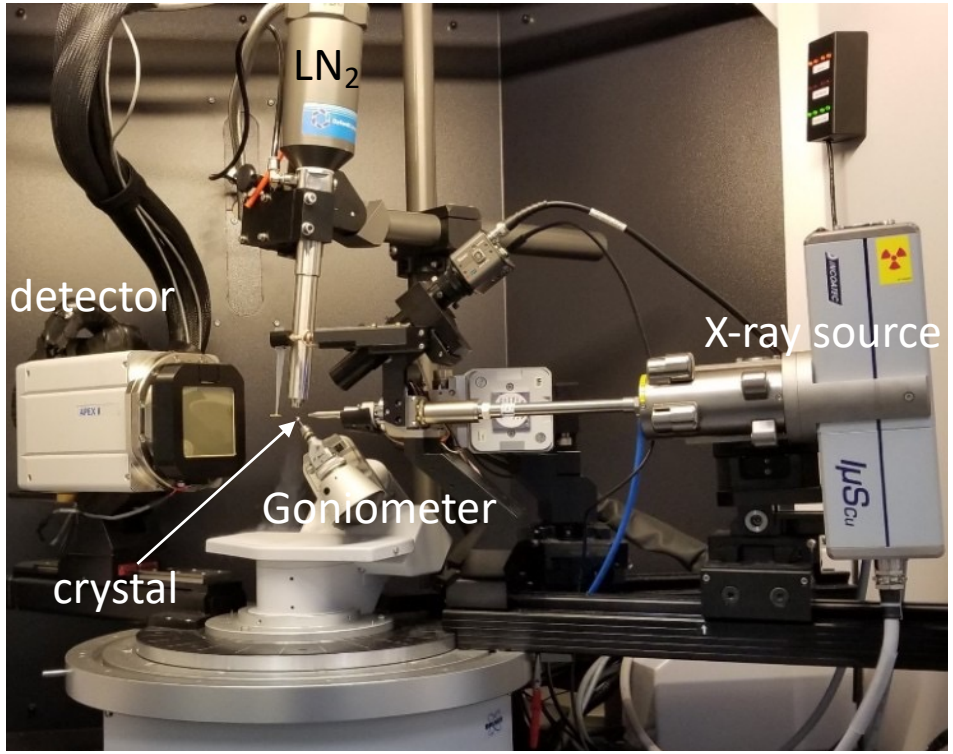
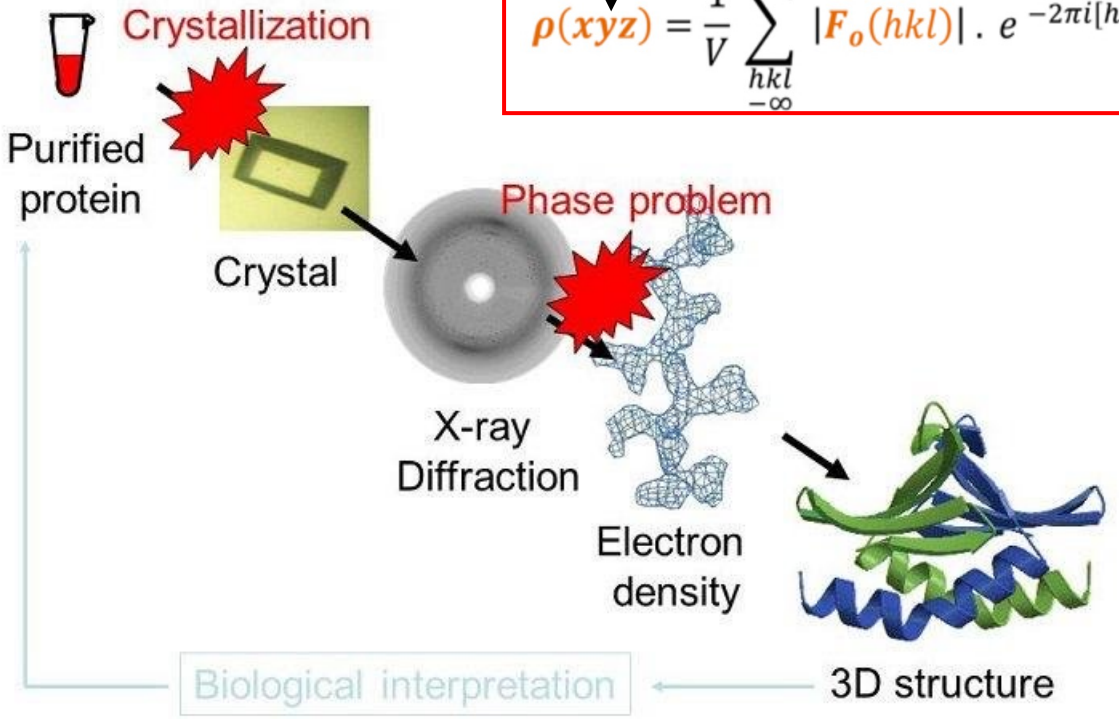
LECTURE 9-2: X-ray Crystallography



