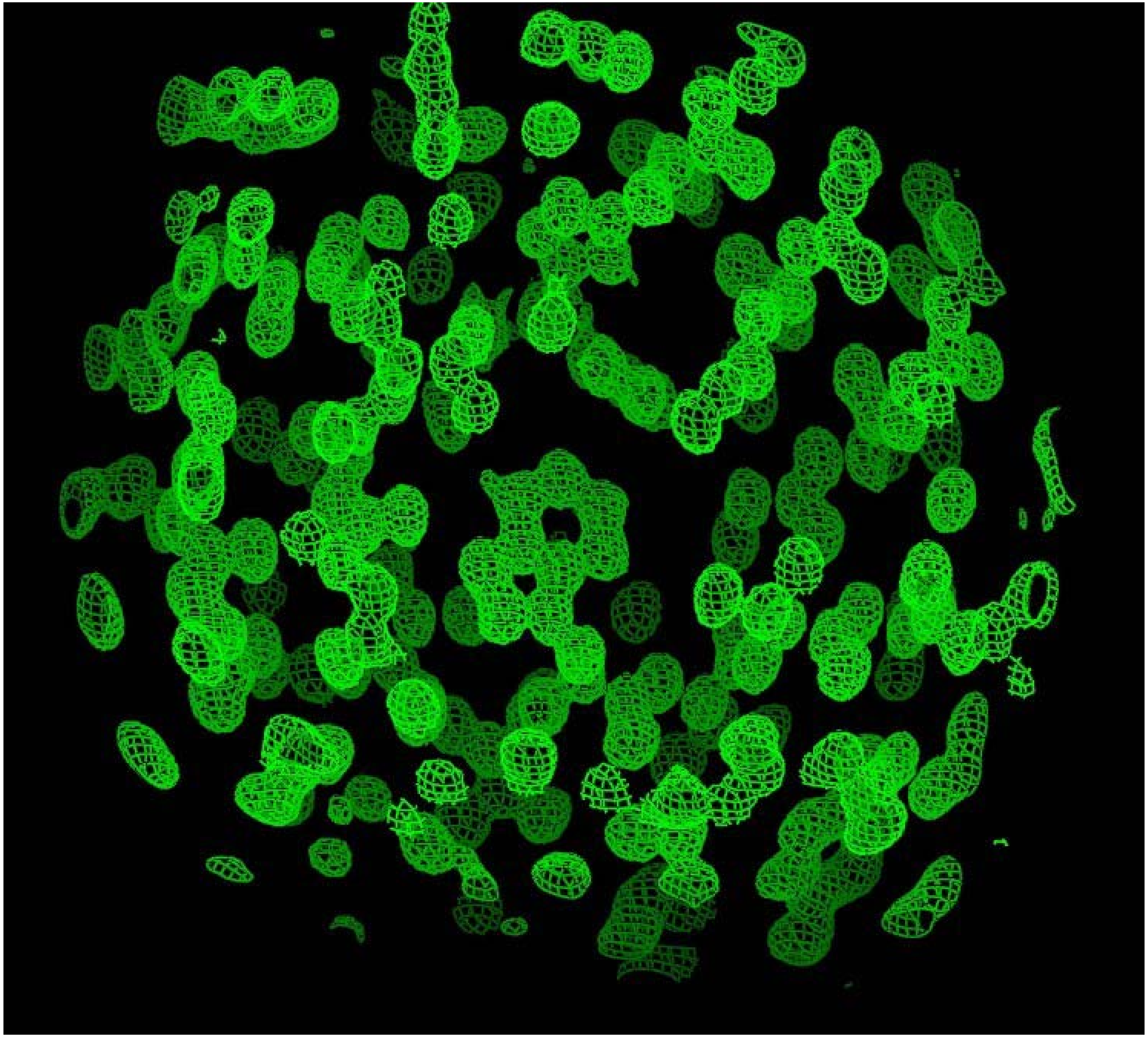
$$\vec{F}(h,k,l) = \sum_{j=1}^{N\,atoms} f_j e^{2\pi i(hx_j + ky_j + lz_j)}$$

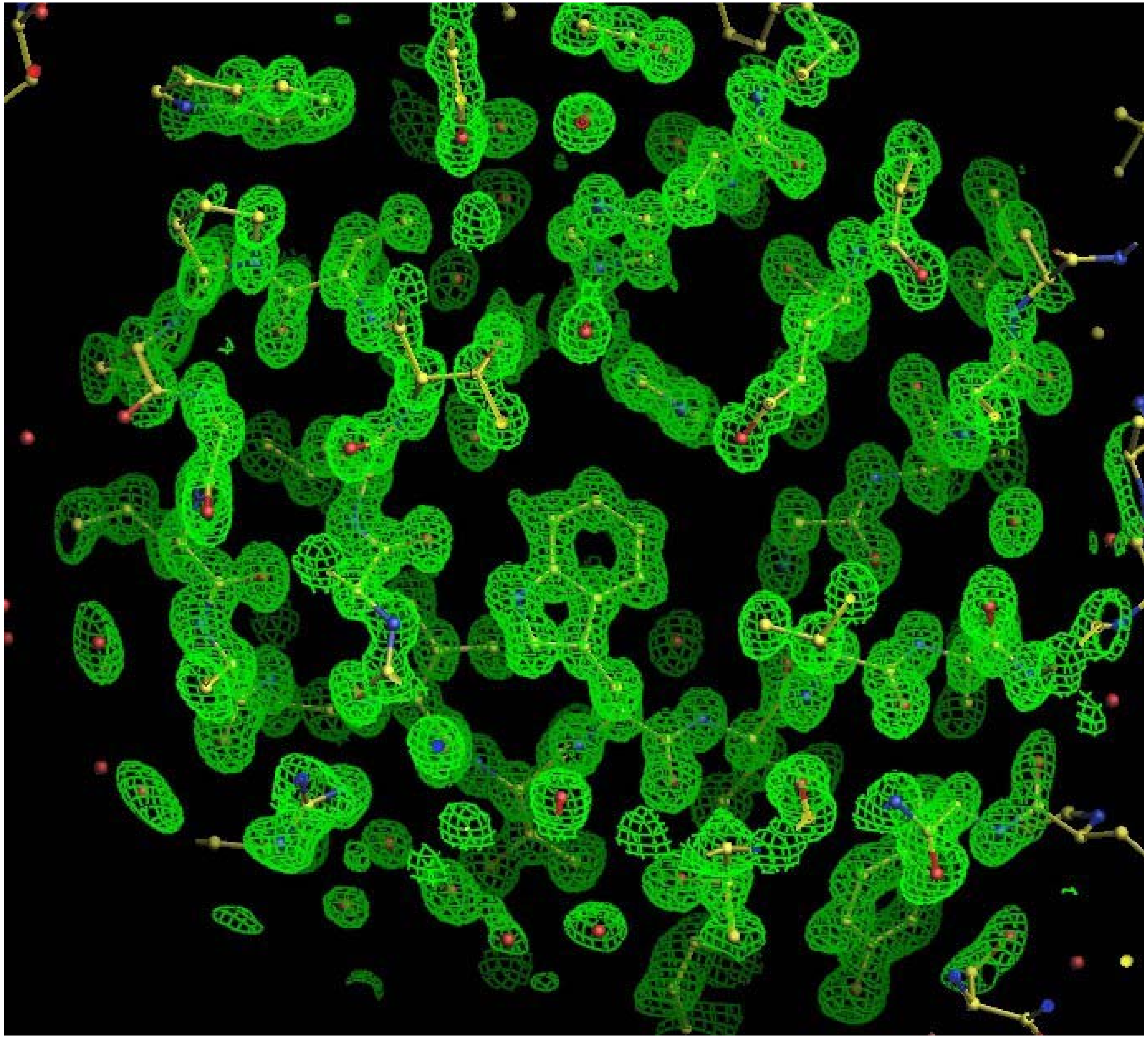
The *atomic structure factors* (f_j) take into consideration the scattering properties of each atom, and depends on the number of electrons (Z). In the structure factor the position and scattering power of each atom is considered (and the use of synchrotrons justified).

