Protein Crystallography

Mitchell Guss
University of Sydney
Australia

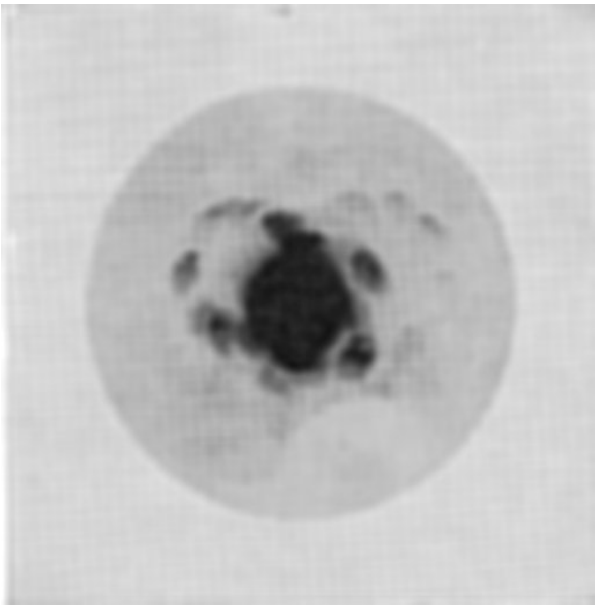
THE UNIVERSITY OF SYDNEY

Outline of the talk

- Recap some basic crystallography and history
- Highlight the special requirements for protein (macromolecular) structure determination
- Emphasize the role played by synchrotron radiation
- Introduce some of the future challenges and opportunities offered by newly emerging technologies

100+ years of X-ray crystallography

- 1912 Max von Laue and his colleagues demonstrate that a crystal (copper sulfate) can diffract X-rays
- Awarded the Nobel Prize 1914

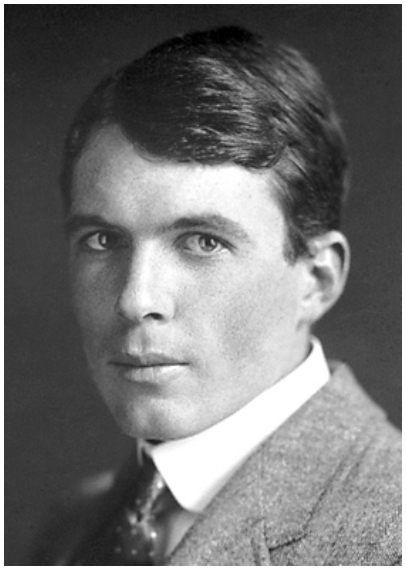


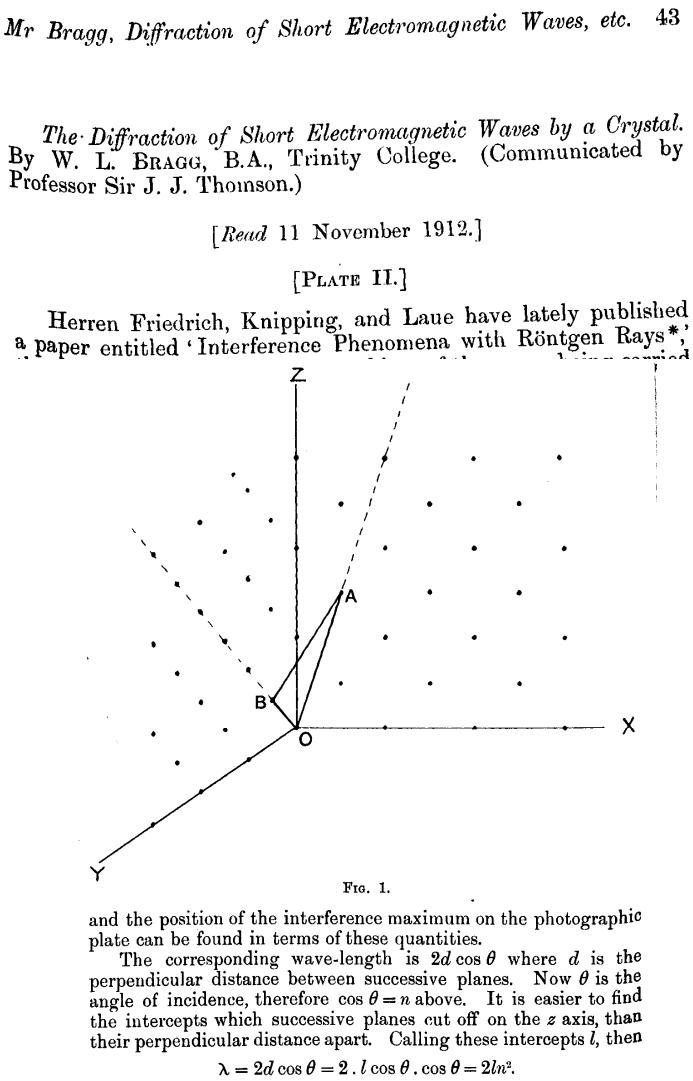
Friedrich, Knipping and von Laue, 1912

- Showed that X-rays behaved as waves
- Failed to correctly relate the scattering to the underlying structure

100+ years of X-ray crystallography

- 1912 William Lawrence Bragg interpreted the diffraction from sodium chloride
- Awarded the Nobel Prize with his father (W.H. Bragg) in 1915
  - W.L. Bragg still remains the youngest science Nobel laureate at 25 years of age





Mr Bragg, Diffraction of Short Electromagnetic Waves, etc. 43

The Diffraction of Short Electromagnetic Waves by a Crystal.
By W. L. Bragg, B.A., Trinity College. (Communicated by Professor Sir J. J. Thomson.)

[*Read* 11 November 1912.]

[Plate II.]

Herren Friedrich, Knipping, and Laue have lately published a paper entitled 'Interference Phenomena with Röntgen Rays*,'

Fig. 1.

and the position of the interference maximum on the photographic plate can be found in terms of these quantities.

The corresponding wave-length is $2d \cos\theta$ where d is the perpendicular distance between successive planes. Now θ is the angle of incidence, therefore $\cos\theta = n$ above. It is easier to find the intercepts which successive planes cut off on the z axis, than their perpendicular distance apart. Calling these intercepts l, then

$$\lambda = 2d \cos\theta = 2 \, . \, l \cos\theta \, . \cos\theta = 2l\mu^2.$$

Braggs, 1912

- Showed that the diffraction was equivalent to the reflection from planes drawn through the crystal
- Could predict the positions of the diffracted spots using Bragg's law: $\lambda = 2d\sin\theta$
- Solved many inorganic structure (e.g. NaCl) in the next few years

Timeline of biological crystallography - 1

- 1895 Röngten: discovers X-rays (Nobel 1901)
- 1912 von Laue: demonstrates diffraction (Nobel 1914)
- 1912 Braggs: first crystal structure (Nobel 1915)
- 1929 Sumner: crystallizes an enzyme (Nobel 1946)
- 1951 Pauling: protein components α-helix and β-sheet (Nobel Chemistry 1954, Peace 1962)
- 1953 Watson, Crick & Wilkins: DNA structure (Nobel 1962)
- 1954 Crowfoot-Hodgkin: vitamin B12, penicillin (Nobel 1964)
- 1959 Kendrew & Perutz: myoglobin and haemoglobin (Nobel 1962)

Timeline of biological crystallography - 2

- 1970+ Klug crystallographic electron microscopy and nucleic acid-protein complexes (Nobel 1982)
- 1985 Michel, Deisenhofer & Huber: the photoreaction centre, first membrane protein (Nobel 1988)
- 1997 Boyer & Walker: ATPsynthase a rotating molecule (Nobel 1998)
- 1998 MacKinnon: K+ ion channel (Nobel 2003)
- 2000 Yonath, Ramakrishnan & Steitz: the ribosome (Nobel 2009)
- 2011 Lefkowitz & Kobilka: G-protein coupled receptors (Nobel 2012)

The beginnings of (X-ray) protein crystallography

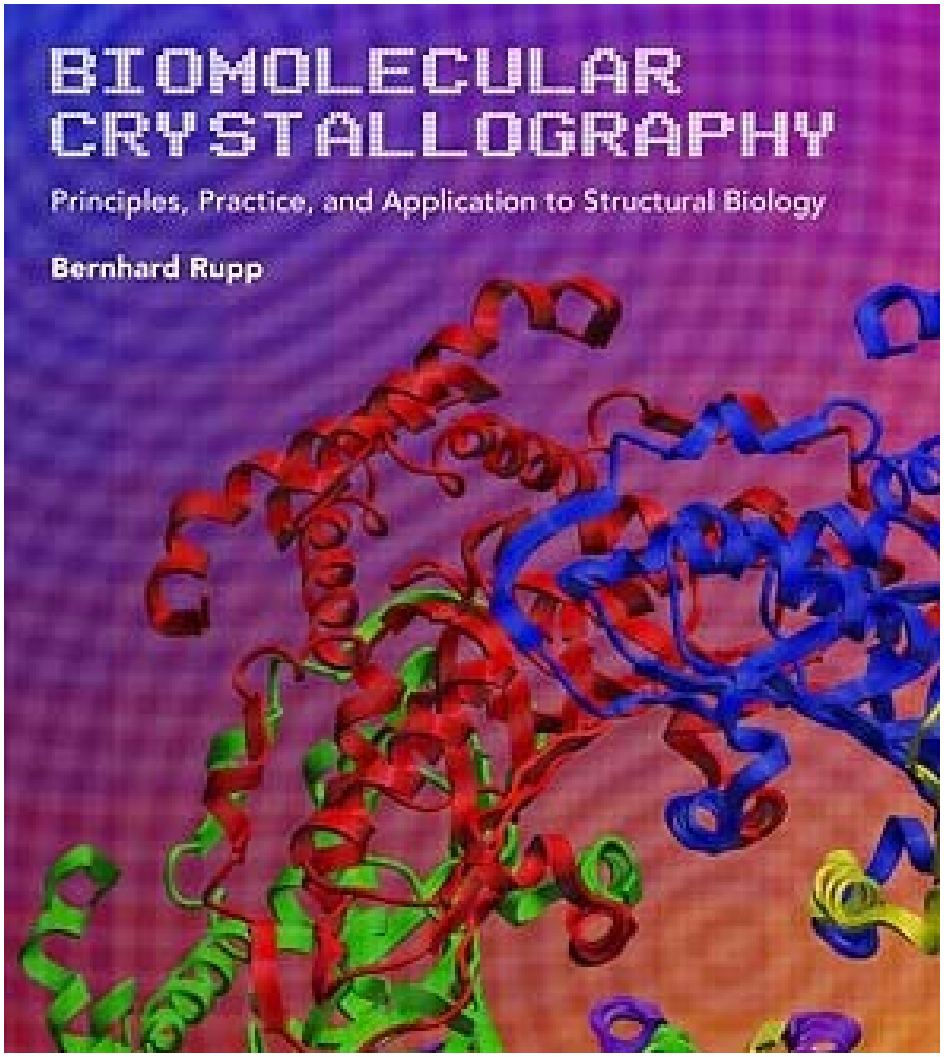
- 1930's Dorothy Crowfoot (later Hodgkin) and J.D. Bernal made the first X-ray diffraction pictures of an enzyme – they realised that the crystals had to be kept wet or they didn't diffract
- 1937 Max Perutz began the structure analysis of haemoglobin with W.L. Bragg as his supervisor – the structure was not solved until 1961!

What has changed from 1937 to today that enables the same structure to be solved in minutes?

Excellent text for biomolecular crystallography

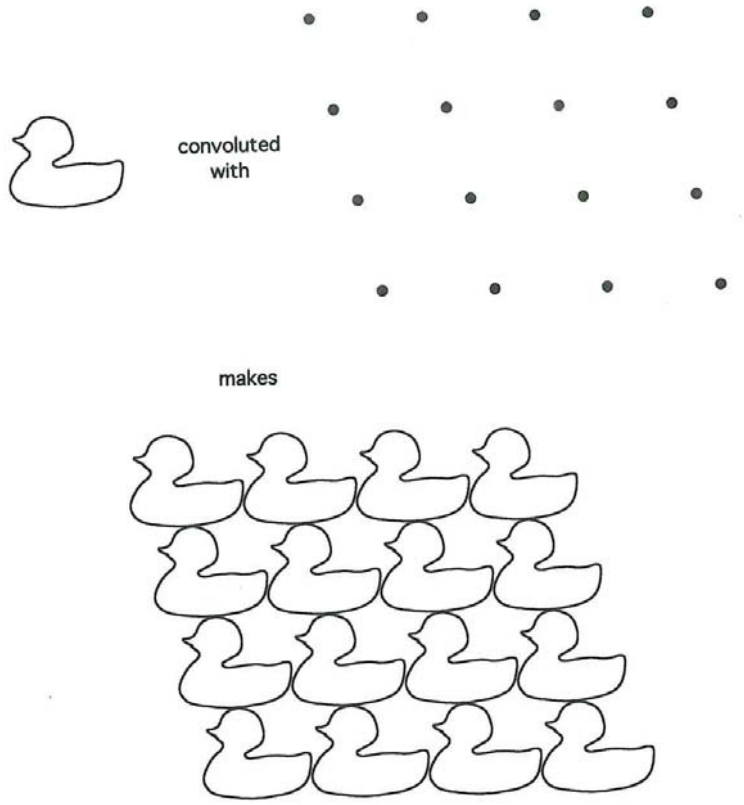
- Excellent coverage of both the theory and practice of the subject
- Many of my illustrations come from this text
- See Bernard Rupp's website for more detail

http://www.ruppweb.org/default.htm



What is a crystal?

- An solid containing of atoms or molecules that repeats in three dimensions
- *Or,* the convolution of a motif (molecule) with a lattice

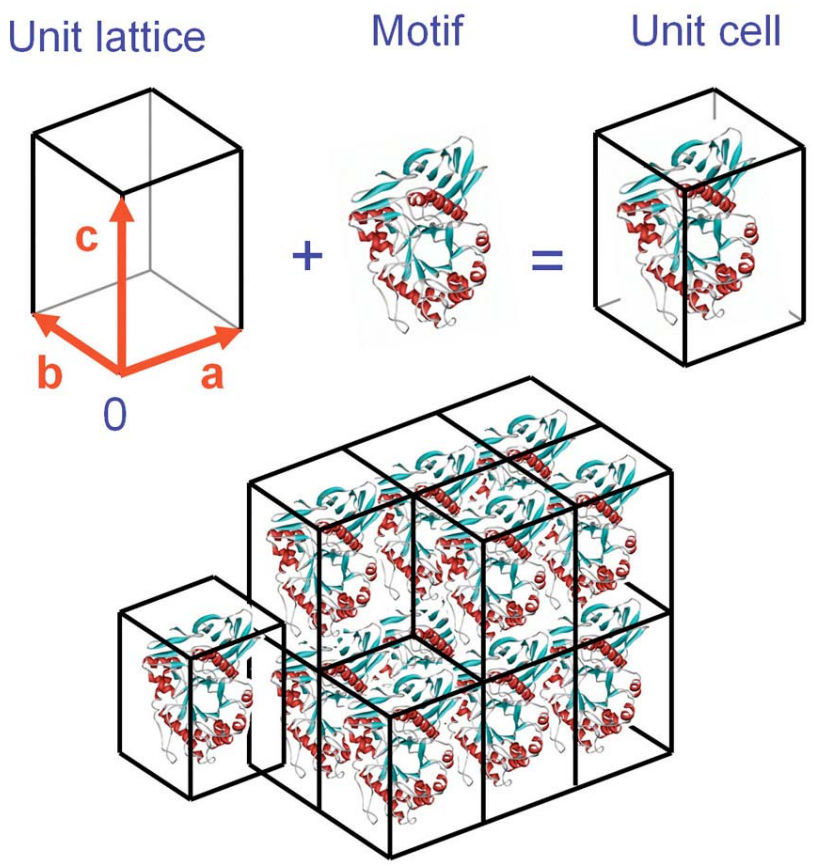


In this case a “duck” is convoluted with a 2-dimensional lattice to give a 2D “duck crystal”

Convolution
- Place an instance of the motif at each point of the lattice

3-dimensional protein crystals

- The unit lattice is the 3D set of points described by the 3 non-collinear vectors that give the repeating units
- The motif is the content of the unit lattice (Note: it may be one or a number of actual molecules)
- The unit cell is the unit lattice with its contents
- The crystal is the 3D array of unit cells that fill space

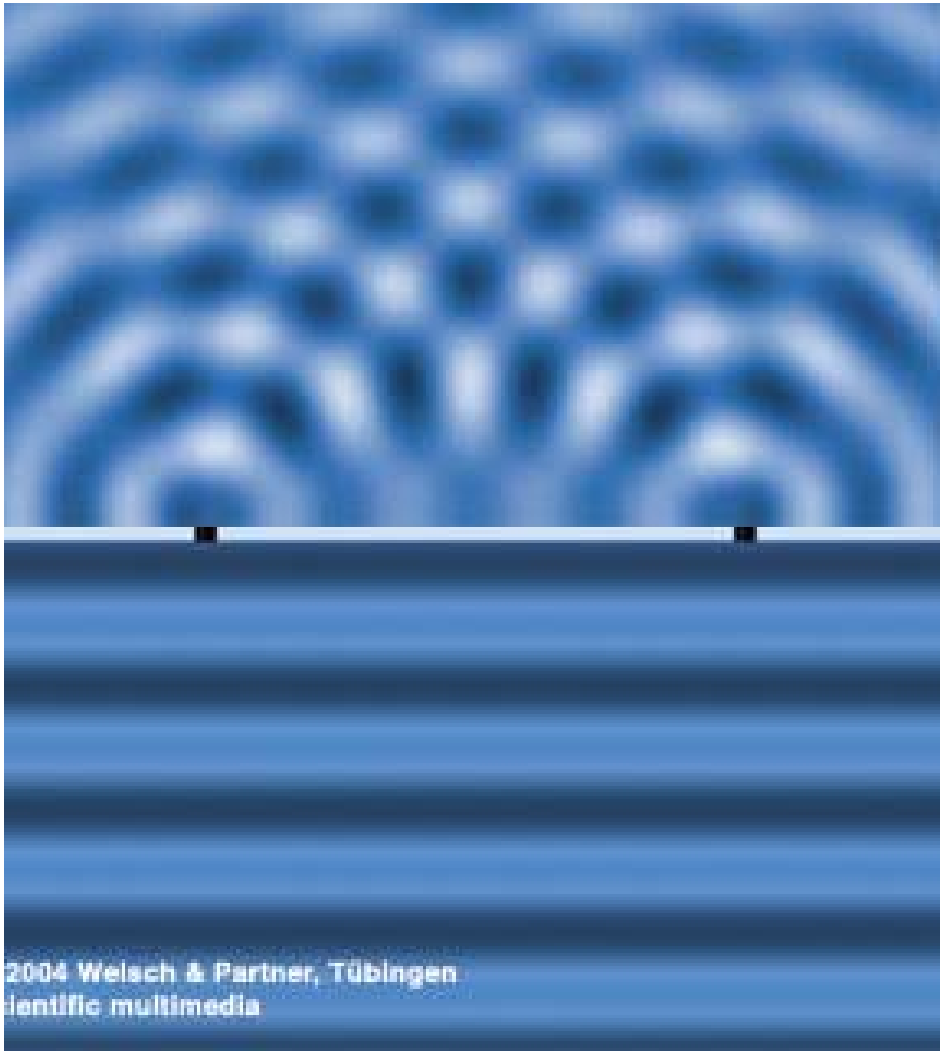


Not true for all systems anymore. Why?