Electron density Magnitude Phase

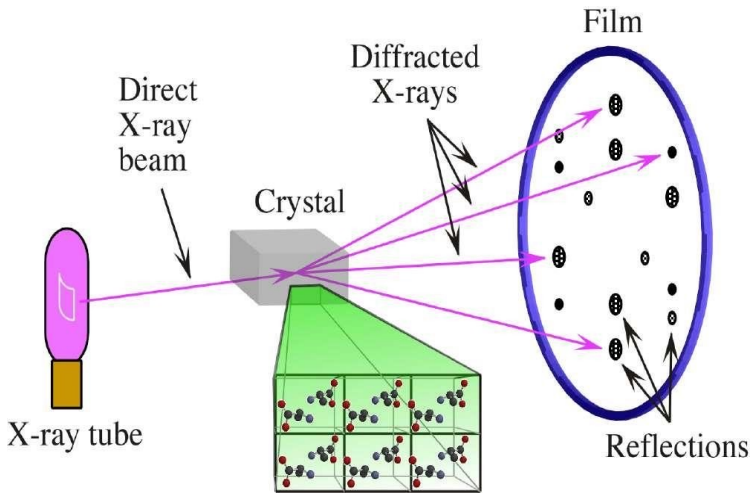
$$\rho(xyz) = \frac{1}{V}\sum_{hkl=-\infty}^{+\infty} |F_o(hkl)| \,.\, e^{-2\pi i[hx+ky+lz-\phi_c(hkl)]}$$



https://ts1.cn.mm.bing.net/th/id/R-C.63a6ede7a118f0c532cf792a9b1a542d?rik=6IS3MWfcDuhLgQ&riu=http%3a%2f%2fwww.bc.edu%2fcontent%2fbc-web%2fresearch%2fsites%2fvice-provost-for-research%2fabout%2fresearch-facilities%2fxray-crystal%2f_jcr_content%2fpar%2fbc_image_content%2fimage.img.jpg%2f1563543458353.jpg&ehk=Axzd9G08ok1yI%2betzQgPp3VK2CHZIuNo4SIGk0GLYUo%3d&risl=&pid=ImgRaw&r=0

https://www.researchgate.net/profile/Athanasios-Valavanidis/publication/339253934/figure/fig2/AS:858407774978048@1581671770050/X-ray-crystallography-diffraction-pattern-and-protein-structure-Between-the-atomic.jpg

X-ray diffraction of a single protein crystal: Review



- The wavelength of X-ray is in the same size regime as covalently bonded atoms (~1.5 Å).
- There are two fundamental reasons why we can't simply take a picture of a protein:
  1. X-rays(0.08-0.6 nm) cannot be focused by lenses
  2. A single protein molecule is a very weak scatterer

- The solutions to these problems are:

1. Arrange the molecules into a 3D repeating array – a crystal.
   - Protein molecules are arranged in precisely the same orientation, which amplifies x-rays scattering. In addition, the radiation scattered from a repeating array can cause constructive interference – **diffraction**.
2. Use a "mathematical lens" – the Fourier Transform – to reform an image of the protein molecule.