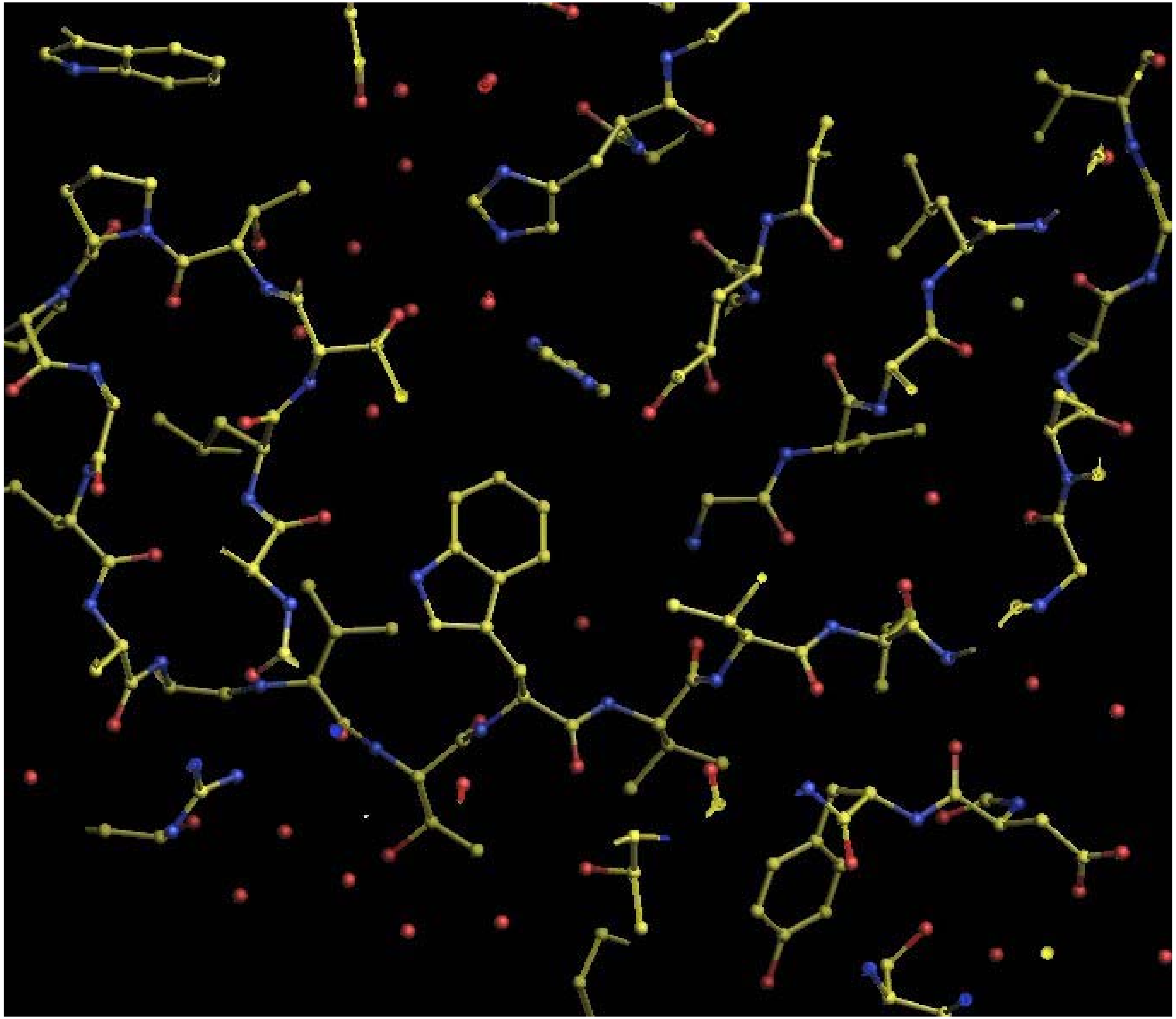
$$\rho\,(x,y,z) = \frac{1}{V_c}\sum_h \sum_k \sum_l \left|F(h,k,l)\right| \cdot e^{-2\pi i(hx+ky+lz)+i\alpha\,(h,k,l)}$$

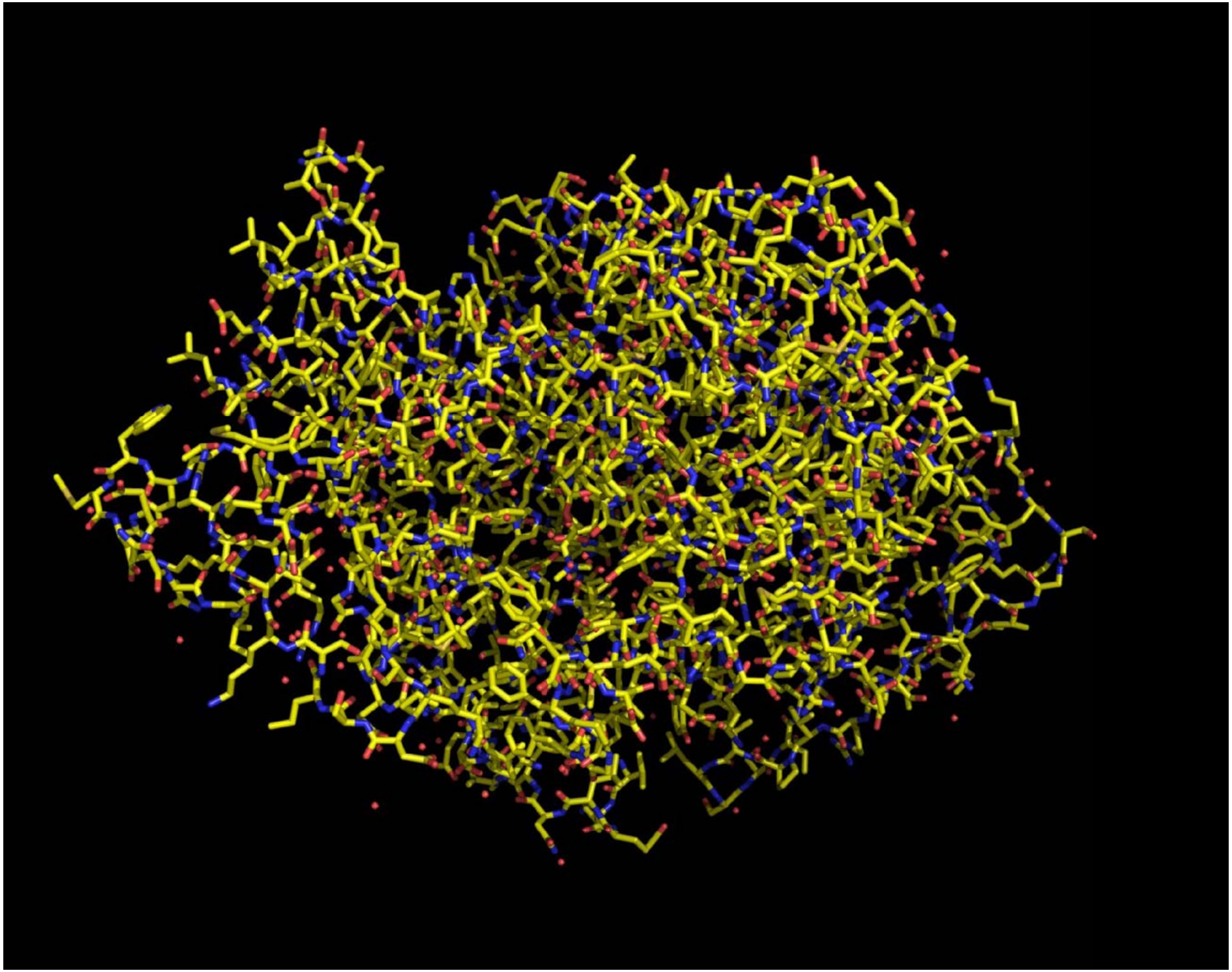
Starting from the electron density it is possible to produce a **model** of the molecule [a small one as well as a protein (thousands of atoms)]