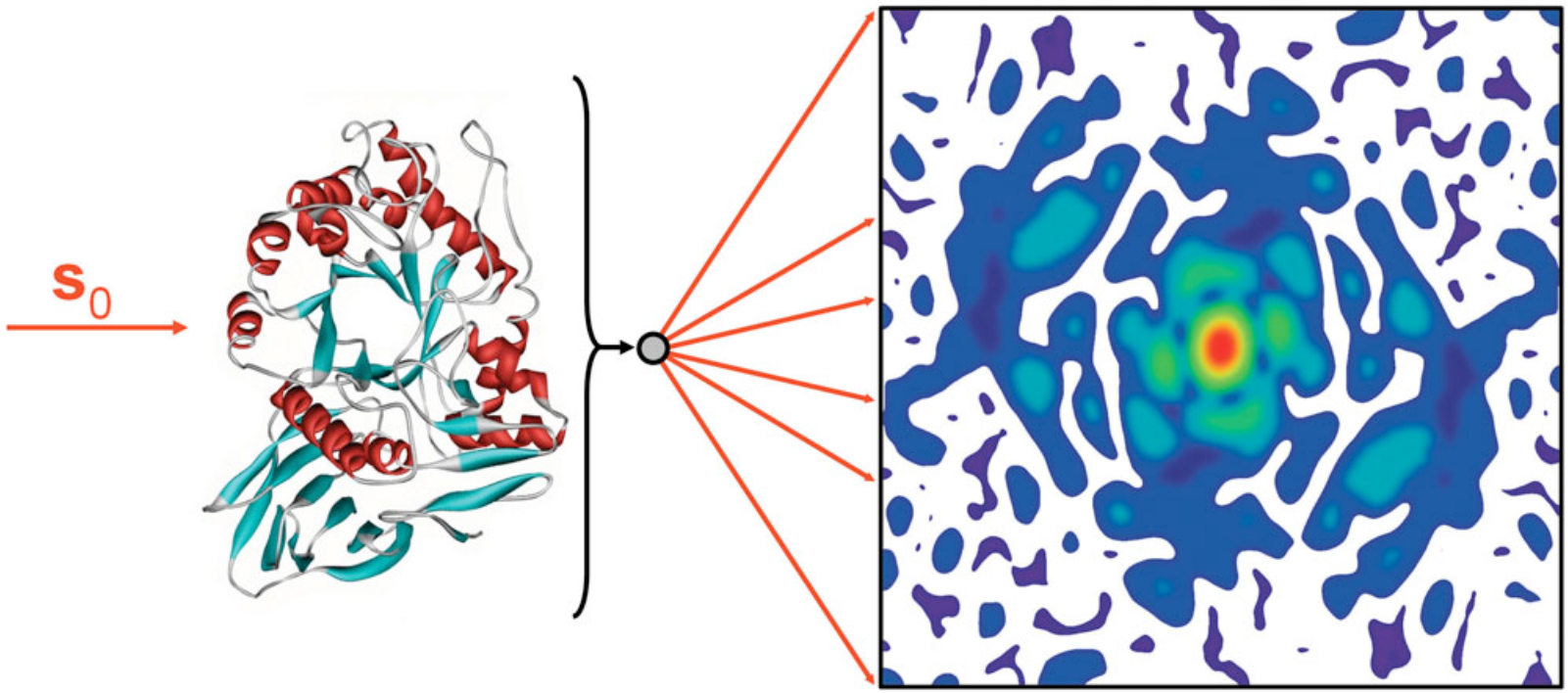
What is diffraction?

- Diffraction takes place with sound; with electromagnetic radiation, such as light, X-rays, and gamma rays; and with very small moving particles such as atoms, neutrons, and electrons, which show wavelike properties.
- The phenomenon is the result of interference (i.e., when waves are superposed, they may reinforce or cancel each other out) and is most pronounced when the wavelength of the radiation is comparable to the linear dimensions of the obstacle.

Since X-rays are of approximately the same wavelength as inter-atomic bonds they will be diffracted by molecules and crystals.



Scattering from a molecule



The image is the scattering of the 2D projection protein molecule on the left. It is the sum of the scattering from each atom (*j* at position *S*). The diffraction (F_S) is continuous and can occur in any direction ***S***.

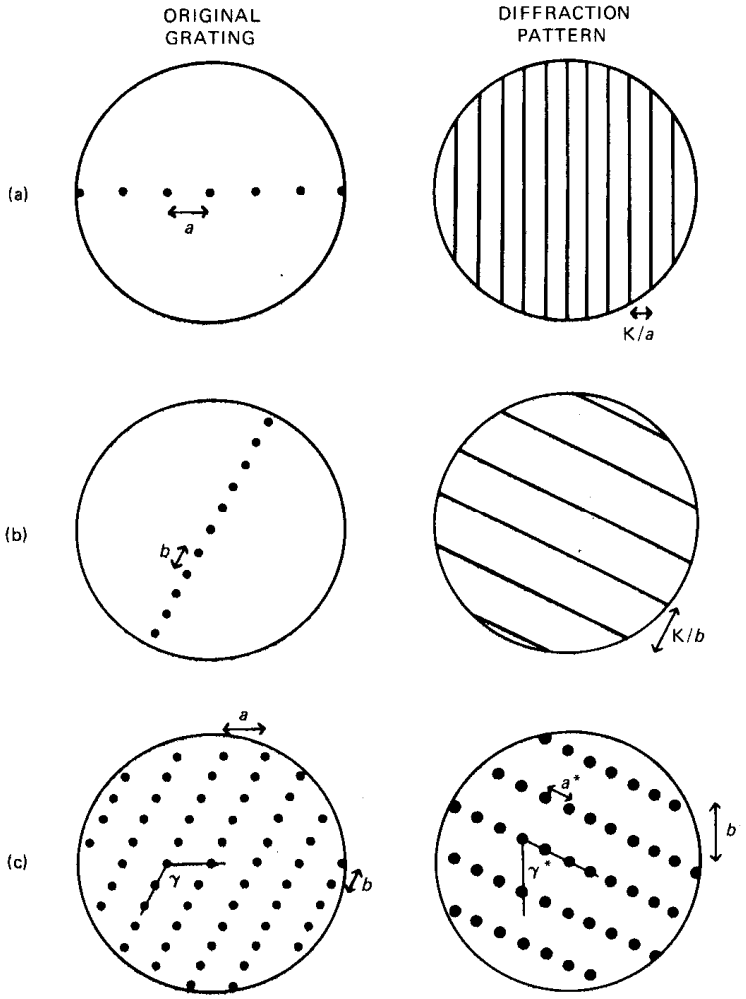
$$F_S = \sum_{j=1}^{atoms} f_{Sj}^0 . \exp(2\pi i S r_j)$$

Diffraction from a lattice

(a, b) Diffraction from a line of evenly spaced dots (or holes) is a series of lines perpendicular to the row of dots. The lines have inverse spacing to the dots.

(c) Diffraction from a 2D array of dots is an array of dots at right angles to the original rows.

The array of dots in (c) is the convolution of the dots in (a) and (b). The diffraction pattern in (c) is the product of the diffraction patterns in (a) and (b). What you see in the (c) is a real lattice and the associated reciprocal lattice

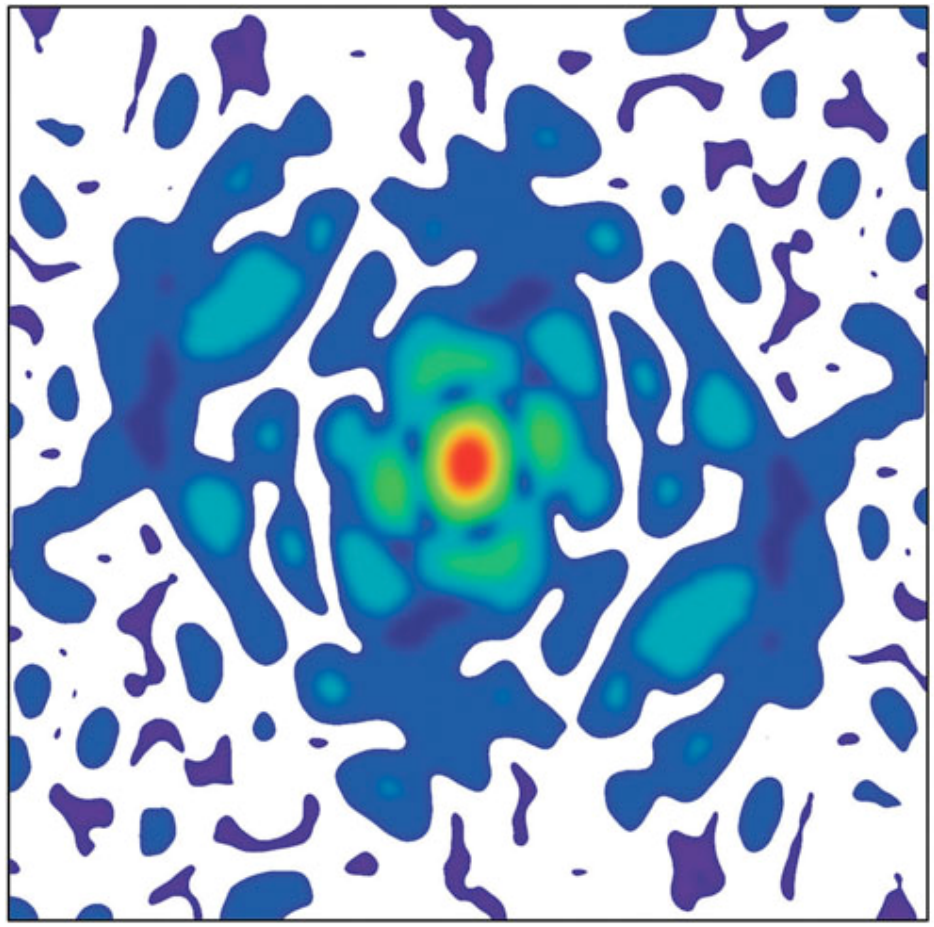


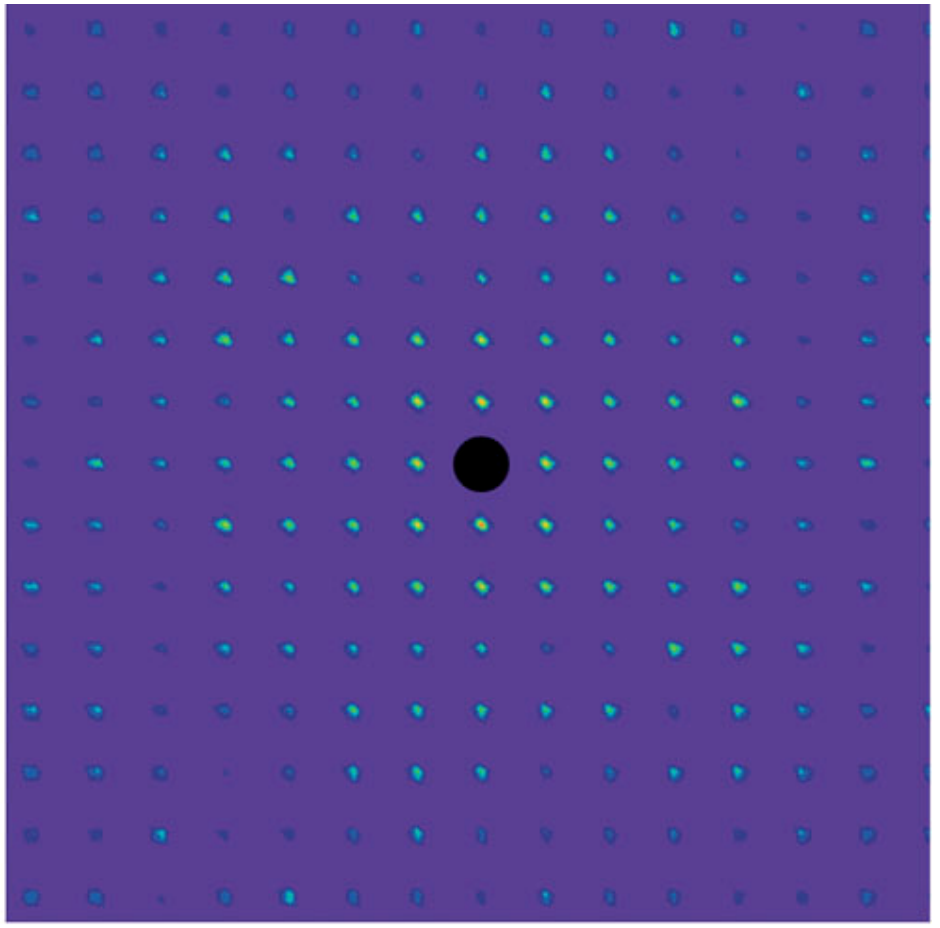
The convolution theorem & diffraction from a crystal

The convolution theorem states that under suitable conditions the Fourier transform of a convolution of two functions is the point wise product of Fourier transforms of the individual functions.

Remember a “crystal” is the convolution of a lattice and a “molecule”. Therefore the diffraction from a crystal will be the product of the diffraction of the lattice and of one molecule.

Scattering from a protein crystal

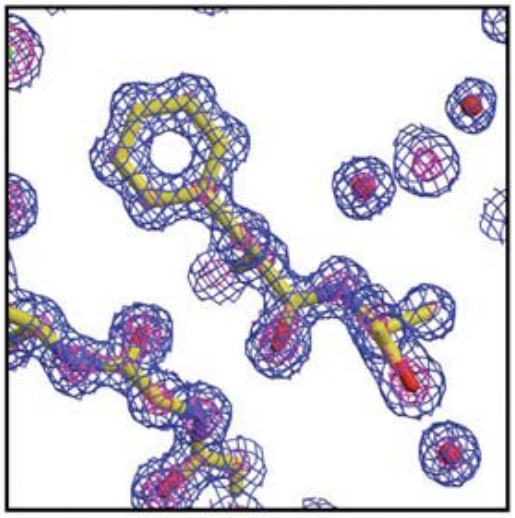


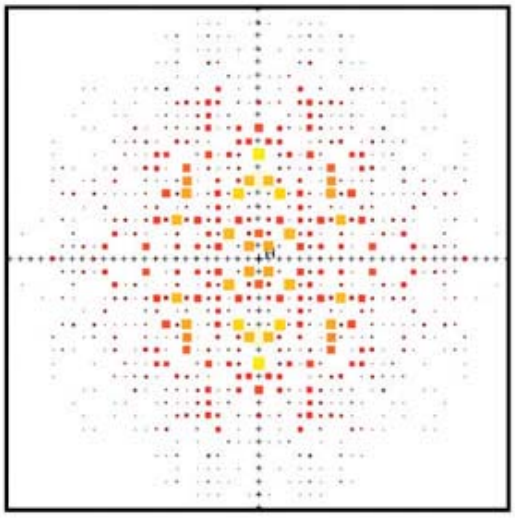


The effect of the crystal is to reinforce the diffraction signal at points where the waves from all the molecules are in phase. This makes the signal sufficiently large to be measured but all the information is theoretically in the scattering from one molecule.