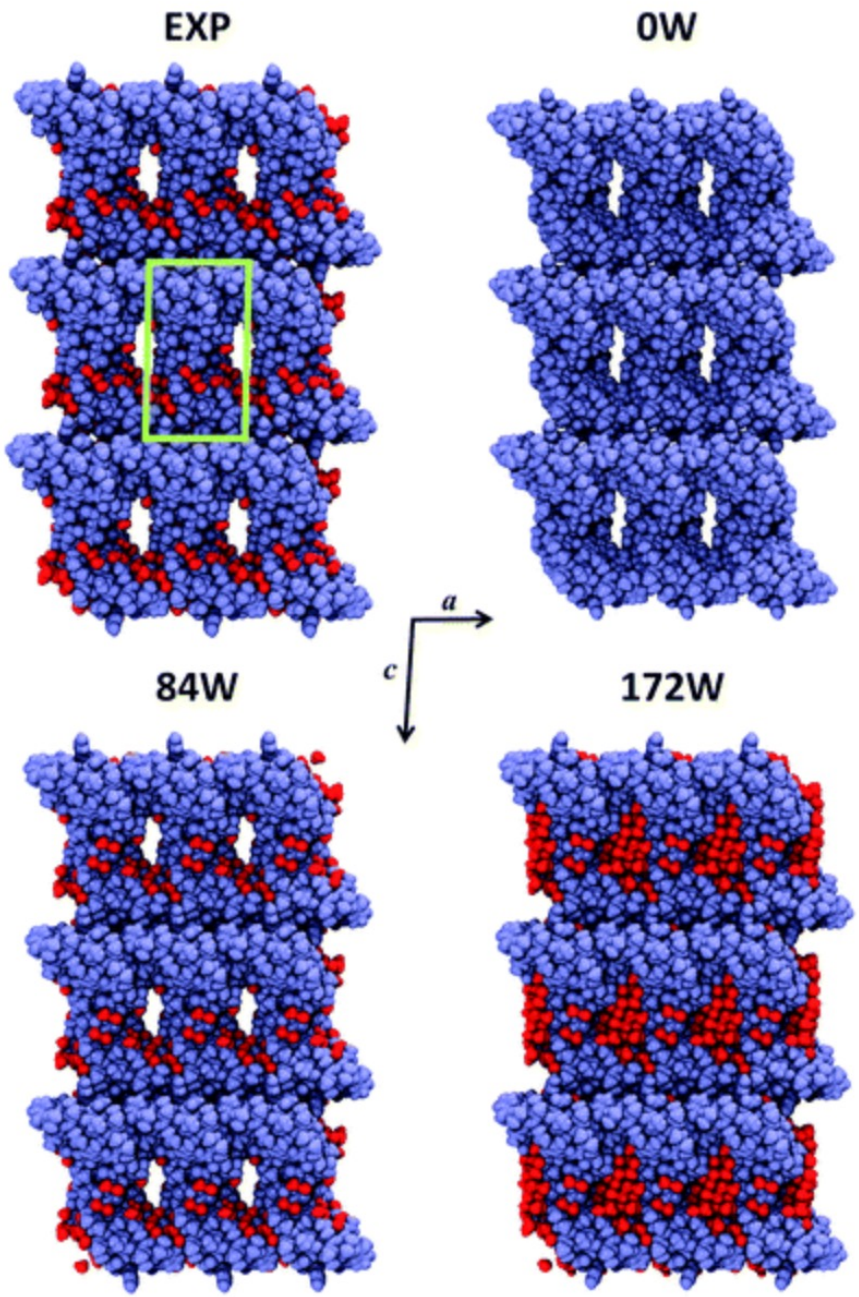
Protein Crystals: Review

- An ordered array of protein molecules making specific intermolecular contacts to form a repeating, 3-dimensional lattice.