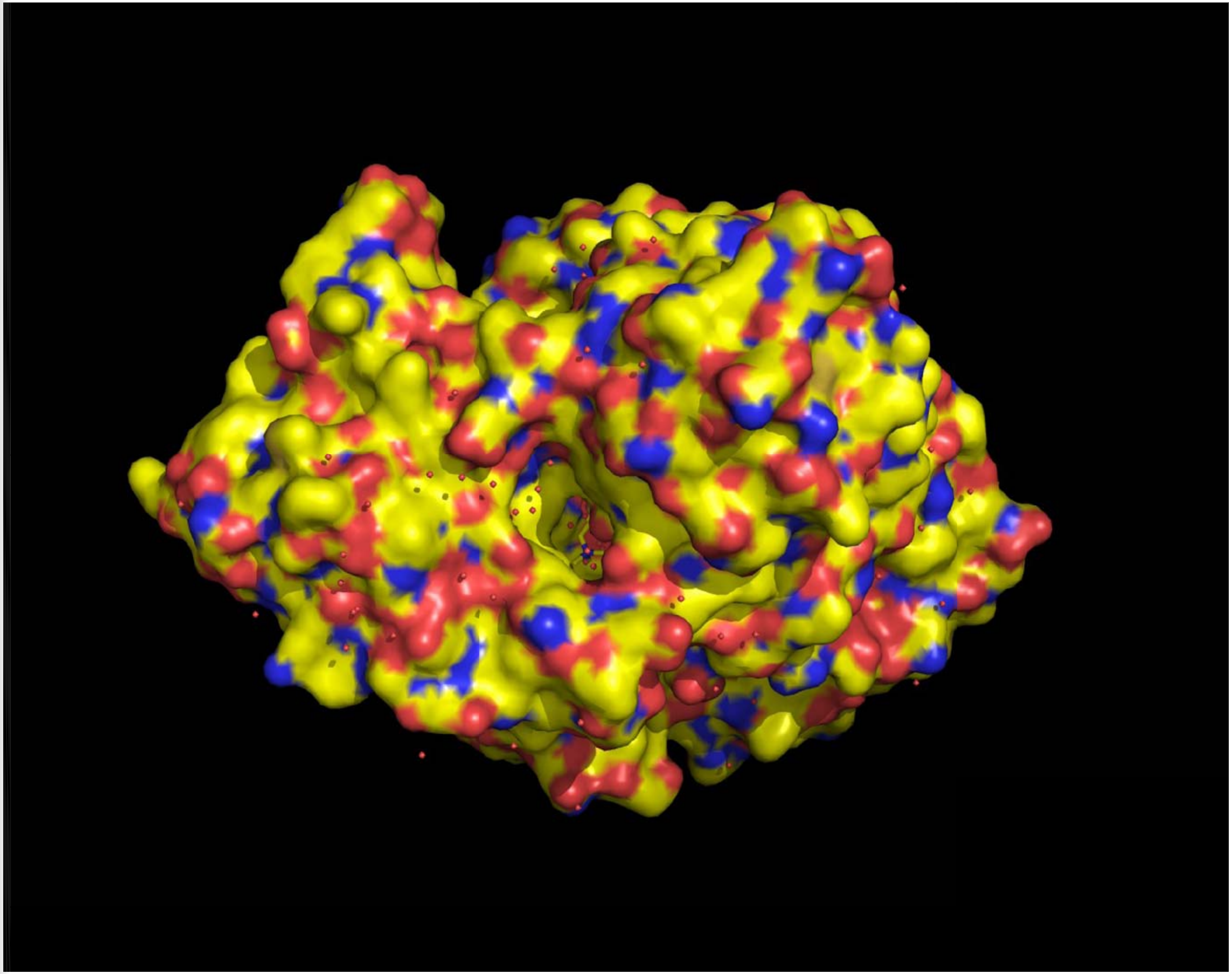


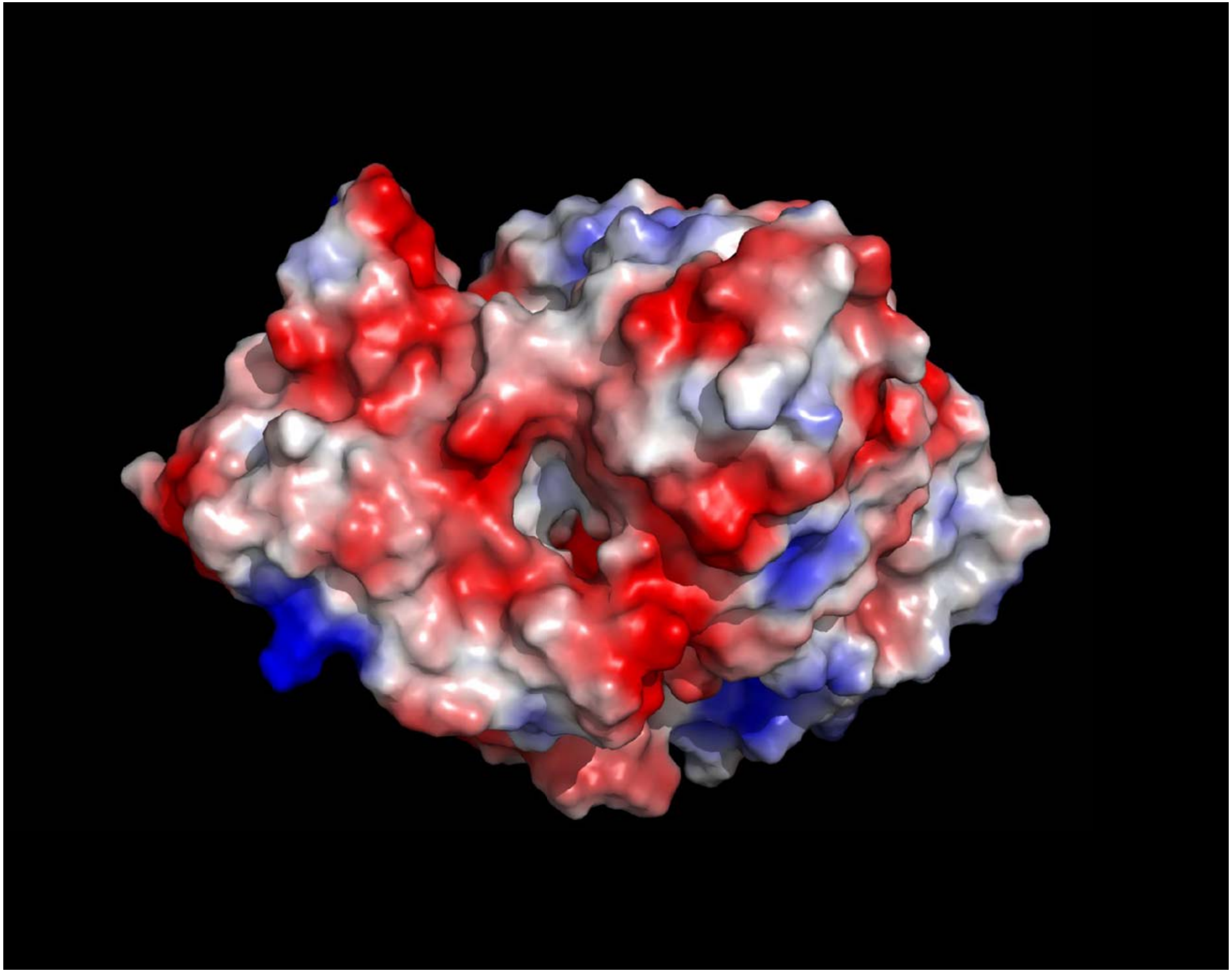












Protein crystallography: (practical aspects)

- How to crystallize your favourite protein?
- an X-ray source (~ 1Å)
- An experimental setup (minimal)

- Data analysis (I): *integration* and *scaling*
- Data analysis (II): *phasing*
- The electron density interpretation: *tracing, docking, improving*

- Into the *model*
- How to represent the model of the protein

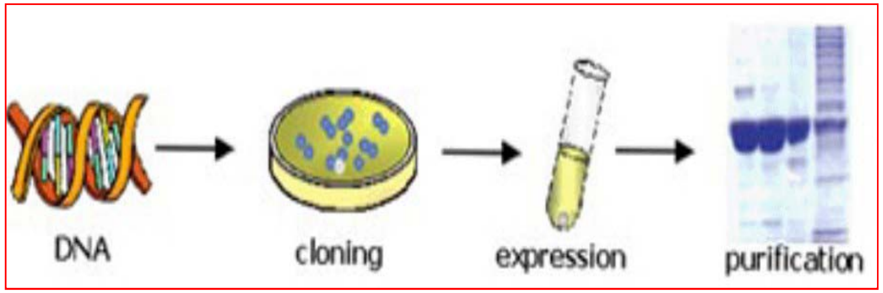
Protein crystallization

•It's mainly a **trial-and-error** procedure in which the protein is slowly precipitated from its solution

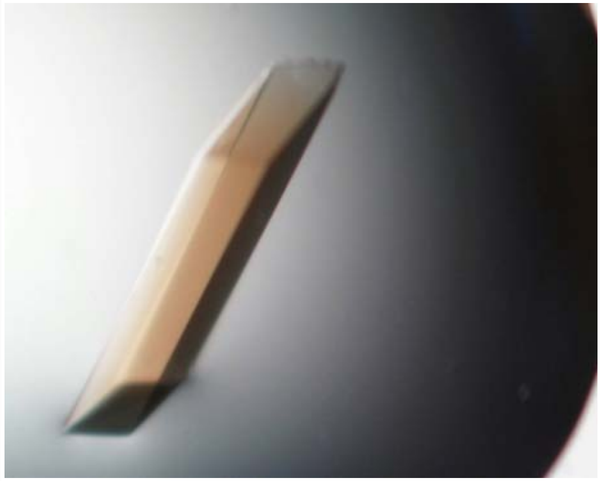
•Crystal growth is solution is a multi parameter process involving 3 basic steps: **nucleation** (possibly having only 100 molecules), **growth** and **cessation of growth** (when the protein in solid phase is in equilibrium with the solution).

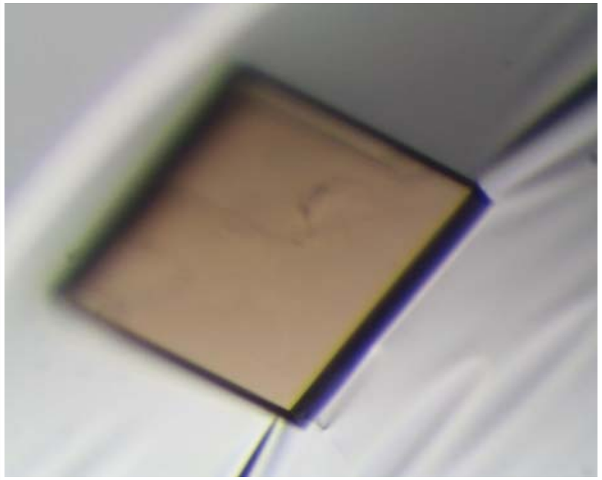
•It's extremely difficult to predict good conditions for nucleation or growth of well-ordered crystals. In practice favourable conditions are identify by screening (hundreds- thousands of solution conditions are generally tried).