The so-called ‘phase problem’

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

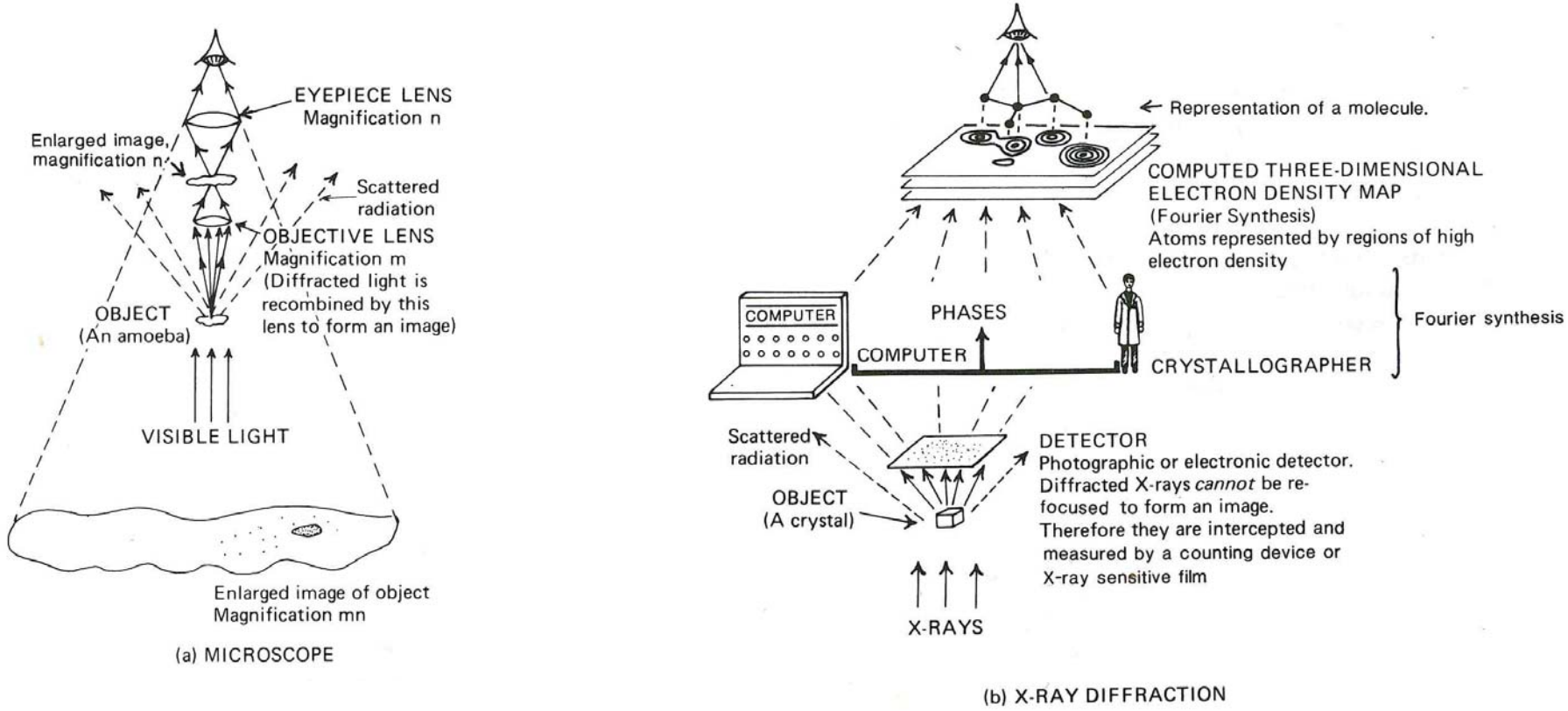




The crystallographic phase problem

$$I(hkl) \propto |\mathrm{F(hkl)}|^2$$

The crystallography experiment

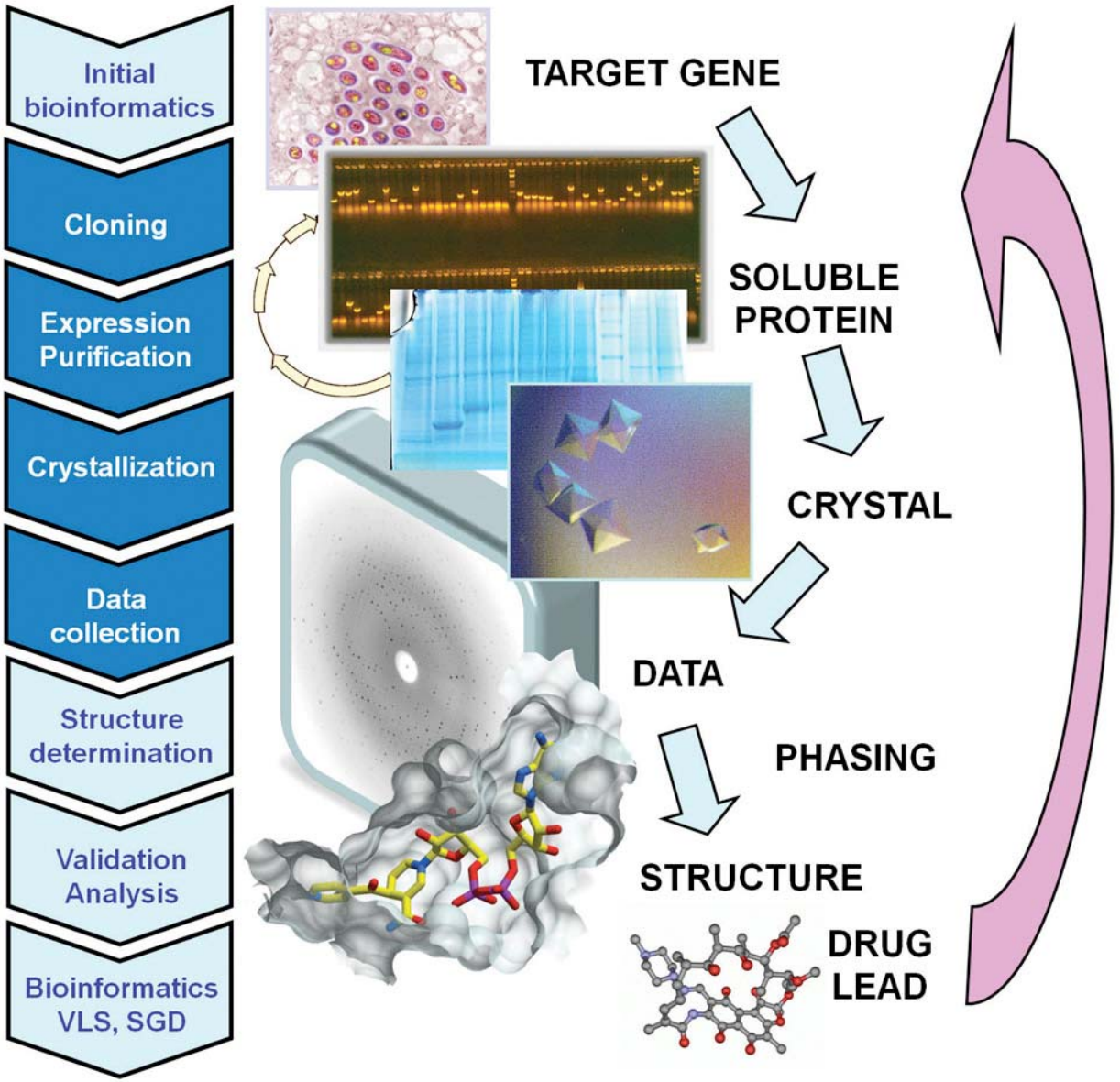


The X-ray crystallography experiment is like a light microscope except (1) use X-rays and not visible light and (2) we have no objective lens for X-rays so we cannot refocus to make the image. In other words while we can measure the amplitudes of the diffracted waves we cannot measure their relative phases. This is the "phase problem" of crystallography.

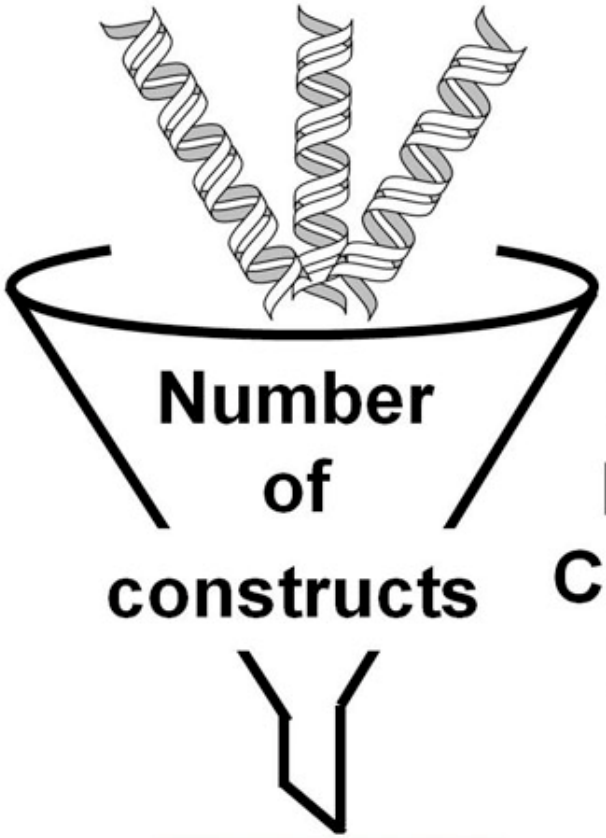
What is so special about protein crystallography?

- Asymmetric unit contains more atoms
  - Even a small protein (50 aa residues) has about 500 non-H atoms. A large protein may of many thousands
- The unit cells are generally larger ($V > 125{,}000 Å^3$)
  - Diffraction is weaker ($I_{hkl} \propto 1/V_c$)
- Crystals diffract to lower resolution
  - Poorly ordered typically 30 – 80% solvent by volume
- Combination of lower resolution, large atom numbers and non-centric space groups
  - direct methods for structure solution applicable to small molecules do no work for proteins

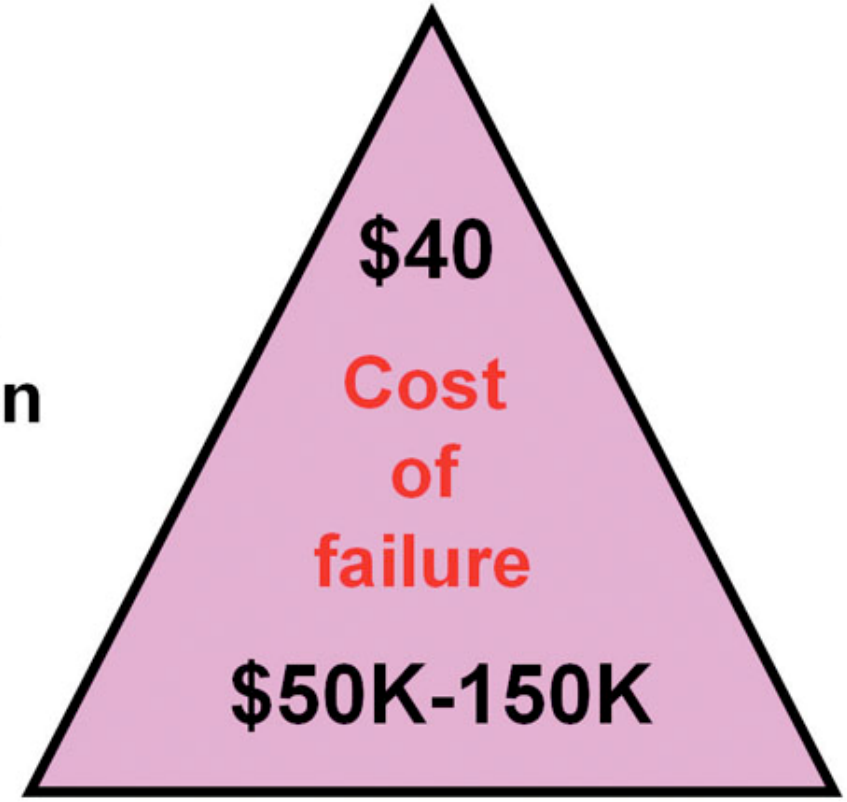
Overview of protein structure determination



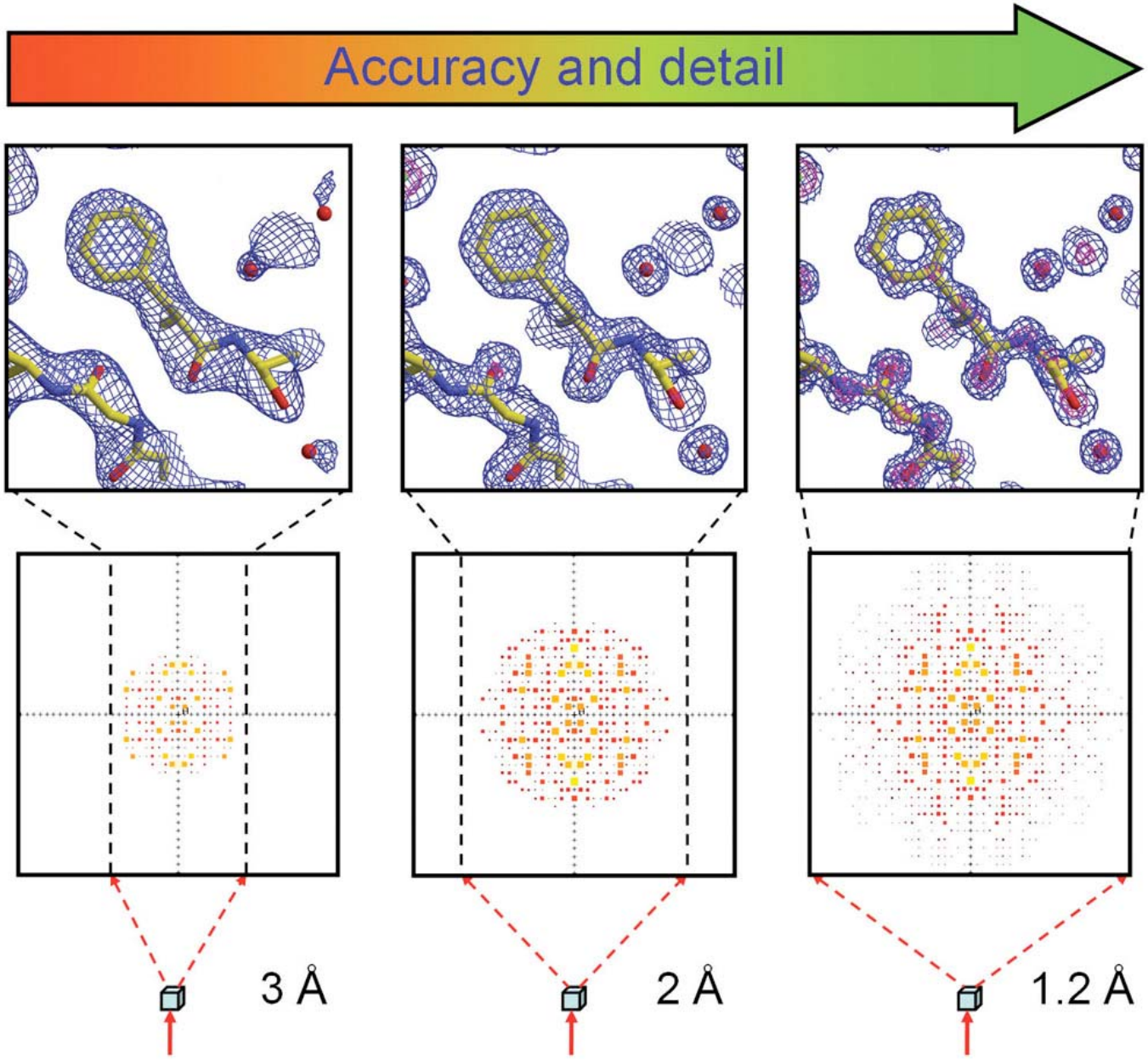
Life wasn't meant to be easy!