- Crystals are stacks of bricks called **unit cell**.

- Protein crystals contain ~40-60% solvent

- Protein molecules are so big but the number of interactions between them is relatively small

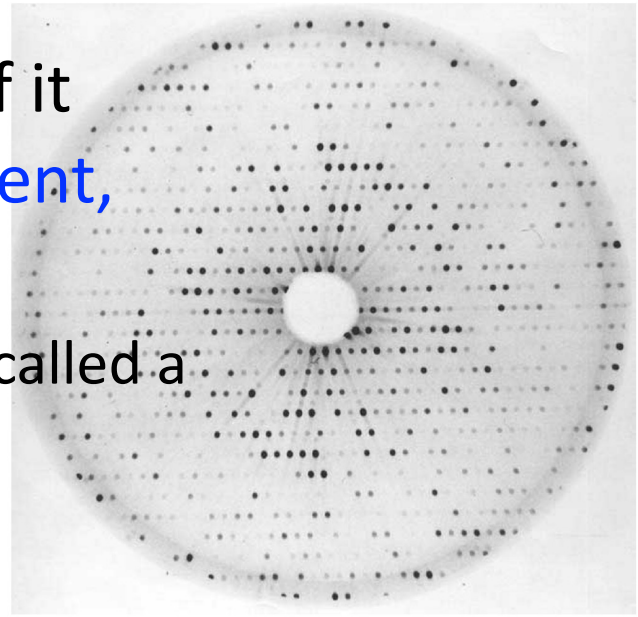


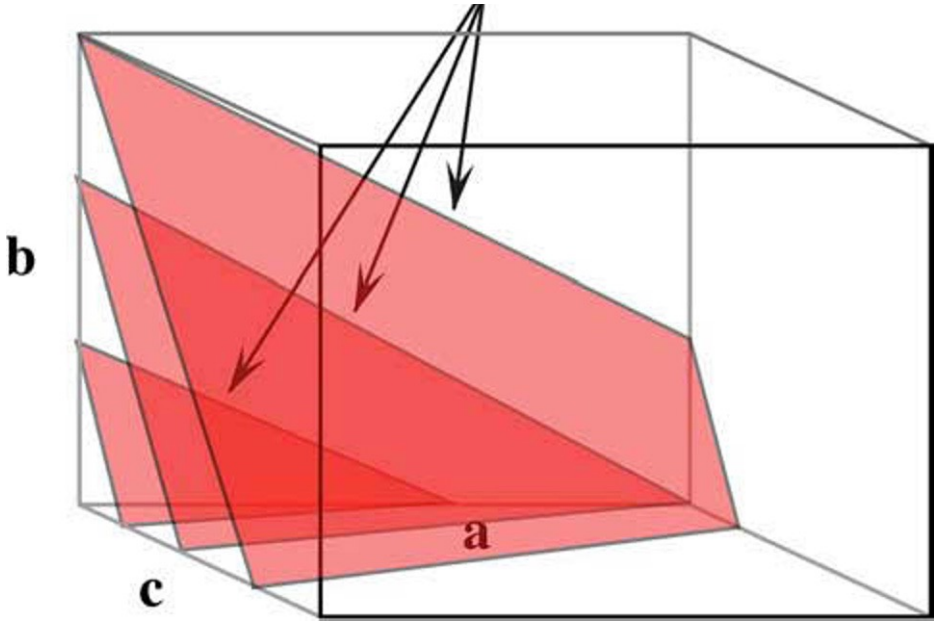
Crystal portion views of crambin.

Delle Piane, M., Corno, M., Orlando, R., Dovesi, R., & Ugliengo, P. (2016). Elucidating the fundamental forces in protein crystal formation: the case of crambin. *Chemical science*, *7*(2), 1496–1507.