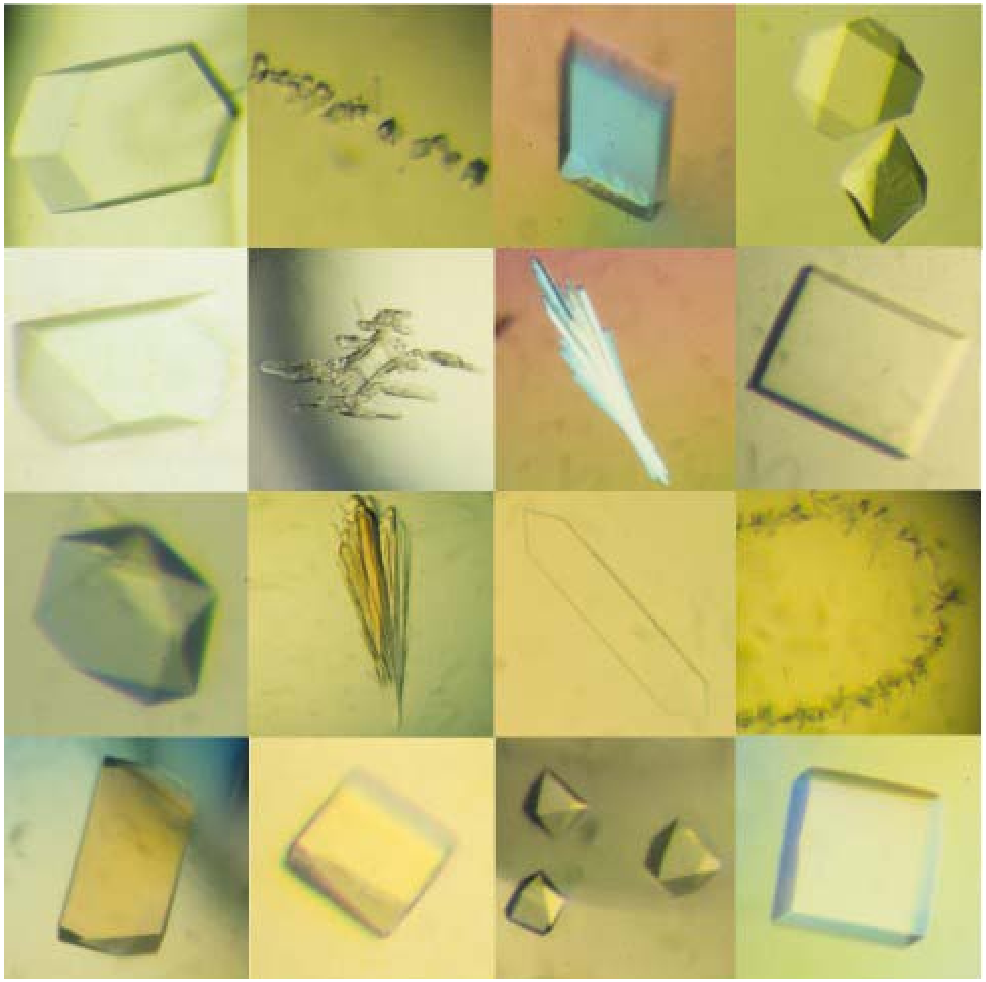
•Large amounts (milligrams!) of the target molecule are required (due to high concentration of the molecule(s) to be crystallized). Techniques of **recombinant cDNA** are used [molecular biology] to produce such amount of protein, which has to be **purified** [biochemistry/biophysics].



Protein crystals - gallery







X-ray sources

Wavelength: 1 Å (10^{-10} m, 12.4 keV)

X-Ray Generators: sealed tube or rotating anode

Targets: Cu (K_α=1.5418Å), Mo (K_α=0.7107Å),
Cr (K_α=2.291Å)

Synchrotron radiation (4.0-25 keV, 0.6-3.1Å)

An experimental setup (minimal)

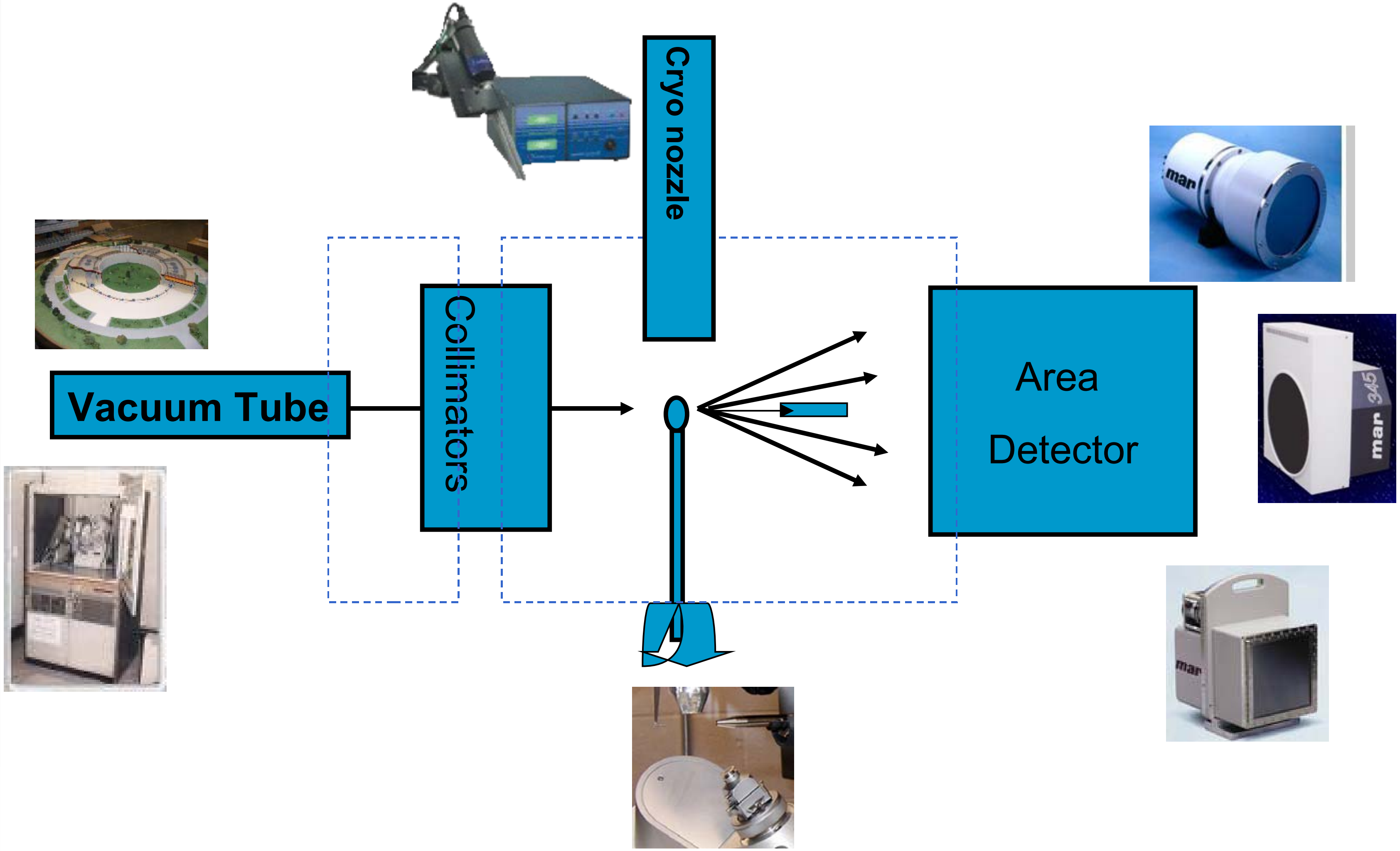
Main components – common to all PX beamlines and home labs (in order of appearance):

- *Slits (beam shapers)*
- *Shutter (related to the time exposure)*
- *Sample (protein single crystal)*
- *Sample cooler system*
- *Sample manipulator system (horizontal spindle axis)*
- *Fluorescence detection system (beam lines only)*
- *(Primary)-Beam stopper*
- *Detector*

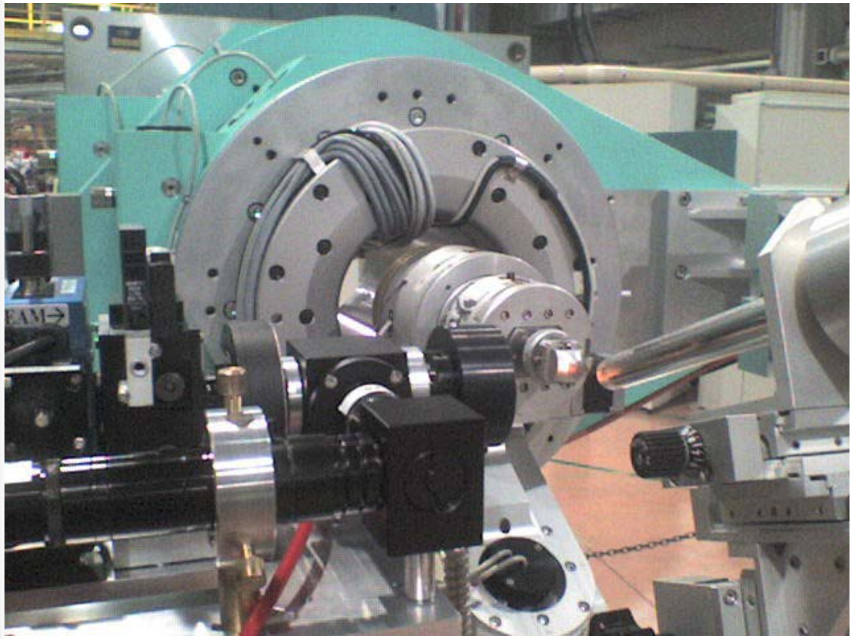
- Experimental key parameters

  Sample-to-detector distance, wavelength, detector surface, beam stopper position, sample macroscopic and unit-cell dimensions, sample orientation, detector angular position, sample rotation per image, exposure time.

An experimental setup (minimal)



LS SERIES


EAM→

DECTRIS

Data analysis (I): *integration* and *scaling*

- Image *integration (Denzo, Mosflm [ccp4], XDS)*

  <u>Results</u>: for each reflection (hkl), get a value for the integrated intensity and relative error . Get the unit cell, the space group, the crystal orientation, effective resolution limit, refine the crystal to detector distance and detector angular positions.