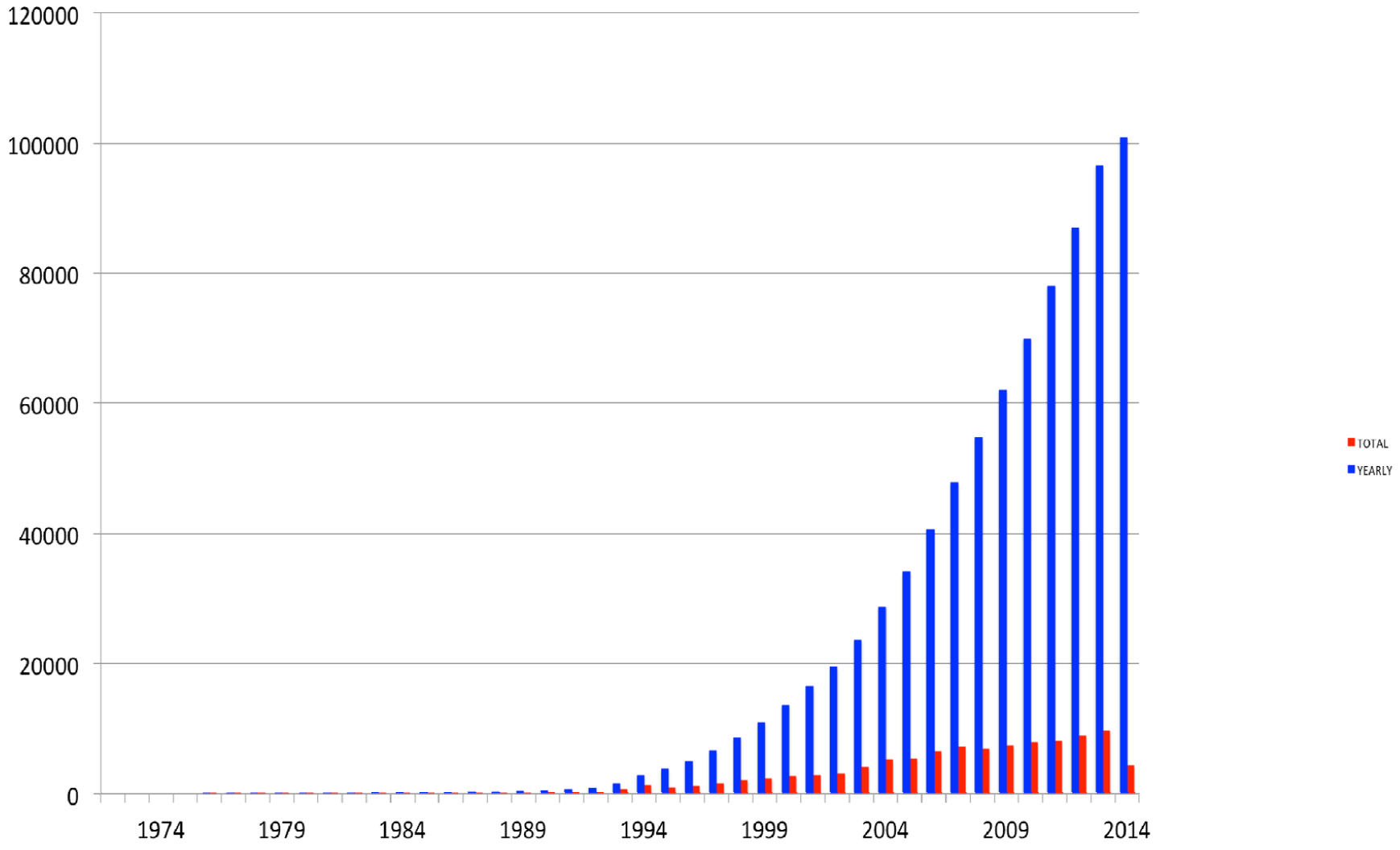
Cloning
Expression
Purification
Crystallization
Harvesting
Data
collection
Structure
determination



Importance of data quality (resolution)

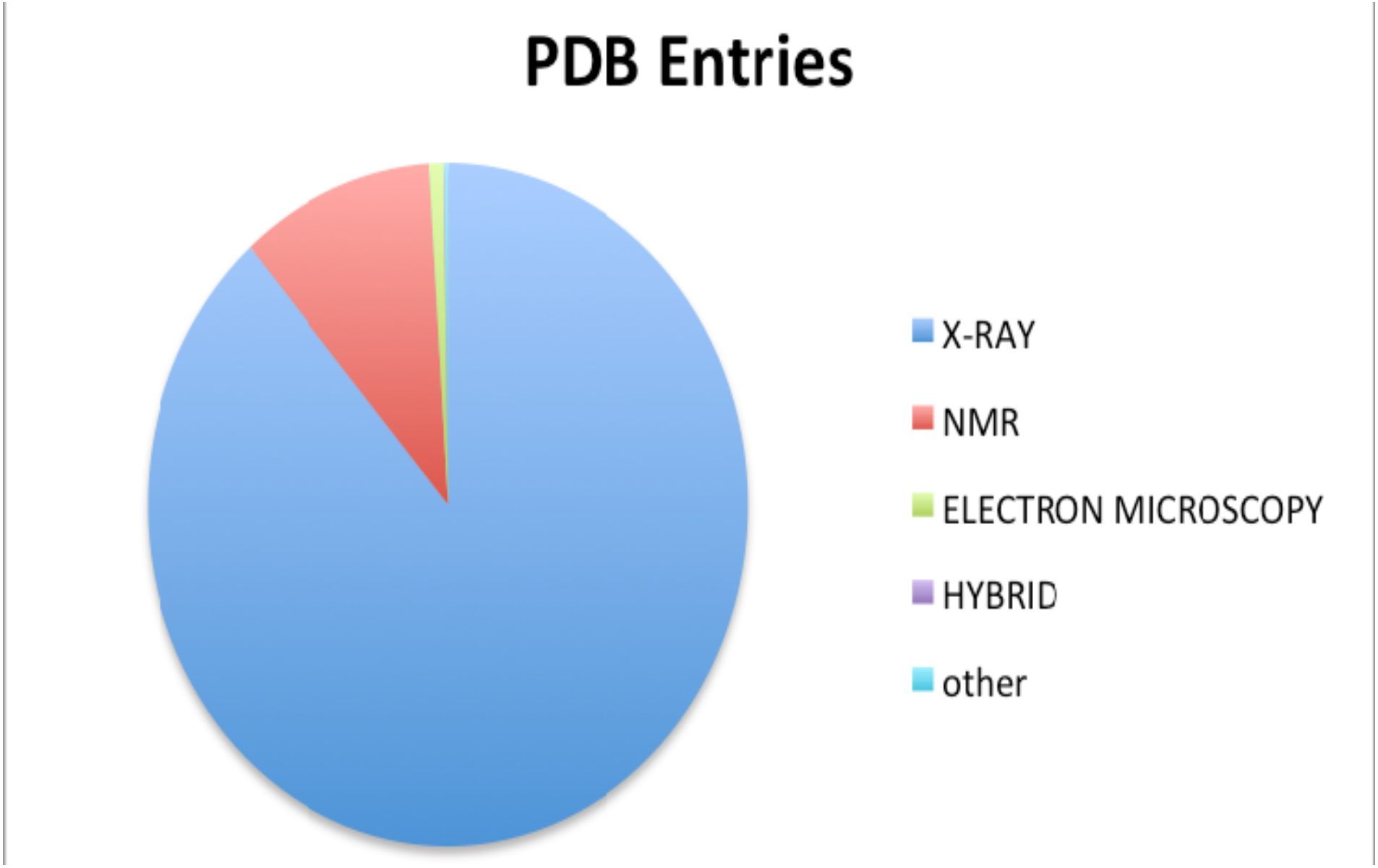


What has been achieved since 1962



Protein Data Bank (PDB) depositions by year

PDB depositions by technique

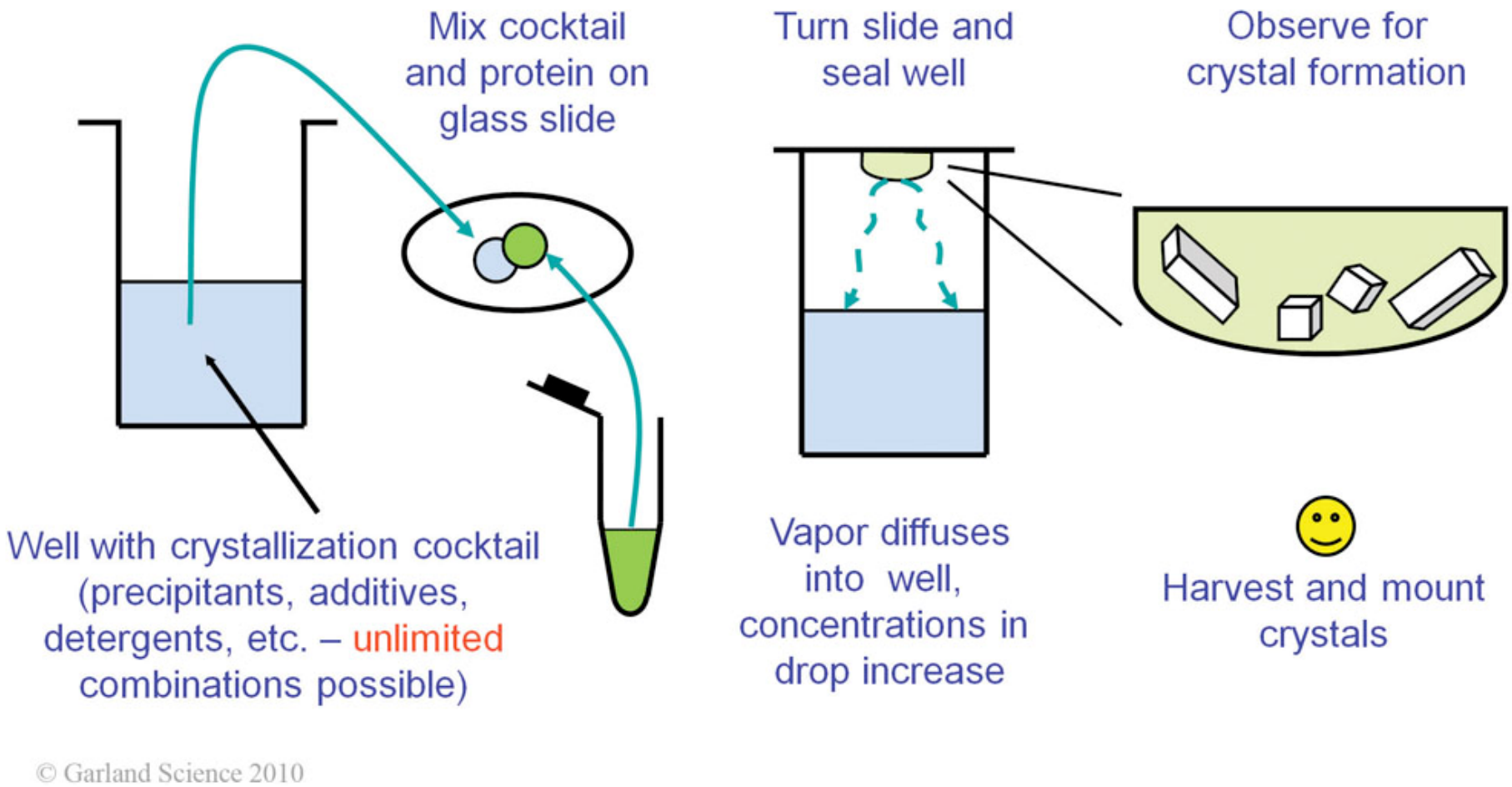


The experiment

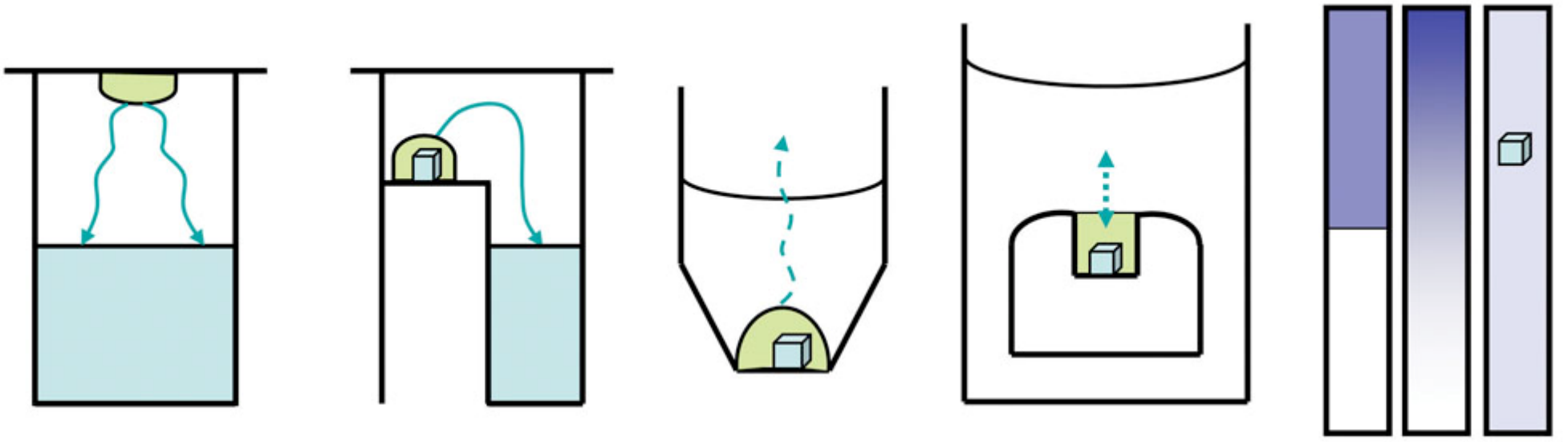
1. Expression of proteins using molecular biology instead of isolation from natural sources – makes more and most importantly purer protein
2. Grow crystals – the most uncertain and often the most time consuming step – use of factorial screens and robotics
3. Record diffraction data – high speed detectors and intense synchrotron sources can reduce what once took months to minutes or even seconds
4. Solve the structure – determine the phases. New methods including anomalous phasing have significantly improved this step although it can still be difficult in some cases
5. Build a model – aided by highly intelligent computer graphics
6. Refine the structure – optimise the model to fit the data. Fast computers have reduced this step from hours to seconds
7. Deposit the structure in the Protein Data Bank, and;
8. Write a paper – often the slowest step

CRYSTALLISATION

Vapour diffusion – in theory!



Different apparatus



The classic: hanging-drop vapor diffusion

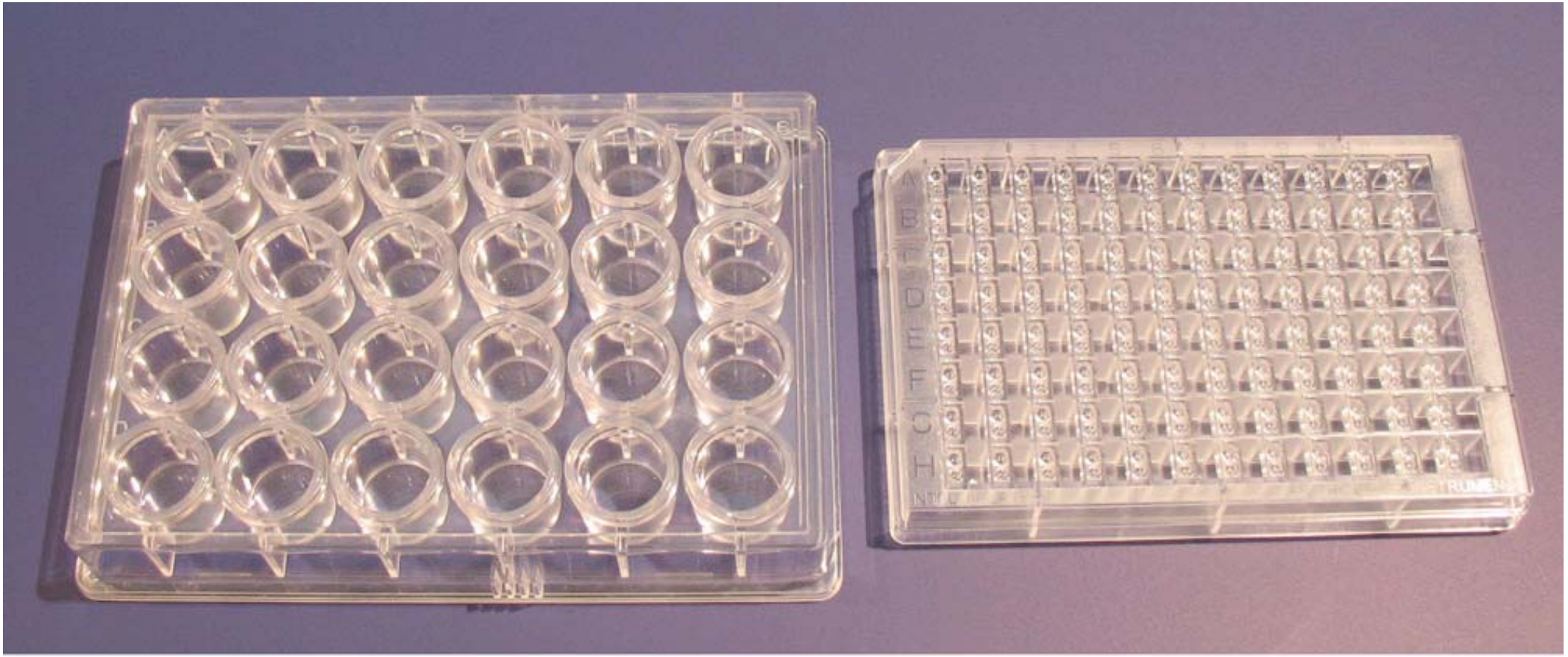
The variant: sitting-drop vapor diffusion