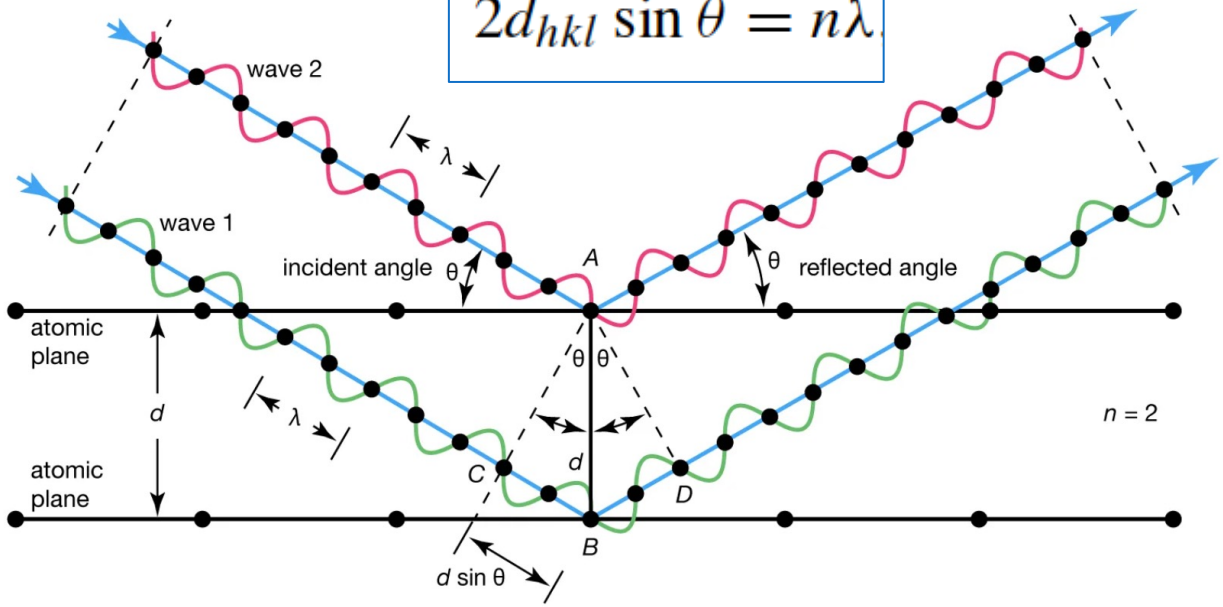
Bragg's Law: Review

- In crystals, diffraction is treated as if it were reflection from sets of equivalent, parallel planes of atoms in a crystal.
  - Each spot in the diffraction pattern is called a *reflection*.