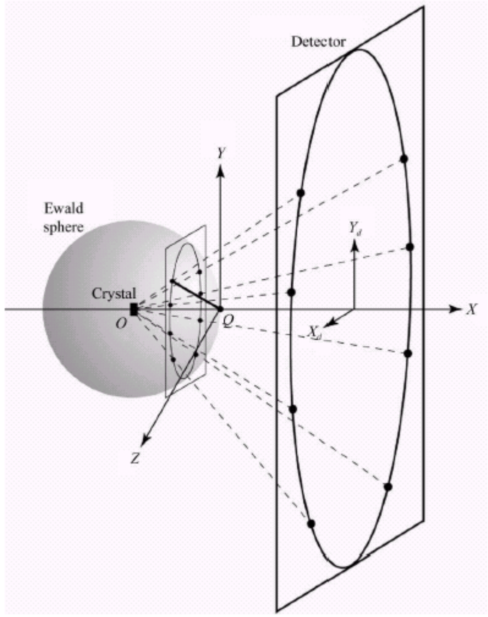
- Data *scaling (Scalepack, Scala [ccp4])*

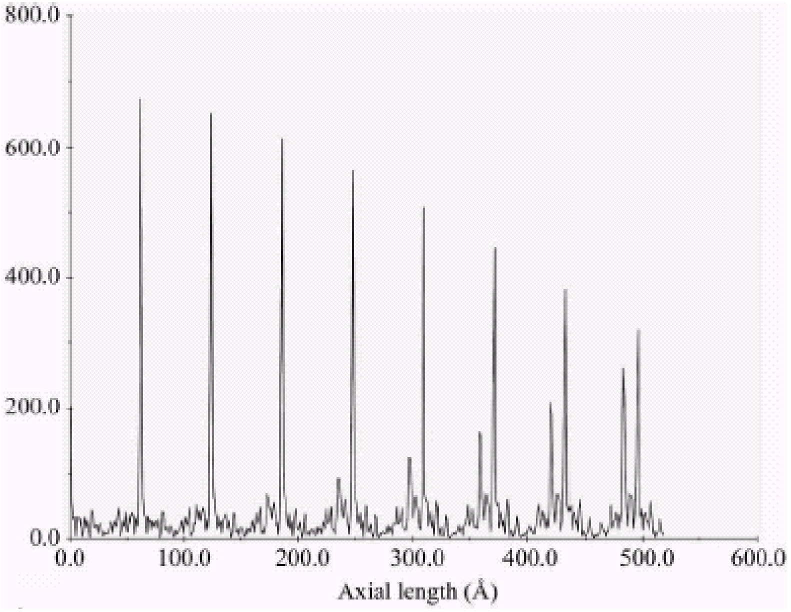
  <u>Results</u>: take into consideration the decay of the beam intensity, sample and air absorption, radiation damage, detector problems (spatial distortion, non-uniformity of response, time stability, bad pixels), changes in diffracting volume, estimation of data quality.

Data analysis (I): *integration* and *scaling*

How the *integration* works:

- If the members of a set of reciprocal-lattice planes perpendicular to a chosen direction **t** are well separated, then the projections of the reciprocal-lattice vectors onto **t** will have an easily recognizable periodic distribution.
- We consider about 7300 separate roughly equally spaced directions.
- The unit of the periodicity is obtained via a Fourier transform.
- The resultant unit cell is then reduced and analyzed in terms of the 44 lattice types (Burzlaff et al., 1992).





Data analysis (I): *integration* and *scaling*

The *scaling* step:

Incident beam related factors

- Synchrotron
  - smooth decay of beam intensity
  - any discontinuities (e.g. beam injection) should be noted and included in scaling model
  - illuminated volume
  - shutter synchronization/goniometer rotation speed

Crystal related factors

- Sample absorption
  - diffracted beam absorption (shape dependent)
  - important for weak anomalous signal
- Radiation damage
  - can be significant on high brilliance sources
  - difficult to correct for
  - modeled as change in relative B-factor
  - extrapolation to zero dose

Data analysis (I): *integration* and *scaling*

Detector related factors

- **calibration errors**
  - spatial distortion
  - non-uniformity of response
  - time stability
  - bad pixels

Miscellaneous factors

- unavoidable
  - zingers
- avoidable
  - beam stop shadow
  - cryo-stream shadow
  - should be dealt with at integration stage

Determination of scale factors

Scales are determined by comparison of symmetry-related reflections, i.e. by adjusting scale factors to get the best internal consistency of intensities. Note that we do not know the true intensities and an internally-consistent dataset is not necessarily correct. Systematic errors will remain

$$\text{Minimize } \Delta\Phi = \Sigma_{hl}\, w_{hl}\, (I_{hl} - 1/k_{hl}\langle I_h \rangle)^2$$

I_{hl} l’th intensity observation of reflection h

k_{hl} scale factor for I_{hl} $\langle I_h \rangle$ current estimate of I_h

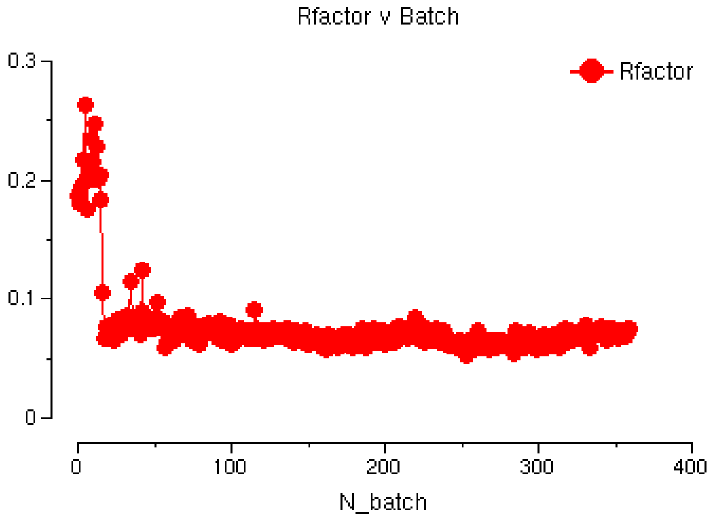
Data analysis (I): *integration* and *scaling*

Data quality indicators

- Rmerge (Rsym) = $\Sigma \mid I_h - \langle I_h \rangle \mid / \Sigma \mid \langle I_h \rangle \mid$

  Values: $R \leq 0.10$ (10%)→Very good; $0.10 \leq R < 0.20$ →Suspect, $R \geq 0.2$ (20%)→ Bad !

Analysis of Rmerge against batch number gives a very clear indication of problems local to some regions of the data. Perhaps something has gone wrong with the integration step, or there are some bad images



- Here the beginning of the dataset is wrong due to problems in integration (e.g. poor orientation matrix in MOSFLM at start of job.)

Data analysis (II): *phasing*

A step back: where did we get the phases?

$$\rho\,(x, y, z) = \frac{1}{V_c} \sum_h \sum_k \sum_l \left|F(h,k,l)\right| \cdot e^{-2\pi\, i(hx + ky + lz) + i\alpha\,(h,k,l)}$$

$$\vec{F}(h,k,l) = \sum_{j=1}^{N\,atoms} f_j e^{2\pi i(hx_j + ky_j + lz_j)}$$

No direct information about the phases comes from the experimental data! We have different methods to get the phases:

1. **Direct** methods
2. Experimental methods (MAD, SAD, **MIR**, …)
3. Previous knowledge (molecular replacement)

Data analysis (II): *phasing* *with Direct methods*

A pure theoretical approach demonstrates the existence of a certain number of relations among the phases:

e.g. $\varphi_{-H-K}+\varphi_{H}+\varphi_{K}+2*2\pi=0$ “**triplet relation**”
Exploited, in particular, in the **tangent formula** (Hauptman)

Excellent data at very high resolution can be treated in this way, obtaining phases of sufficient quality to obtain an electron density map.

The method applies in particular for small proteins.

Data analysis (II): *phasing* *with experimental methods*

The MIR (multiple isomorphous replacement) case:

Basic principle: Binding of heavy atoms to the macromolecules ***does not*** change its structure (*Isomorphism between native and DERIVATIVE structures*)

- The presence of the heavy atom(s) introduces differences to the diffraction pattern with respect to the diffraction pattern of the native crystal: The **differences** are in the ***intensities*** of the diffracted X-rays
- When an heavy atom binds ***isomorphously***, then the difference between the two samples are due ***only*** to the presence of the heavy atoms(s)

Frequently used heavy atoms

- Pt, Au, Hg, Pb, Th, U, Re, Os, Ir,
- Pd, Ag (small atomic number) - for small proteins
- J, iodinated tyrosine, modified nucleic acid bases (J, Br)
- Lanthanides (La-Lu) - can substitute Mg^{+2} or Ca^{+2}
- Noble gasses (Xe, Kr)
- Cryo halides (NaBr, KI)

Data analysis (II): *phasing* *with experimental methods*

Determination of heavy atom positions: Patterson Map

The Patterson function is essentially the Fourier transform of the intensities rather than the structure factors:

$$P(u,v,w)=1/V\sum_h\sum_k\sum_l\left|\vec{F}_{hkl}\right|^2\exp[-2\pi i(hu+kv+lw)]$$

u,**v** and **w** are relative coordinates in the unit cell.