Micro-batch under oil

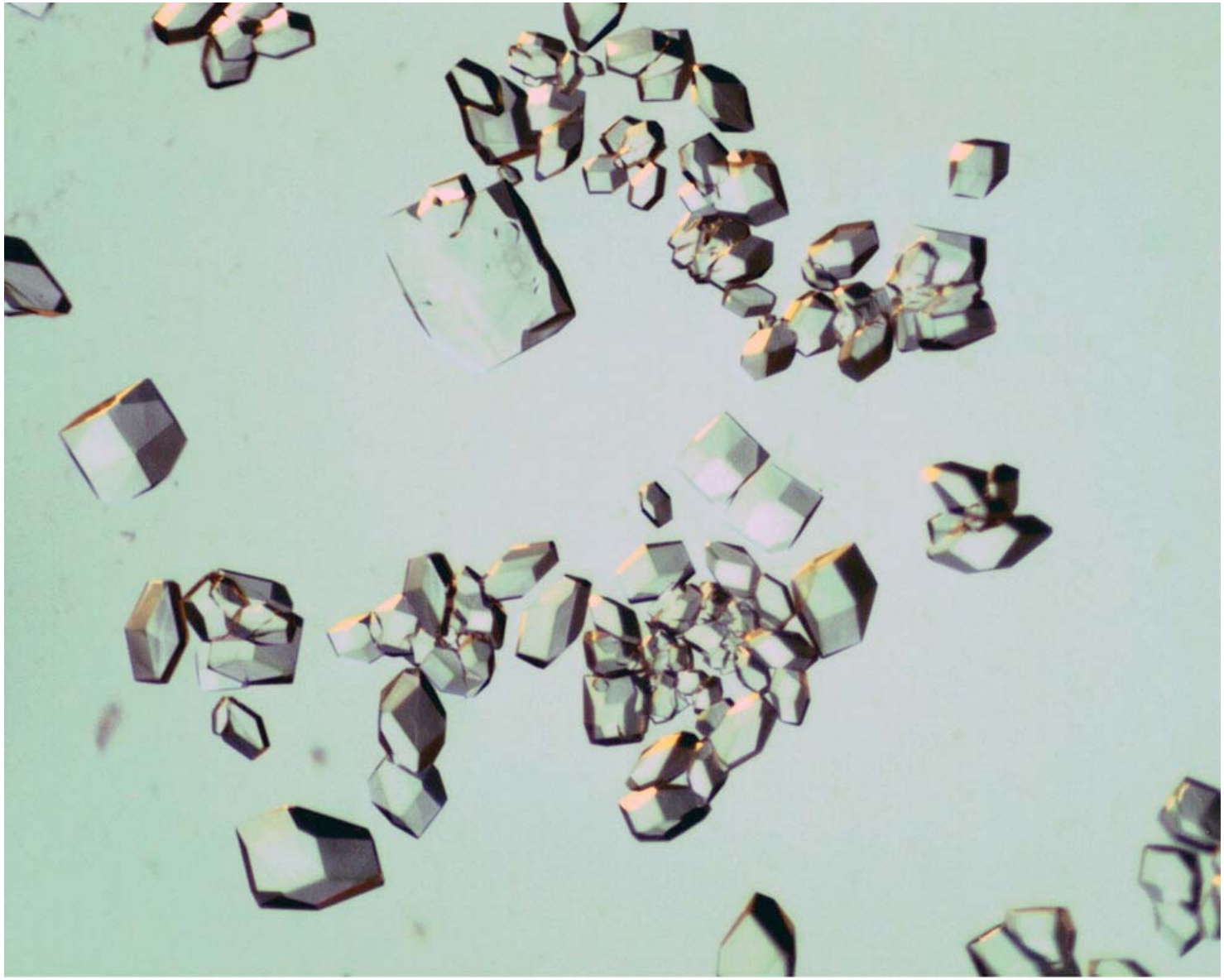
Micro-dialysis

Free-interface diffusion

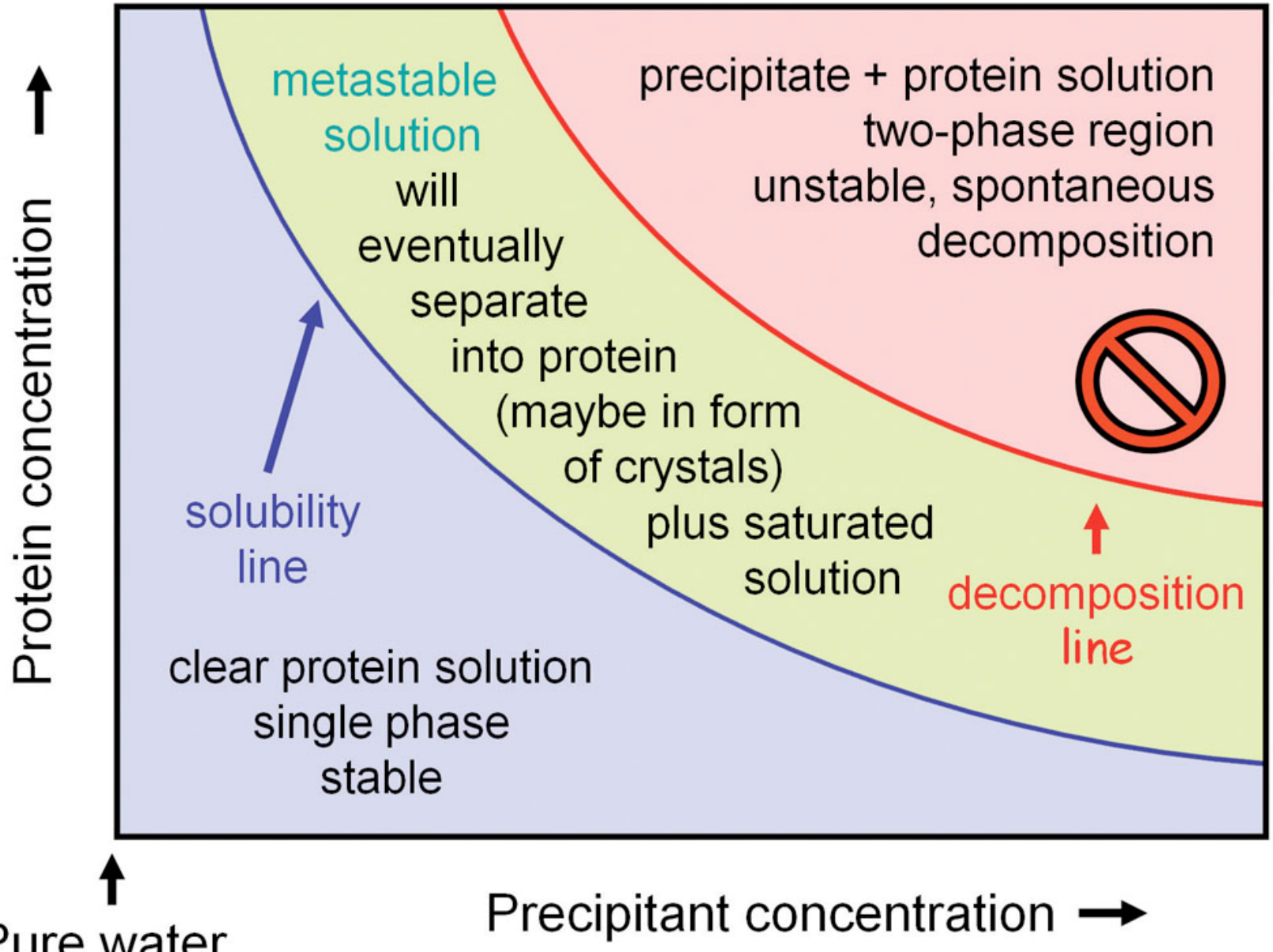
Crystallisation plates – for vapour diffusion



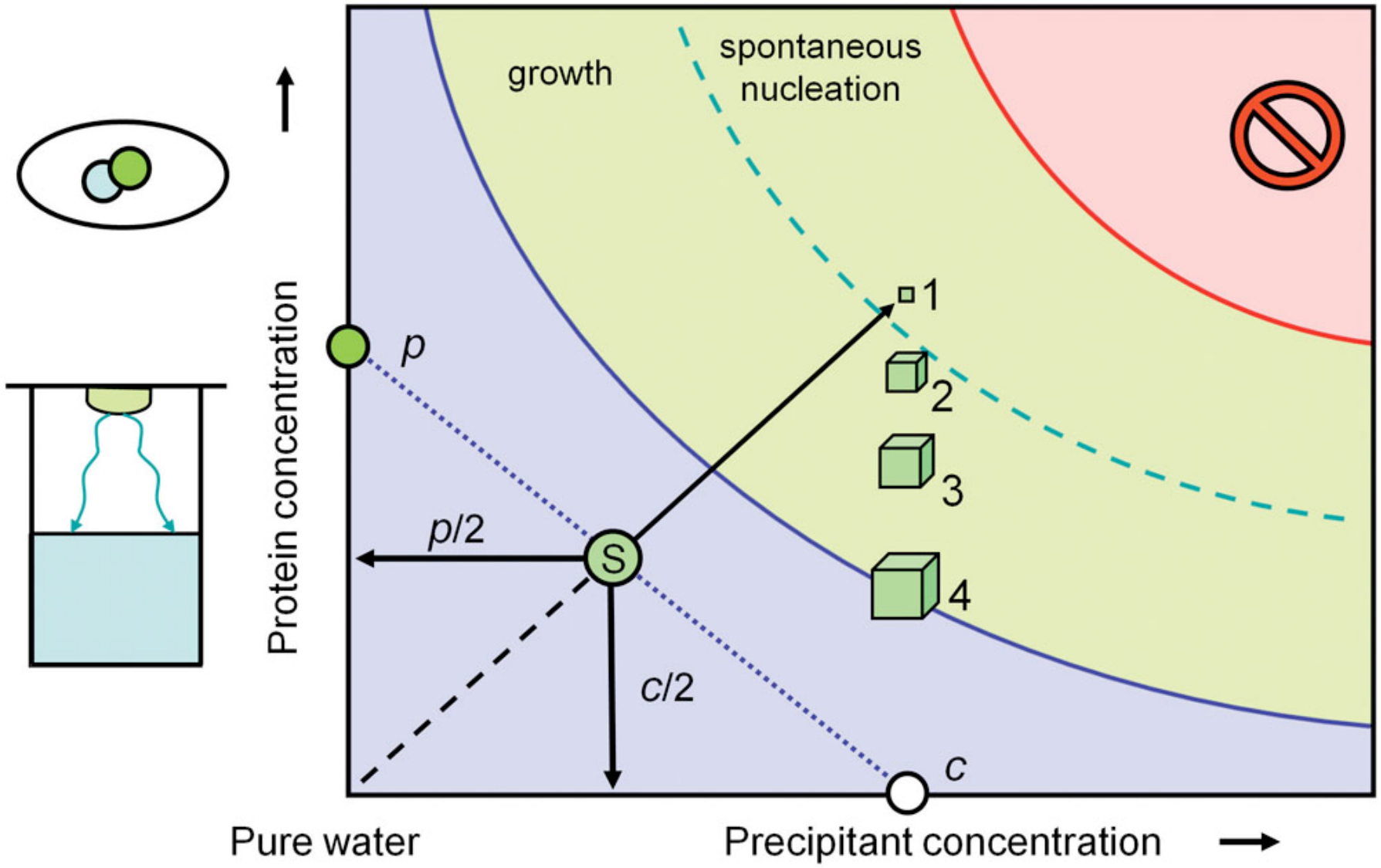
What you are looking for!



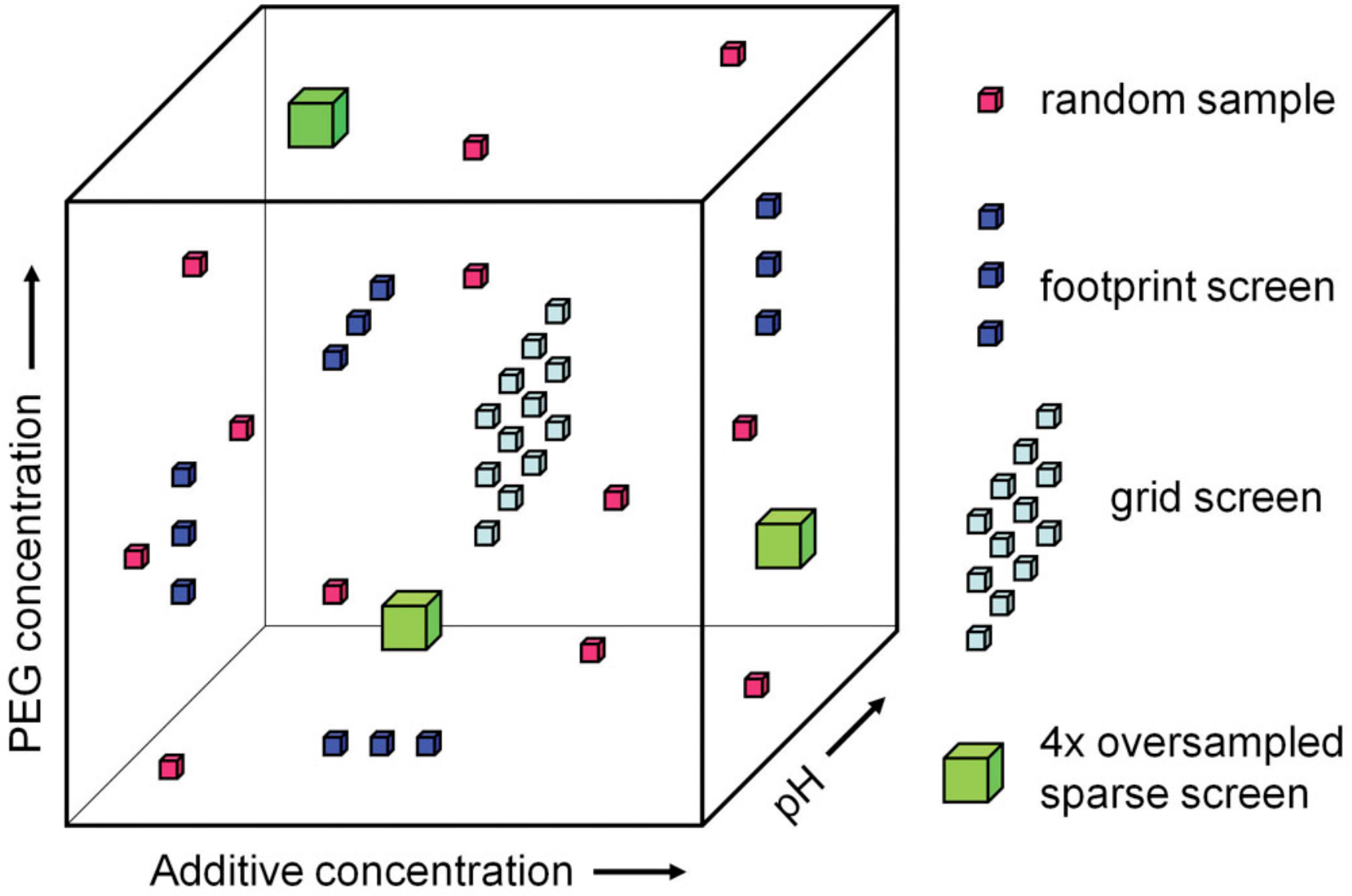
The basic crystallisation diagram



Successful vapour diffusion experiment

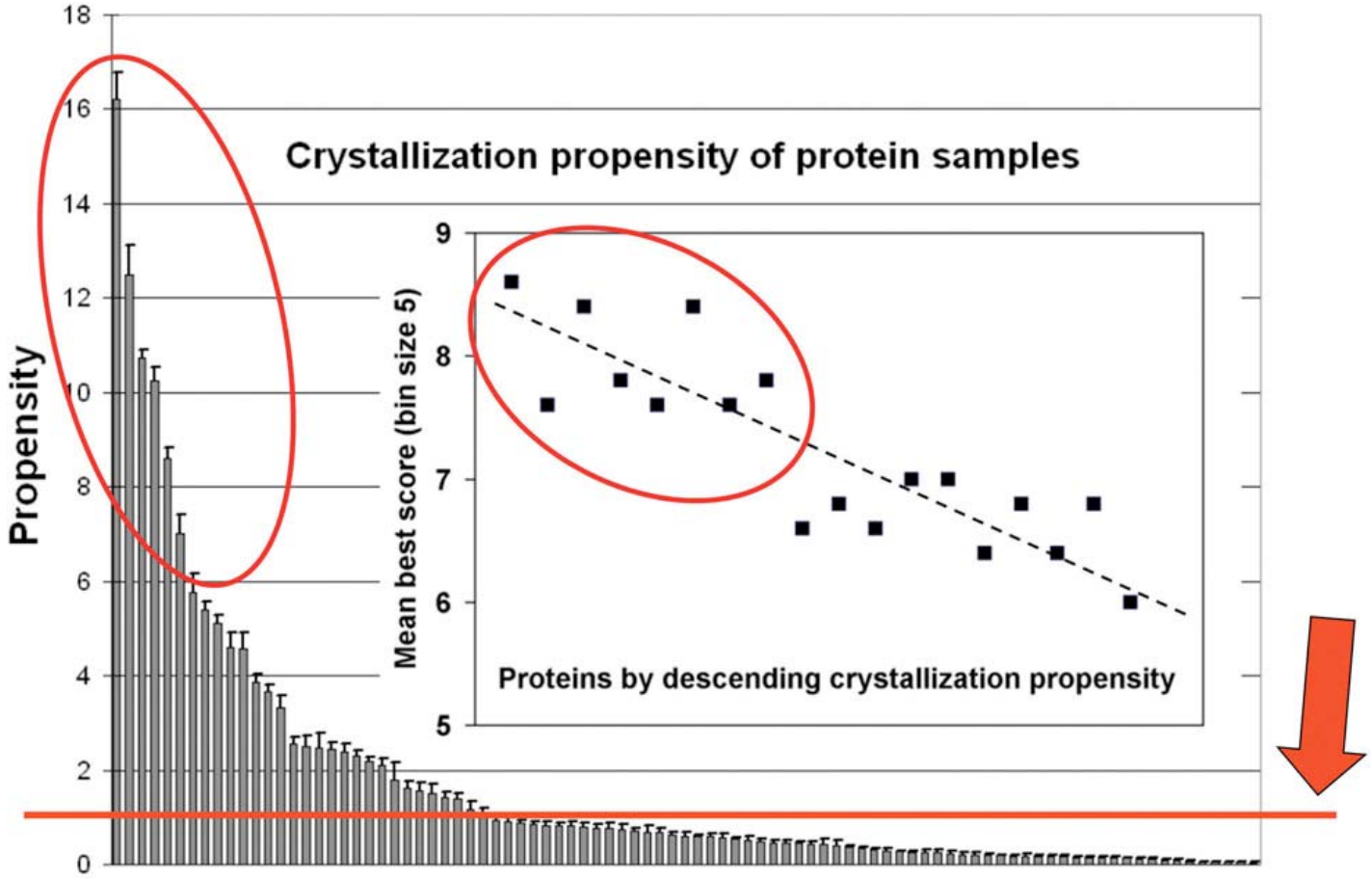


Screening – can't do everything



Can you predict whether a protein will crystallise

The xtalpred server: http://ffas.burnham.org/XtalPred-cgi/xtal.pl



"More efficient to improve the inherent crystallisability than to do further screening" Rupp, 2010

DATA COLLECTION (SOURCES)

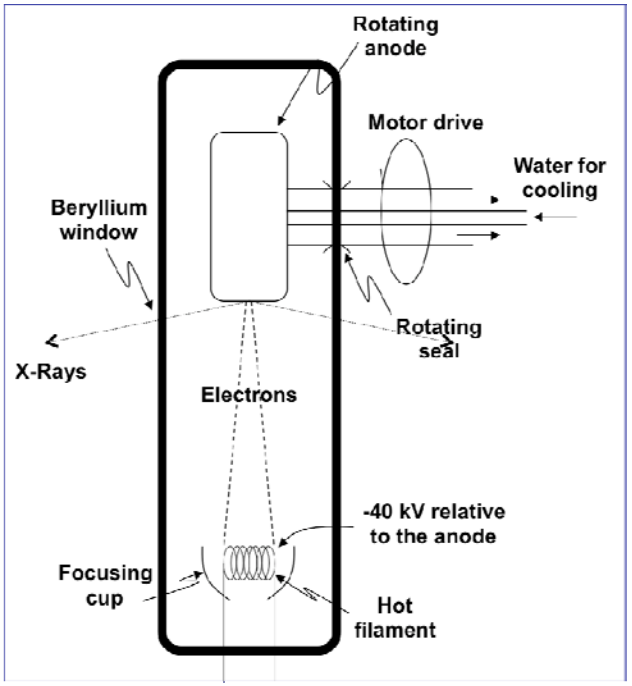
Laboratory X-ray sources

Table 1
Approximate X-ray beam brilliance for the main types of in-house sources with optics.