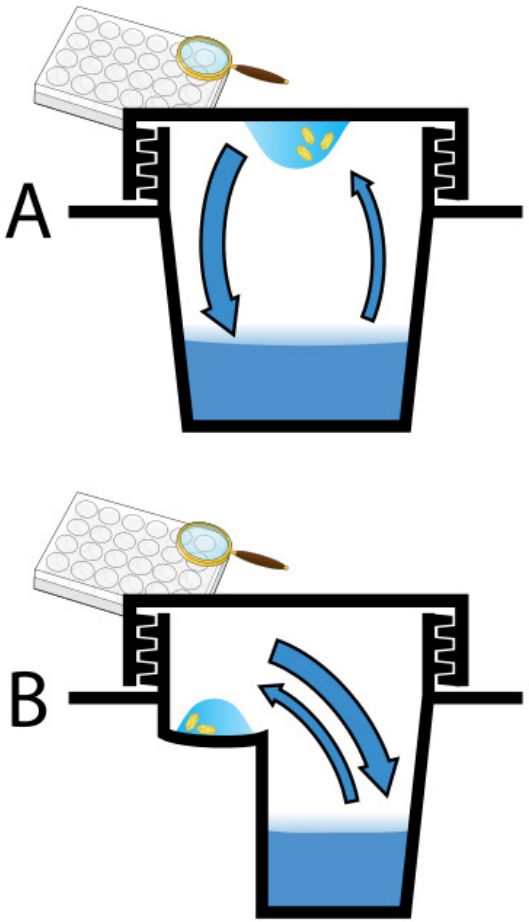


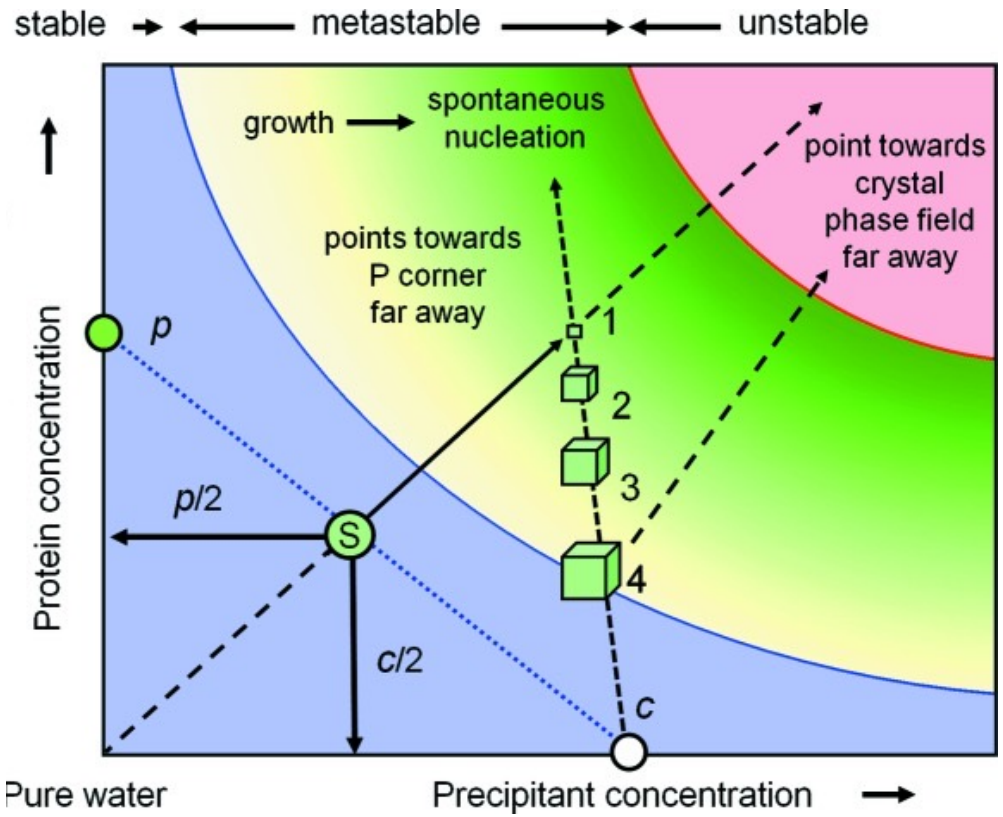
$$2d_{hkl} \sin\theta = n\lambda$$



Protein Phase Diagram: Review

Vapor diffusion





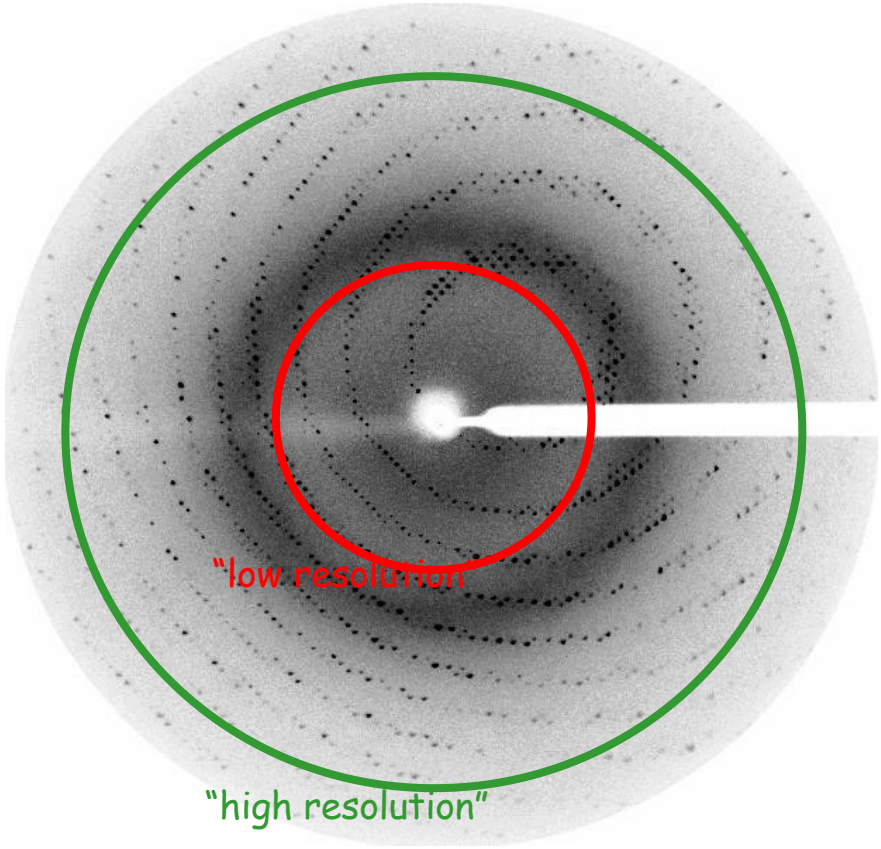
- As vapor diffusion proceeds, both the protein and precipitant concentrate until both reach their stock value, which induces the protein nucleate spontaneously.

Rupp B. Acta crystallographica. Section F, Structural Biology Communications. 2015 Mar;71(Pt 3):247-260. DOI: 10.1107/s2053230x1500374x. PMID: 25760697; PMCID: PMC4356298.