- It can always be calculated from the experimental diffraction data (no phase information is needed)

The Patterson map can be written as the convolution of the electron density:

$$P(u,v,w)=\int_0^1\int_0^1\int_0^1\rho(x,y,z)*\rho(x+u,y+v,z+w)\,dV$$

- A patterson map of N points has (N^2-N) peaks and a maximum at the origin;
- The distances of the maxima from the origin depend on the length of the inter-atomic vector
- The intensity of the maxima depends on the product of the atomic numbers of the atoms i and j: Z_iZ_j
- **The peaks are the interatomic distances weighted by the product of the number of electrons in the atoms concerned**

Data analysis (II): *phasing* *with experimental methods*

If we use a **difference Patterson function** we can obtain the Patterson function solely for the heavy atoms in a derivative crystal:

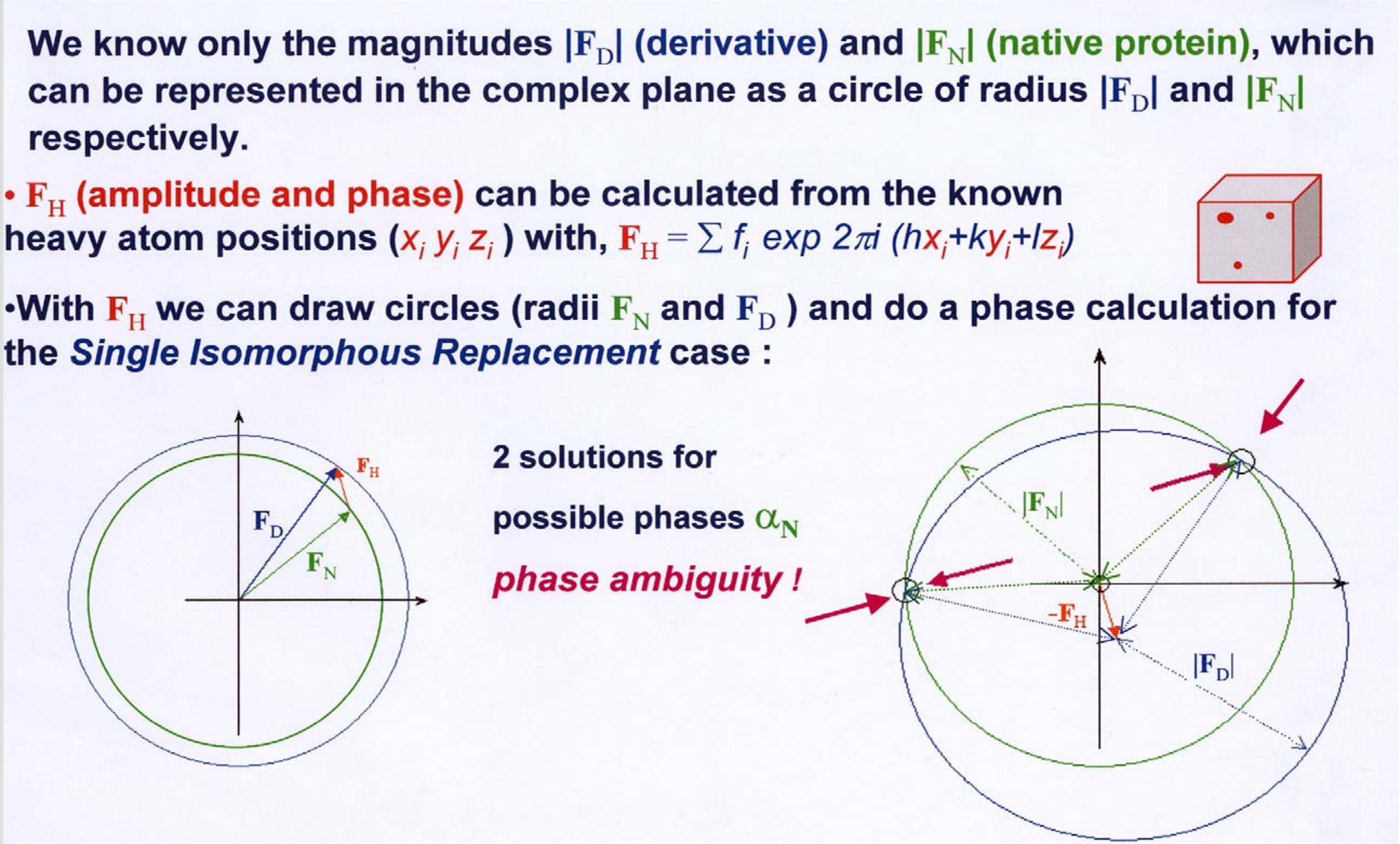
$$P(u,v,w)=1/V\sum_{h}\sum_{k}\sum_{l}\Delta\vec{F}_{hkl}^{\,2}\exp[-2\pi i(hu+kv+lw)]$$

where $\Delta\vec{F}_{hkl}=(|\vec{F}_{PH}|-|\vec{F}_{P}|)_{hkl}$

Deconvolution of the difference Patterson function allows the calculation of heavy atom positions in the crystal unit cell .
It' so possible to get $|\mathbf{F}_{\mathbf{H}}|$ and the phases for **H!**

Data analysis (II): *phasing* *with experimental methods*

The MIR (multiple isomorphous replacement) case:



We know only the magnitudes $|F_D|$ (derivative) and $|F_N|$ (native protein), which can be represented in the complex plane as a circle of radius $|F_D|$ and $|F_N|$ respectively.

- F_H (amplitude and phase) can be calculated from the known heavy atom positions ($x_i\ y_i\ z_i$) with, $F_H = \sum f_i \exp 2\pi i\,(hx_i+ky_i+lz_i)$

•With F_H we can draw circles (radii F_N and F_D) and do a phase calculation for the *Single Isomorphous Replacement* case :

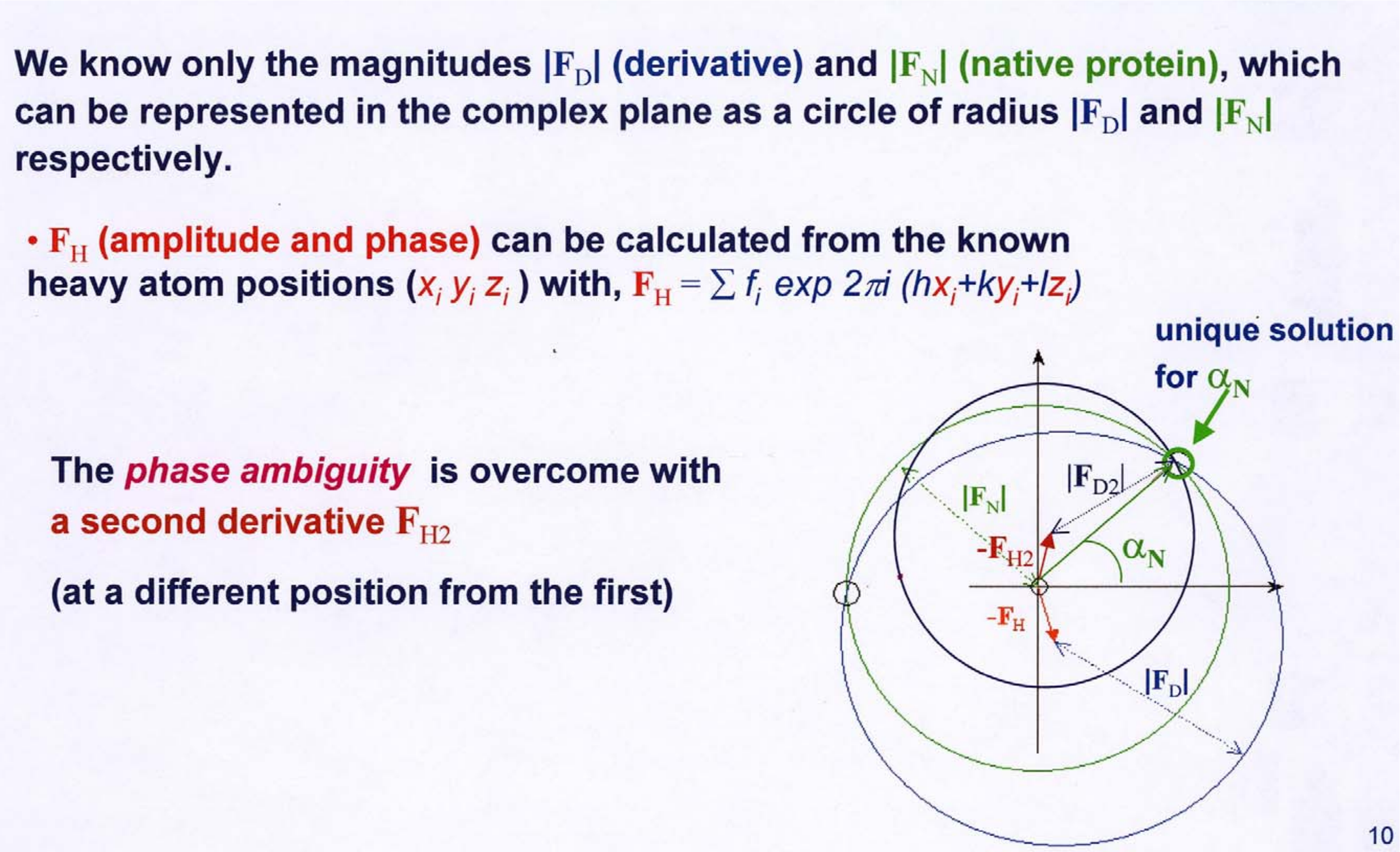
2 solutions for

possible phases α_N

phase ambiguity !

Data analysis (II): *phasing* *with experimental methods*

The MIR (multiple isomorphous replacement) case:

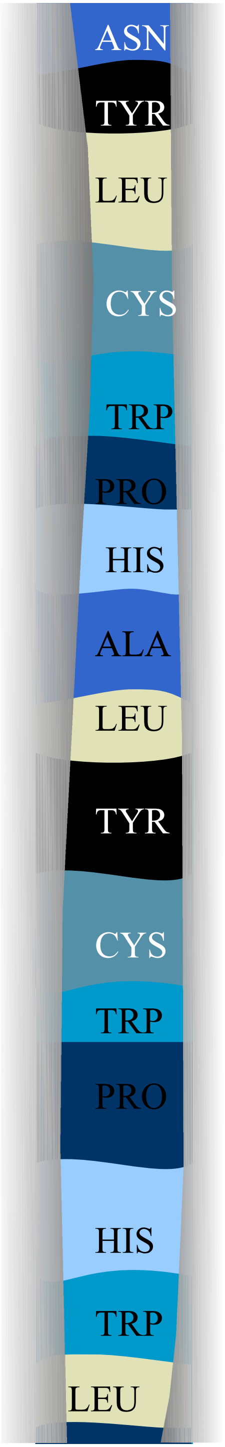


We know only the magnitudes $|F_D|$ (derivative) and $|F_N|$ (native protein), which can be represented in the complex plane as a circle of radius $|F_D|$ and $|F_N|$ respectively.