System	Power (W)	Actual spot on anode (μm)	Apparent spot on anode (μm)	Brilliance (photons s^{-1} mm^{-2} $mrad^{-1}$)
Standard sealed tube	2000	10000 × 1000	1000 × 1000	0.1×10^9
Standard rotating-anode generator	3000	3000 × 300	300 × 300	0.6×10^9
Microfocus sealed tube	50	150 × 30	30 × 30	2.0×10^9
Microfocus rotating-anode generator	1200	700 × 70	70 × 70	6.0×10^9
State-of-the-art microfocus rotating-anode generator	2500	800 × 80	80 × 80	12×10^9
Excillum JXS-D1-200	200	20 × 20	20 × 20	26×10^9

Beam size at the sample 100 - 200μ depending on optics

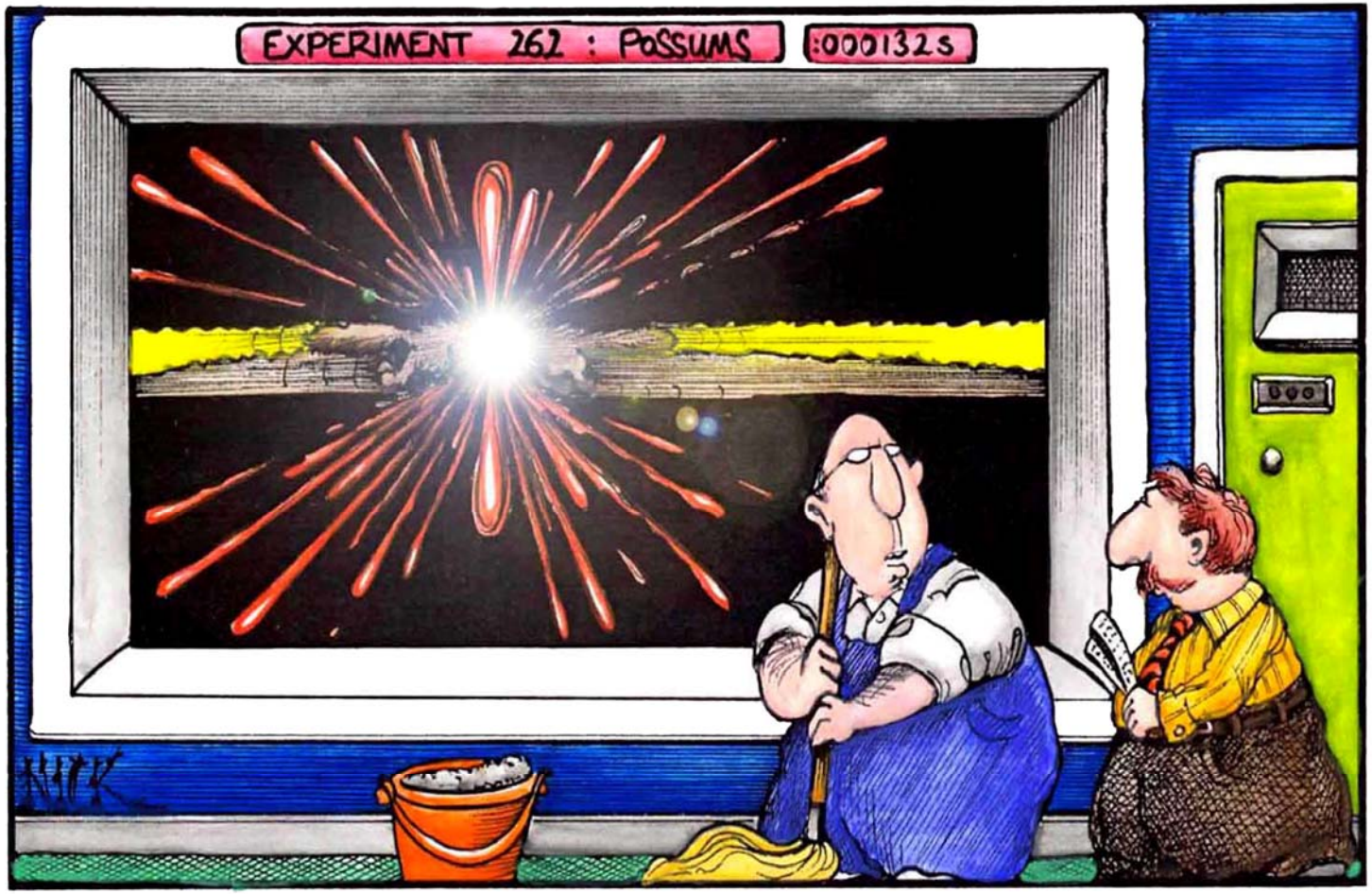
In the 1960s and 1970s typical problems were considered insoluble unless crystals at least 300μ could be grown. Even with modern optics most protein crystals suitable for a laboratory source are > 100μ.



Synchrotron radiation and protein crystallography

- Brilliance, small beam
  - Use of small crystals; rapid data collection
- Low Divergence
  - Resolve close reflections from large unit cells e.g. viruses
- Tunability
  - Enhance anomalous signal for phasing
- Pulsed beam
  - Perform time-resolved experiments on enzymes

Biologists were not always welcome at synchrotrons



"Sure been a heap more work for ME around here since those Biologists got granted research time on the ol' Synchrotron ."

Protein crystallography at the Australian Synchrotron

- Two beamlines
  1) Bending magnet, high throughput, MAD phasing, small molecule crystallography
  2) Undulator, high brilliance, small crystals
- Extensive use of remote access
  - large country, only one synchrotron

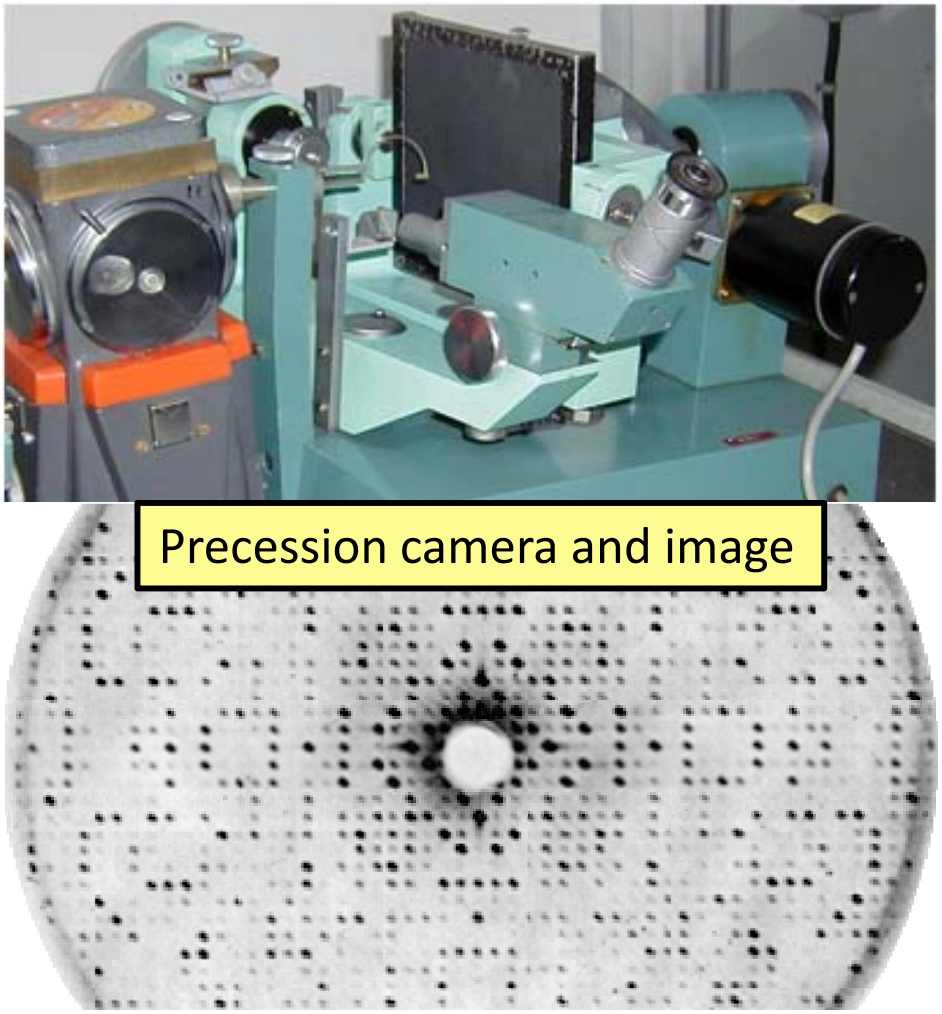
Protein crystallography endstations

- Robot
- CCD detector
  - Would like Dectris!
- Overhead frame
  - Easy access
  - But less stable
- On-line software
  - Processing on-the-fly allows for rapid evaluation of crystals

DATA COLLECTION (DETECTORS)

Detectors - film

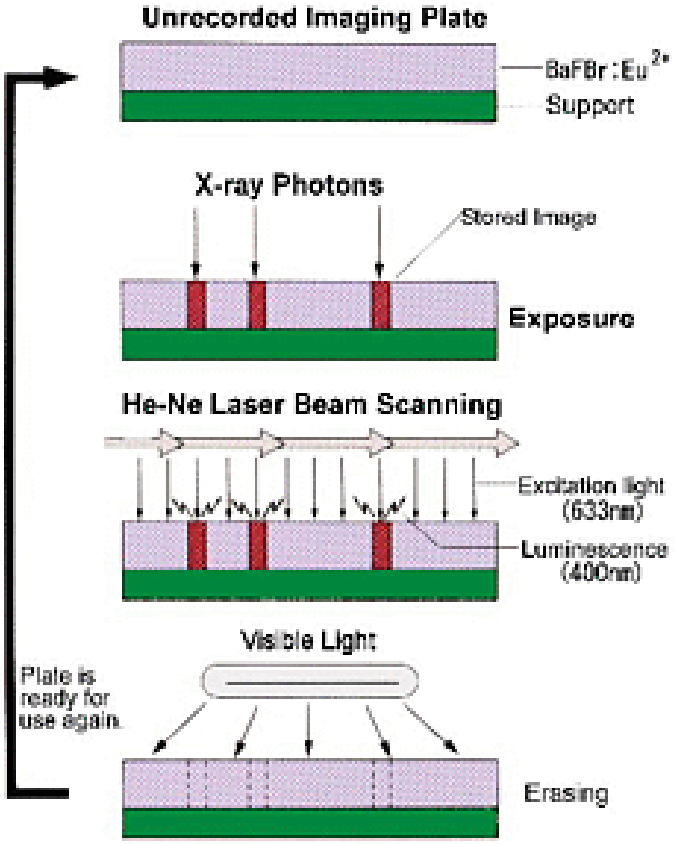
- Film
  - Cheap and linear response – no longer used except for teaching
  - Requires use of darkroom
  - Requires separate scanner to digitize the images
  - Can provide overall view illustrating the crystal symmetry



Precession camera and image

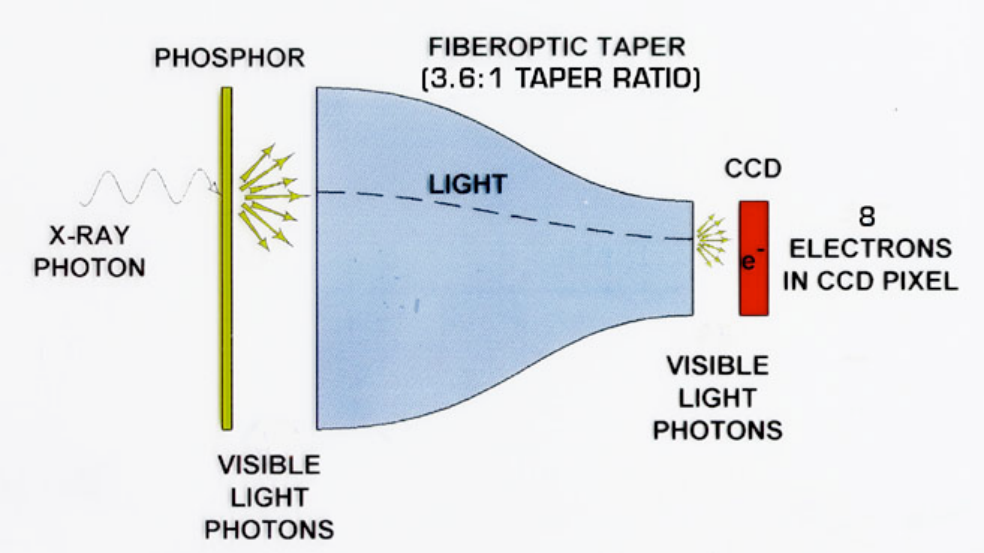
Detectors – image plates

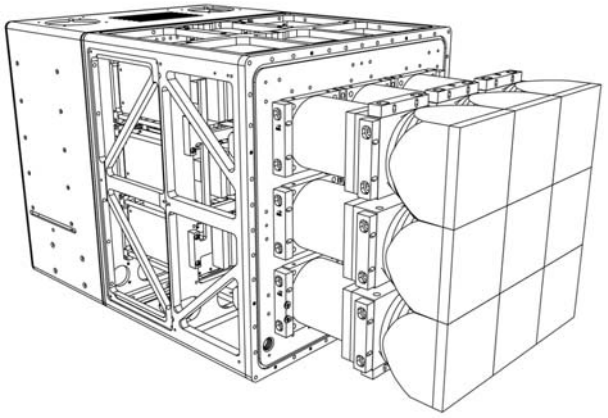
- Image plates
  - Essentially a reusable plastic film
    - Read out by laser that detects excited centres created by X-rays
    - Erased by light
  - Relatively cheap for large area. Wide dynamic range. Low noise.
  - Relatively slow readout (about 120 s).
    - Restricts usefulness at a synchrotron where an exposure is about 1s.
    - Still used at home sources where exposures are 1-10 minutes



Detectors - CCD

- CCD
  - CCD. Need to be tiled for large area. Requires cooling to reduce noise. Relatively fast readout (< 1s). Need to demagnify the image to match CCD size.
  - Relatively expensive. Need careful calibration.
  - Physically large

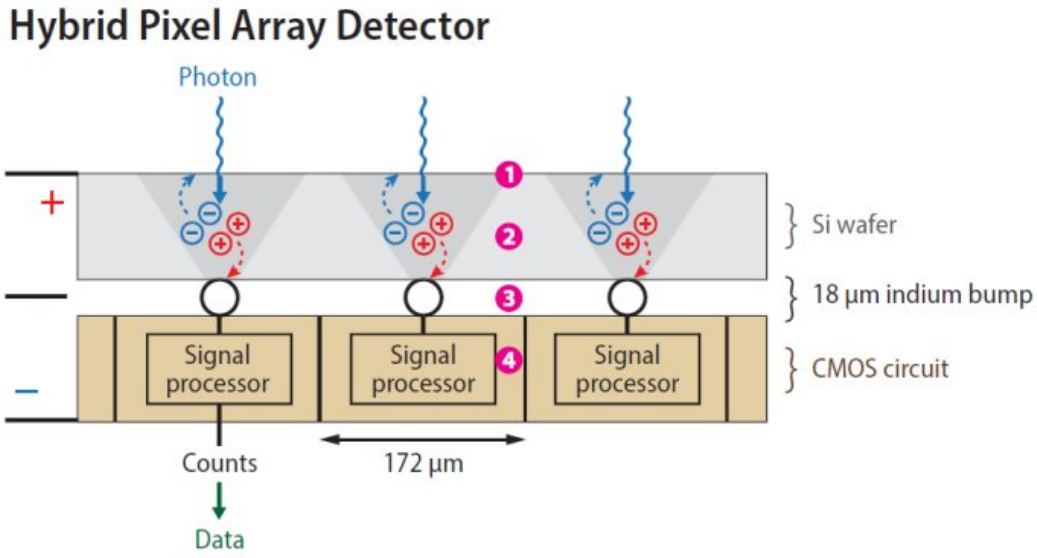


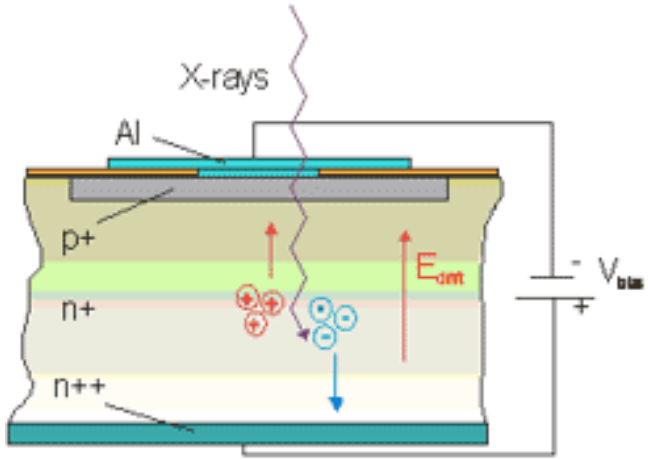


ADSC CCD detector

Detectors – pixel arrays

- Pixel arrays
  - Can tile many modules.
  - Virtually zero noise.
  - Operates at room temperature.
  - Ultrafast readout permits continuous exposure.
  - Very expensive – early adoption of new technology.