- **F_H (amplitude and phase) can be calculated from the known heavy atom positions (x_i y_i z_i) with, $F_H = \sum f_i \exp 2\pi i\,(hx_i + ky_i + lz_i)$**

The *phase ambiguity* is overcome with a second derivative F_{H2}

(at a different position from the first)



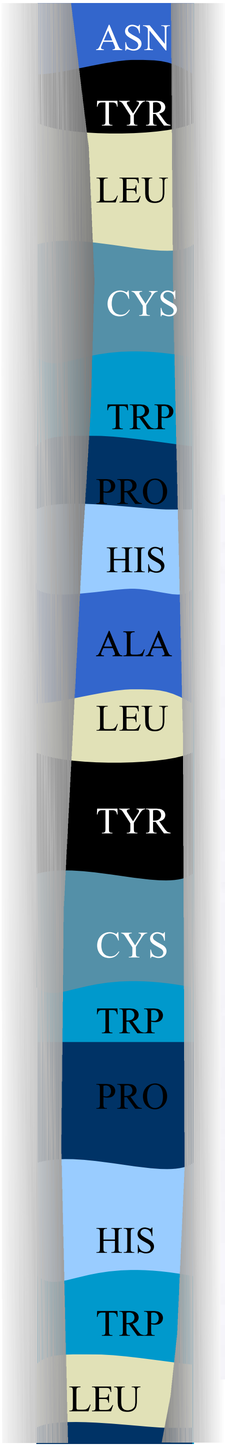
Data analysis (II): *phasing* *using prior knowledge*

•The basic idea of **molecular replacement** is to use a known model of protein which is very similar to the unknown one.

•The amino sequence identity is used to select few candidates (30% sequence identity at least)

•Homologus protein from the **protein data bank**

•Model phases are grafted onto the intensities which are experimentally determined

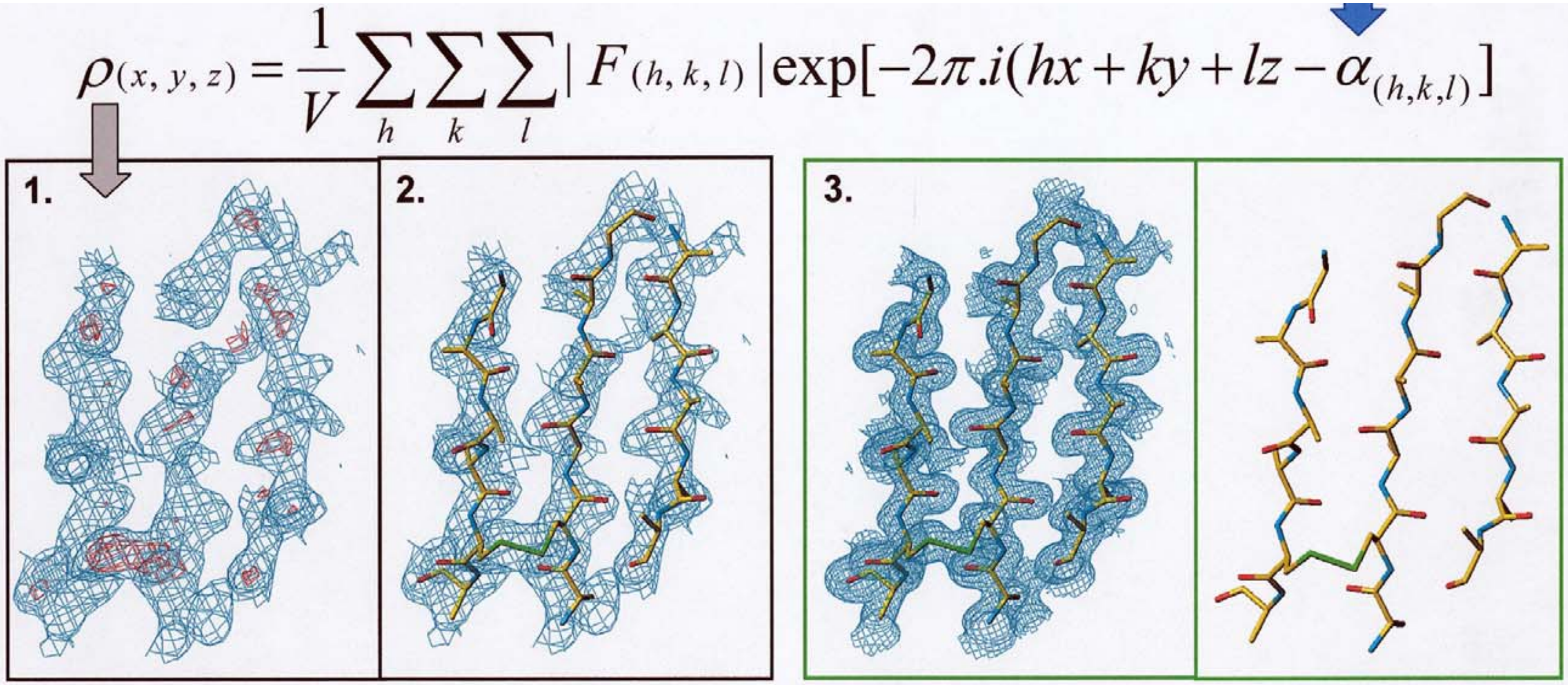


The electron density interpretation:

tracing, docking, electron density improving.

A protein model is fitted into the initial experimental electron density map:

$$\rho(x,y,z) = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}|F(h,k,l)|\exp[-2\pi.i(hx+ky+lz-\alpha_{(h,k,l)}]$$



1. What we “see”
contour map
of the e⁻ density

2. We build a model
(that makes sense)

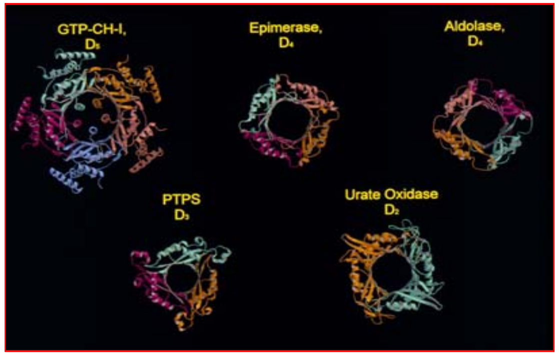
3. The final refined
model (better phases)

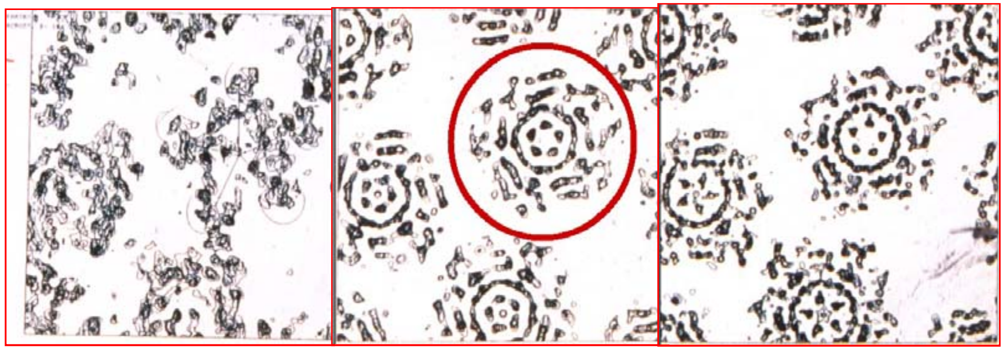
Model
(what is published)

The electron density interpretation:

tracing, docking, electron density improving.

Among the methods to improve initial experimental phases for symmetrical oligomers **cyclic averaging** is the most important



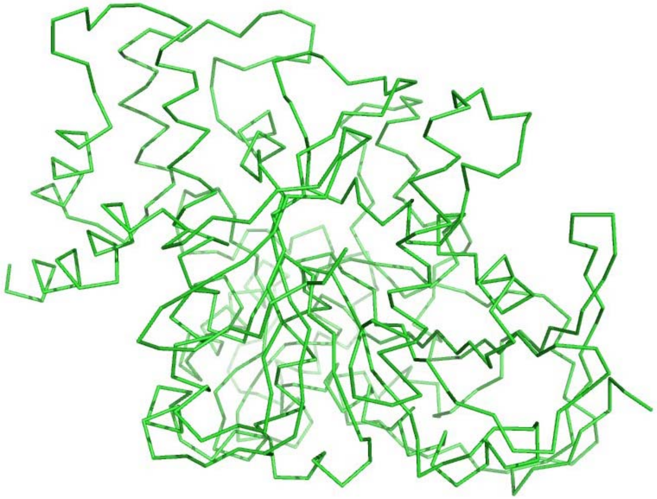


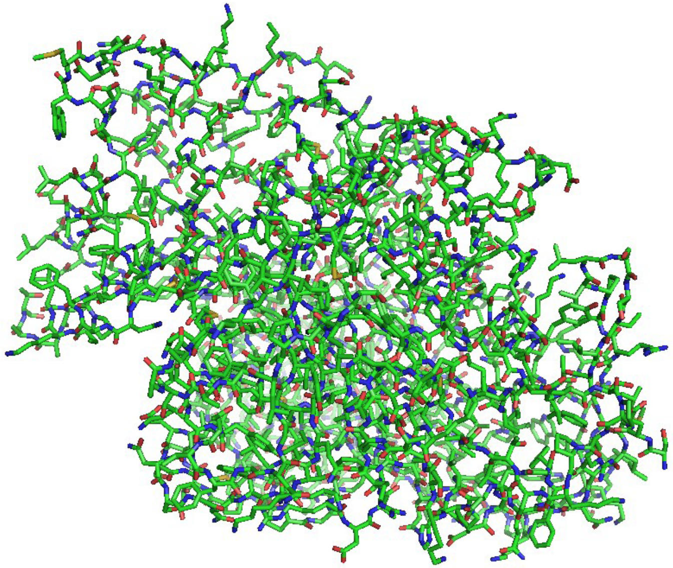
Add the (ordered!) **waters** to improve the density
•Add the **heavy atoms** to your model
•Add **other ligands** (*e.g.* from the crystallization conditions)