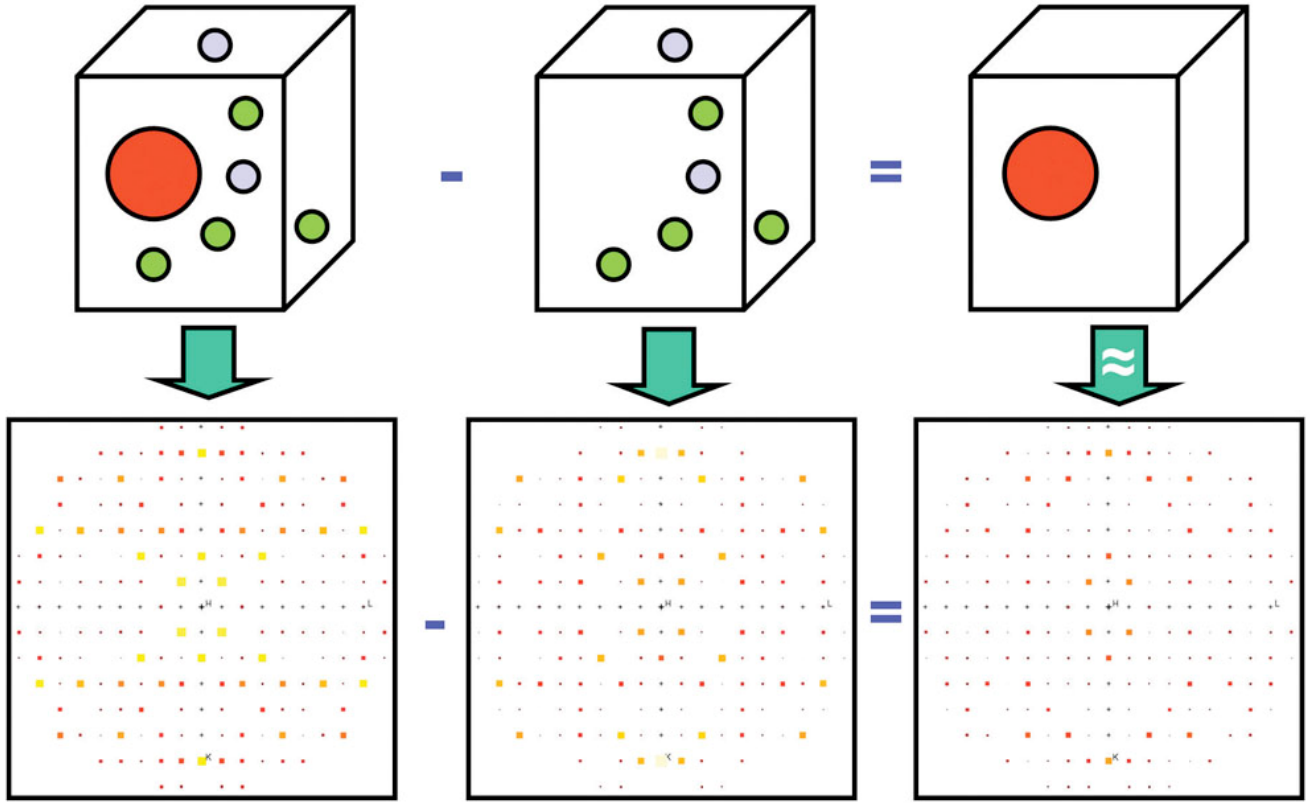
STRUCTURE SOLUTION

Solving the phase problem

- **Marker atom substructure methods** do not depend on prior structural knowledge about the protein (other than the sequence and the structures of the amino acids). Once the marker atoms (heavy metals, selenium) have been located by Patterson or direct methods starting phases for the modelling can be calculated. These are *de novo* or experimental phasing methods.
- **Density modification** are powerful methods for *improving* the initial phases and are used in practically all *de novo* phasing experiments. These methods included solvent flattening, and non-crystallographic symmetry averaging. As they do not require any model information they may also be considered experimental techniques.
- **Molecular replacement** require a similar structure as a molecular search probe. *Replacement* is to be understood as the placing of the search model in the crystal structure not as "substitution". Phase bias can be a serious problem with this technique.
- **Direct methods** exploit the fact that relationships exist between certain sets of structure factors. They require very high resolution (1.2 Å or better) for *ab initio* determination of a protein structure and have been limited to very small proteins. They are commonly used for small molecule structures. They are, however, very important in sub-structure (marker atom) determination.

The concept of isomorphous difference

If we add an atom to a protein crystal and change nothing else (isomorphous) then the difference in the diffraction patterns of the "derivative" structure with added heavy atom and the "native" protein structure approximates the diffraction pattern from the heavy atom alone.



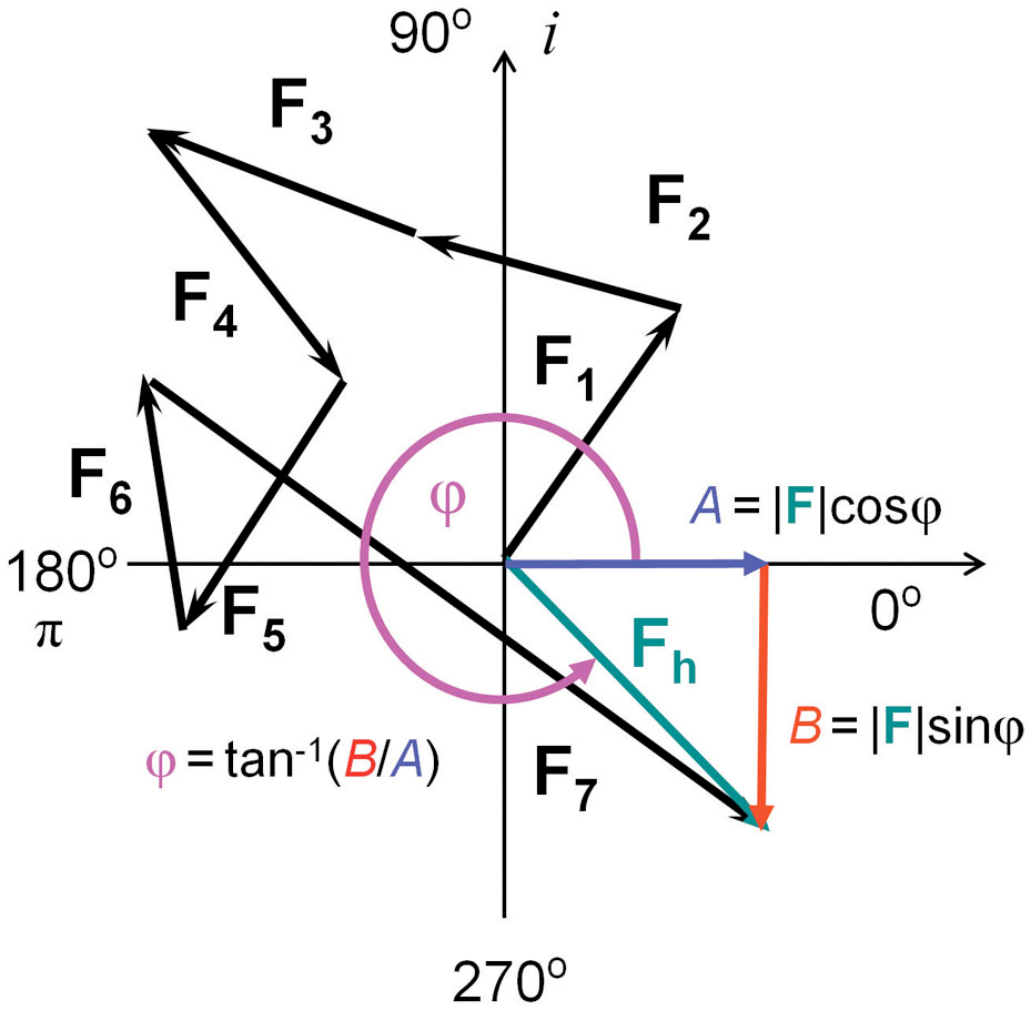
$\mathbf{F}_{PA} = \mathbf{F}_P + \mathbf{F}_A$ and thus

$\mathbf{F}_A = \mathbf{F}_P - \mathbf{F}_A$ but these are vectors and the amplitudes (things we can measure) don't add up: $F_A \neq F_{PA} - F_P$ but we can still use the differences to solve the sub-structure.

Revision – graphical construction of the structure factor

Graphical illustration of the summation that yields one structure factor for a 7-atom structure – with 6 relatively light atoms and one heavier atom. The resultant structure factor is $\mathbf{F}_h$.

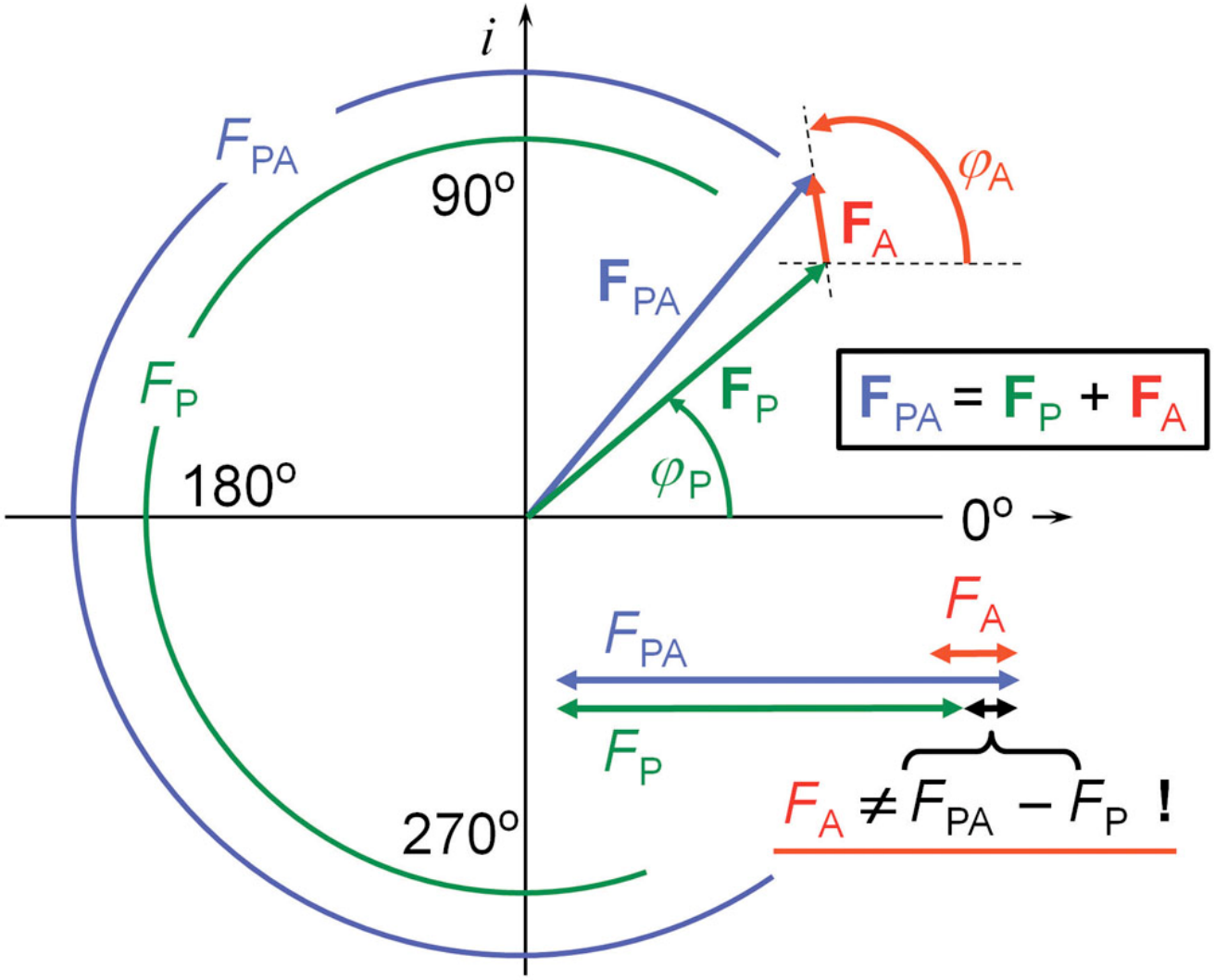
$$F_h = \sum_{j=1}^{atoms} f_{hj}^{0}.\exp(2\pi i h x_j)$$



The “true” isomorphous differences

When computing the total structure factor the contributions from the individual atoms are summed *vectorially.* Thus, if we add one more atom to a protein we simply add its structure factor to that of the total for the protein.

The radii of the circles are the amplitudes for the protein and heavy-atom derivative structure factors.

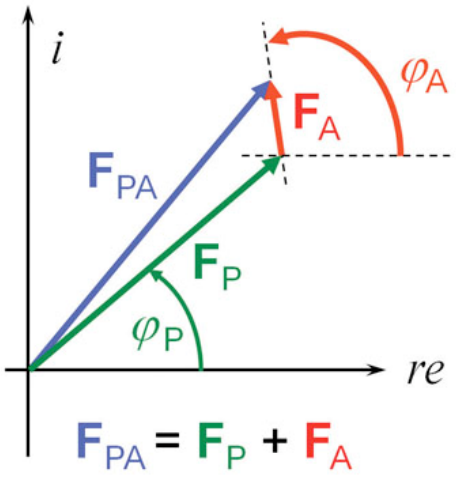


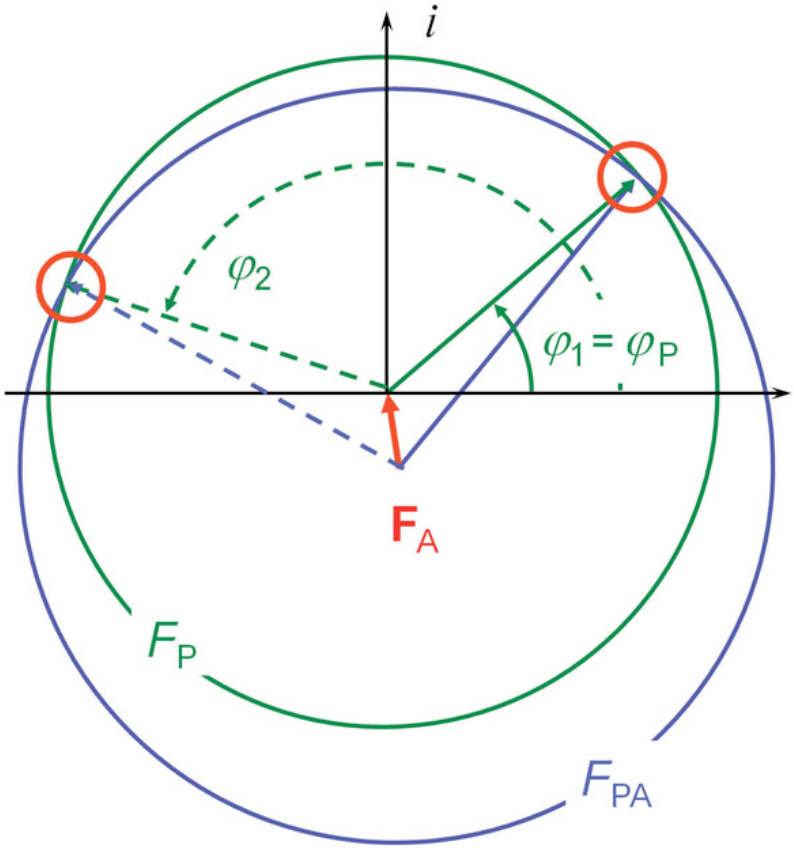
Graphical illustration of phase determination

Once we know where the heavy atom is we can calculate $\mathbf{F}_A$ – the amplitude AND phase of its contribution.

For each reflection we can make the construct closing the vector triangle. We know the amplitudes F_P and F_{PA} from the experimental measurements and the total contribution $\mathbf{F}_A$.

If we draw a circle radius, F_{PA}, at the origin and then move along $\mathbf{F}_A$ and draw a circle radius, F_P, there should be TWO points of intersection that satisfy the phase addition giving two solutions for ϕ_P.





Resolving the phase ambiguity

- We have a phase ambiguity if there is: a single isomorphous derivative (SIR) or anomalous scattering from the native crystal (e.g. from sulfur) (SAD).

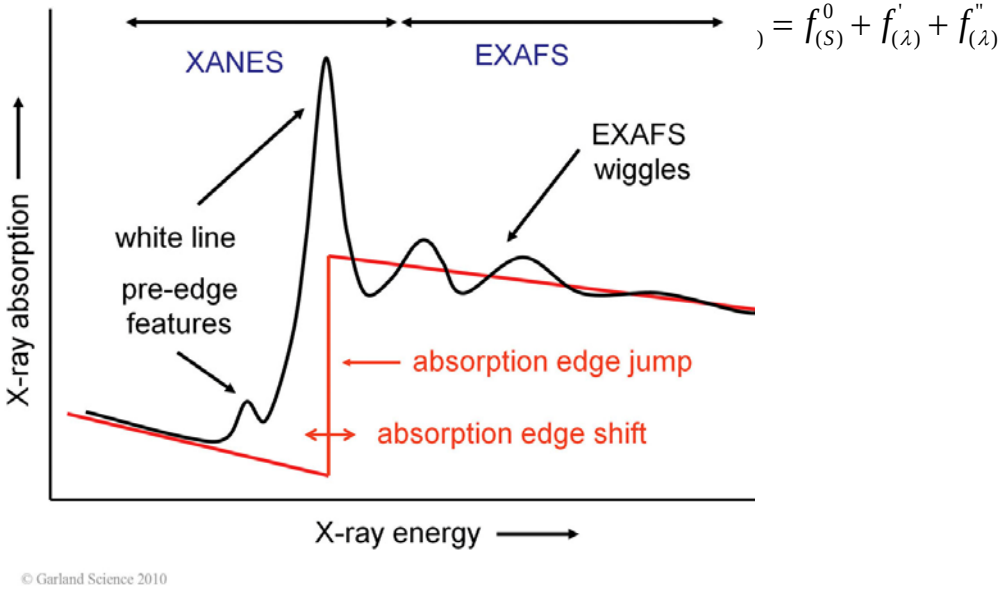
This problem can be resolved:

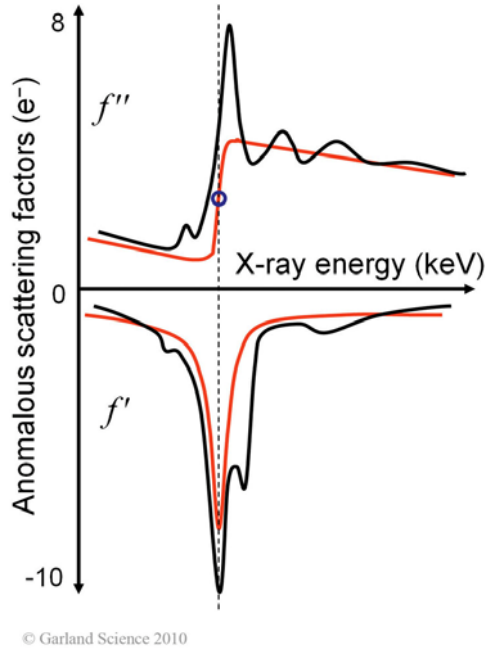
- By adding more derivatives – multiple isomorphous replacement (MIR)
- By using the anomalous signal – single isomorphous replacement with anomalous scattering (SIRAS) or multiple isomorphous replacement with anomalous scattering (MIRAS)
- By using the dispersive and anomalous signal together – multiple wavelength anomalous diffraction (MAD)
- With the help of density modification
- With dedicated direct methods programs

Anomalous scattering

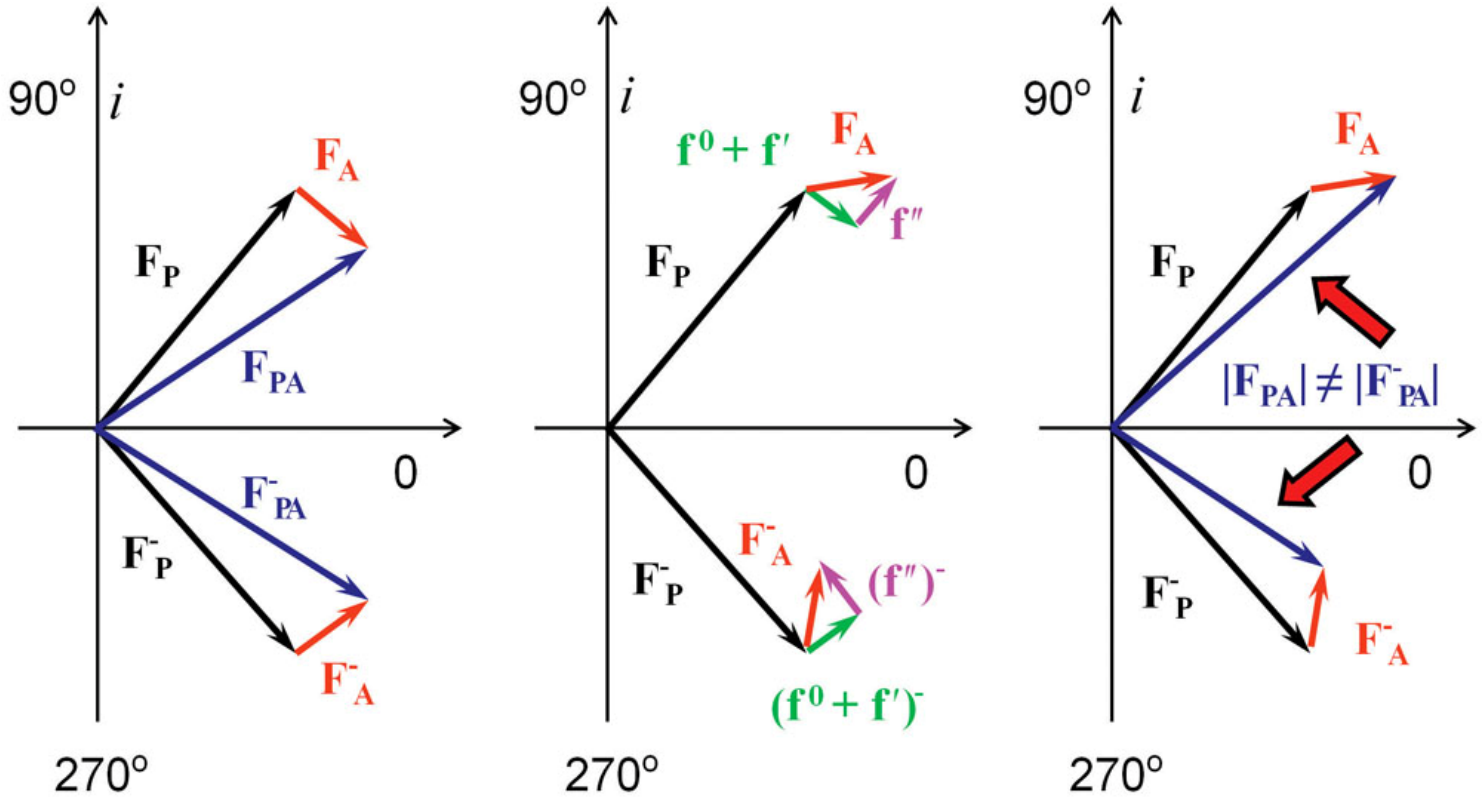
$$f_{(S,\lambda)} = f^{0}_{(S)} + f^{'}_{(\lambda)} + f^{''}_{(\lambda)}$$

The scattering from an atom actually contains two wavelength (energy) dependent terms. These vary rapidly near an absorption edge.



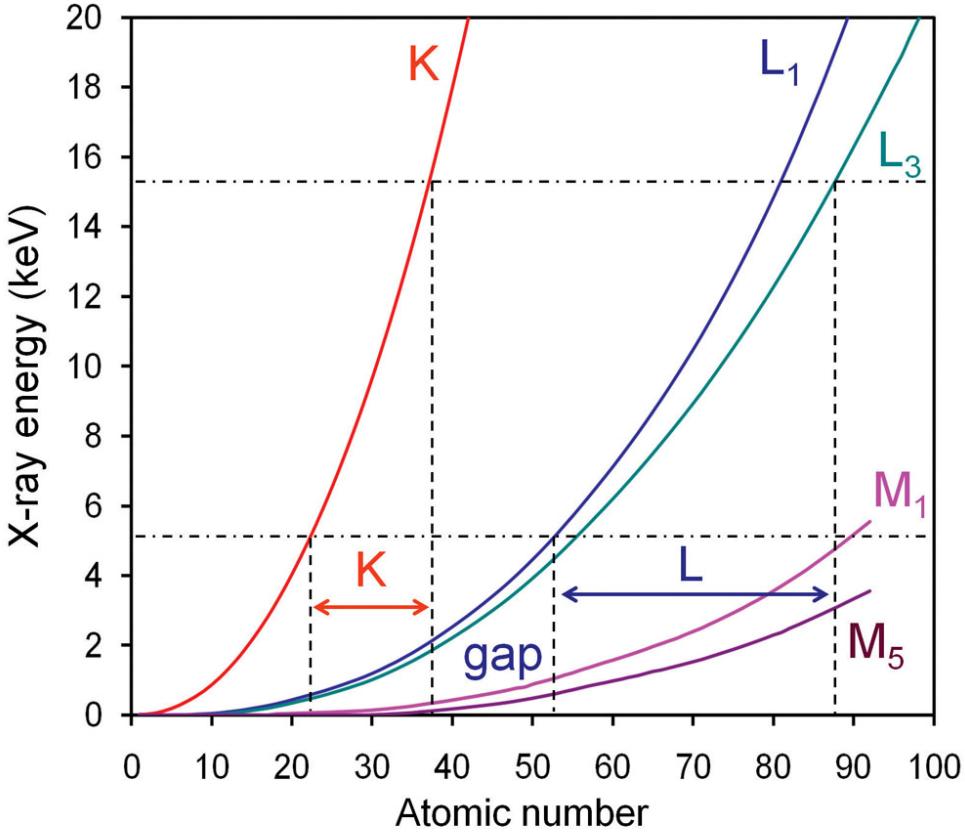


Anomalous scattering for phasing



Anomalous scattering only significant for heavier elements in a protein

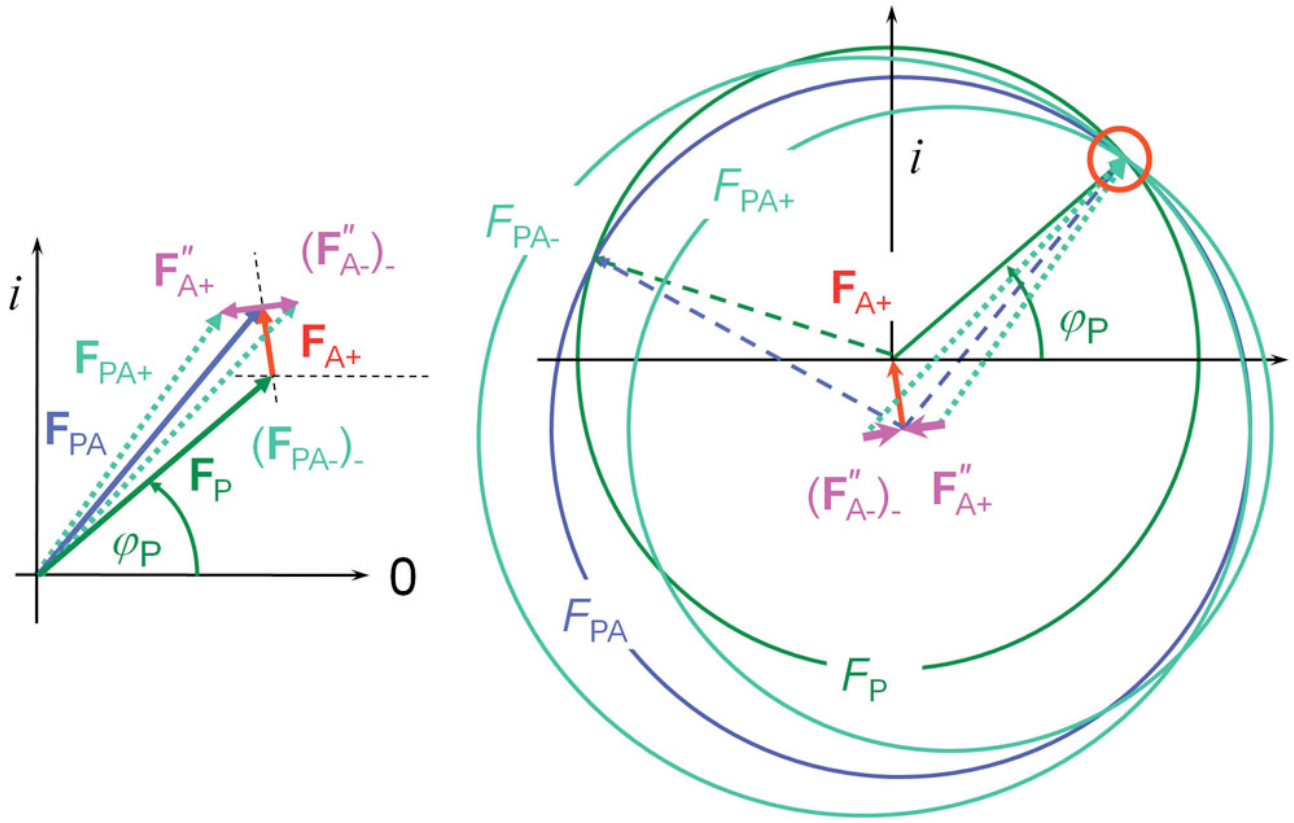
- At most synchrotrons PX beamlines one can access the L-edges for elements form iodine to uranium and the K-edges for transition metals
- Most importantly the edge for selenium is at 0.979 Å (1.27 KeV)
  - Se can be introduced into recombinantly expressed proteins replacing the sulfur in the amino acid methionine



Single Isomorphous Replacement with Anomalous Scattering (SIRAS)

In this case for each reflection we have 1 observation of the native amplitude, F_P, and two for the derivative, F_{PA-} and F_{PA-}.

In the absence of anomalous scattering (blue circle) there are two solutions with anomalous scattering (light green circles) there is only one.



Summary of structure solution methods

Phasing method	Phasing marker	Remarks	Available data	Marker	Difference data	Data to be phased
SAD via sulfur atoms (S-SAD)	S in Met, Cys residues	Highly redundant data collection, must be combined with density modification	Anomalous pairs F_{PS}^{+}, F_{PS}^{-} also serve as native data	Sulfur positions, $\mathbf{F}_S$	ΔF_{ano} from $\lvert F_{PS}^{+} - F_{PS}^{-}\rvert$	Merged F_{PS}
SAD via naturally bound metals	Naturally bound metal ion, cofactor	Requires density modification for resolution of phase ambiguity	Anomalous pairs F_{PA}^{+}, F_{PA}^{-} also serve as native F_P	Anomalous scatterer positions, $\mathbf{F}_A$	ΔF_{ano} from $\lvert F_{PA}^{+} - F_{PA}^{-}\rvert$	Merged F_{PA}
SIR(AS) via isomorphous metals	Heavy atom ion, specifically bound anions Br^-, I^-, I_3^-, also Xe	Isomorphous phasing power proportional to $z_{(H)}$, anomalous signal or density modification needed to break phase ambiguity	Native data F_P, isomorphous data F_{PA}, in pairs F_{PA}^{+}, F_{PA}^{-} for SIRAS	Isomorphous/ anomalous scatterer positions, $\mathbf{F}_A$	ΔF_{iso} from $\lvert F_{PA} - F_P\rvert$ and ΔF_{ano} from $\lvert F_{PA}^{+} - F_{PA}^{-}\rvert$	Native F_P
MIR(AS) via isomorphous metals	Heavy atom ions, clusters, specifically bound anions Br^-, I^-, I_3^-	As above, except multiple derivatives or anomalous signal break phase ambiguity. Hg, Pt, Au, etc. phase several hundred residues, heavy atom clusters more.	Native data F_P, isomorphous data $n\cdot(F_{PA})$ in pairs $n\cdot(F_{PA}^{+}, F_{PA}^{-})$ for MIRAS	Isomorphous/ anomalous scatterer positions, $n\cdot(\mathbf{F}_A)$	$n\cdot\Delta F_{iso}$ from $n\cdot\lvert F_{PA} - F_P\rvert$, and ΔF_{ano} from $\lvert F_{PA}^{+} - F_{PA}^{-}\rvert$ pairs	Native F_P
MAD via Se	Se in Se-Met residues	1 Se phases 100–200 residues, introduced by expression host	Bijvoet pairs at n wavelengths, $n\cdot(F_{PA}^{+}, F_{PA}^{-})_{\lambda n}$, optional native data	Anomalous scatterer positions $\mathbf{F}_{Se}$	ΔF_{ano} from $\lvert F_{PA}^{+} - F_{PA}^{-}\rvert_{\lambda}$ pairs ΔF_{λ} from $\lvert F_{\lambda i} - F_{\lambda j}\rvert$ pairs	Best merged data F_{PA}, optional native F_P
MAD via isomorphous metals	Heavy atom, specifically bound	Strong signal on XAS "white lines," particularly at L-edges, can phase several hundred residues	Bijvoet pairs at n wavelengths $n\cdot(F_{PA}^{+}, F_{PA}^{-})_{\lambda n}$, native data F_P (not needed for phasing)	Anomalous scatterer positions $\mathbf{F}_A$	ΔF_{ano} from $\lvert F_{PA}^{+} - F_{PA}^{-}\rvert_{\lambda}$ pairs ΔF_{λ} from $\lvert F_{\lambda i} - F_{\lambda j}\rvert$ pairs	Native F_P, optional best merged F_{PA}
Direct methods	None	Near atomic resolution data (1.2 Å or better), relatively small proteins	F_P	All non-H atom positions *ab initio*	N/A	Native F_P
Density modification	None	Needs multiple copies of motif in asymmetric unit for *ab initio* phasing	F_P	Multiple copies of a subunit	N/A	Native F_P
MR via model structure	Positioned search model	Needs search model with structural similarity, subject to model bias, particularly at low resolution	Native data F_P and search model structure factors F_C	Entire model serves as search probe	N/A	Native F_P