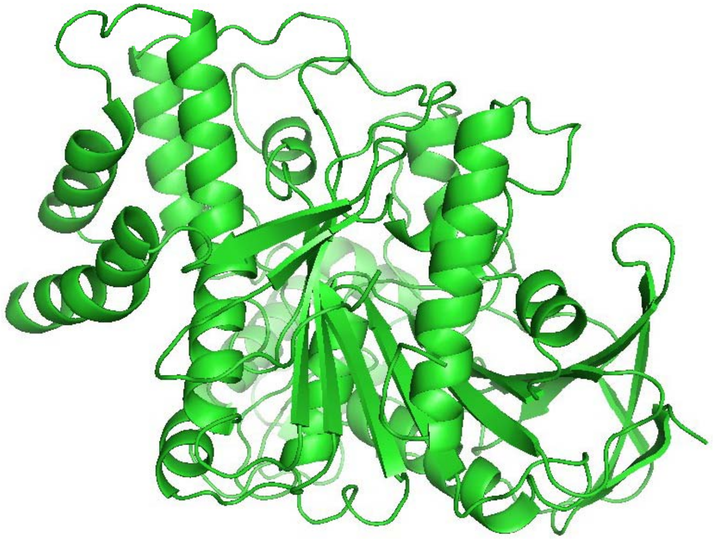
Is your electron density map fully interpreted (is something missing in your model?/ Is there something in your model not present in your maps?)

Keep in mind: Disordered parts will never be seen using diffraction methods!!. Double conformation have to be considered too.

Different ways to represent your model:



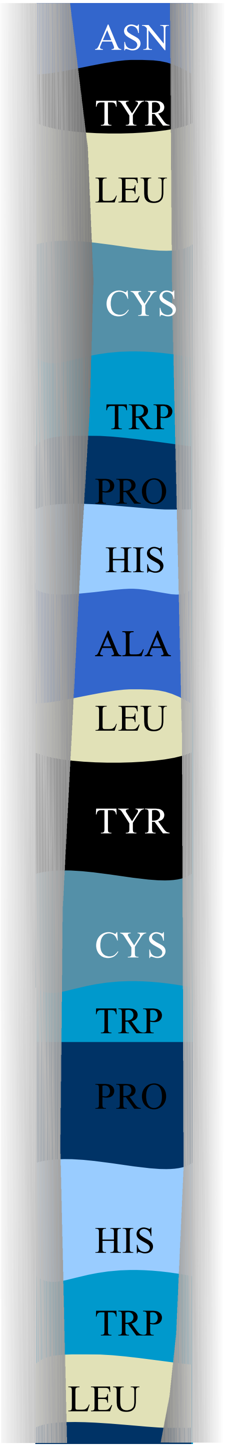




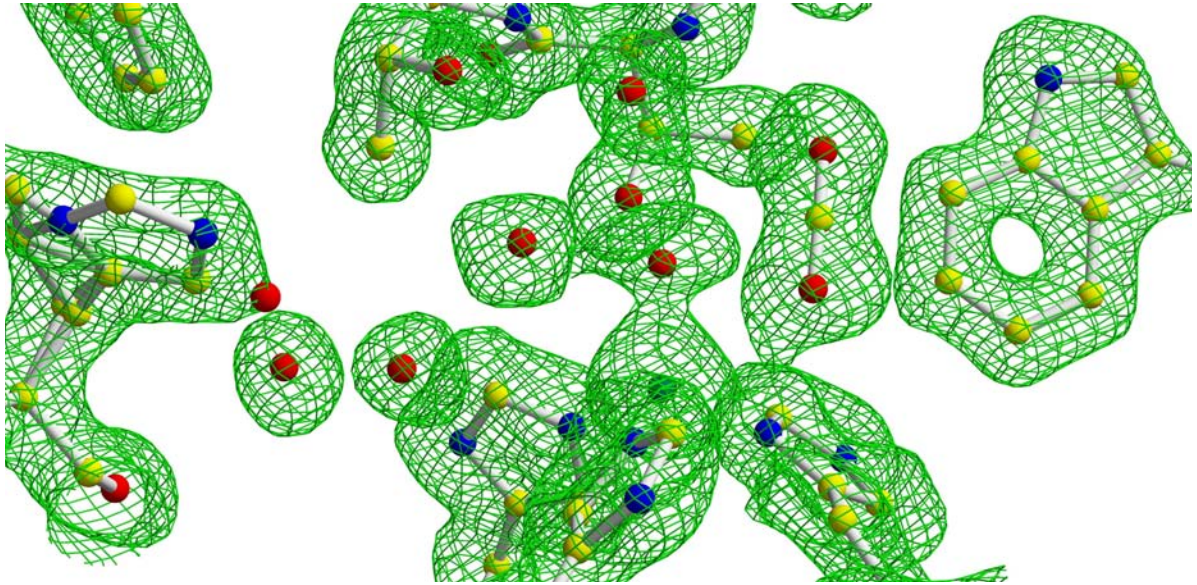
Into the *model*

Take a look into it:

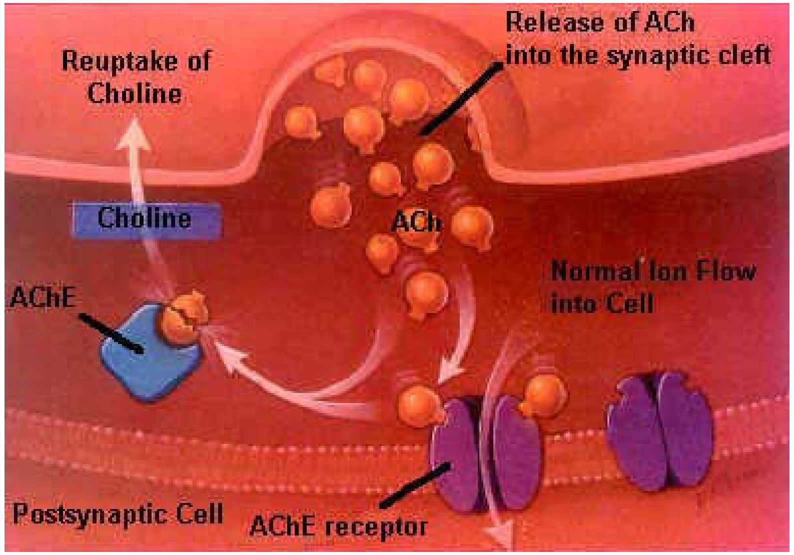
•Is there a clear active site?
•How does it work? Which are the most important ammino acids?
•Which are the physical-chemical properties of the active site?
•Trace the surface: How does it interact with other proteins?
•How can we build a small molecule to inhibit it (partially or in a total manner) [**pharma**]
•How can we engineer it in order to use it in a industrial process? [**biotech**]

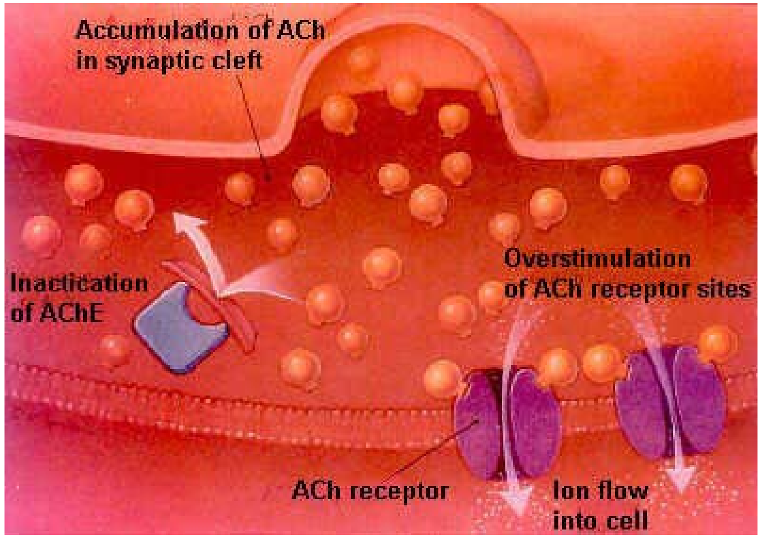


… and applications.



Rational drug design for Alzheimer's desease


Release of ACh
into the synaptic cleft
Reuptake of
Choline
Choline
ACh
AChE
Normal Ion Flow
into Cell
Postsynaptic Cell
AChE receptor


Accumulation of ACh
in synaptic cleft
Inactivation
of AChE
Overstimulation
of ACh receptor sites
ACh receptor
Ion flow
into cell