Data Collection: Review

- Two Pieces of Data
  - The position of a reflection point on the reciprocal lattice, given by coordinates *h, k, l*. Determined by the direction reflected.
  - The intensity of the reflection.

- The degree of order in the crystal determines the quality of the diffraction data and ultimately the quality of the final atomic model.
  - Mosaic spread: misalignment, lower S/N

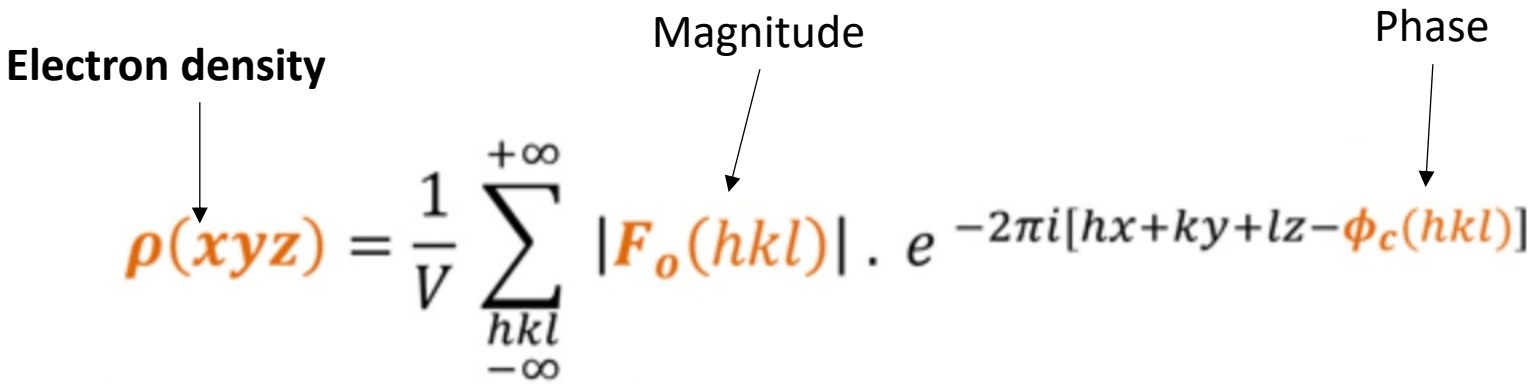


Data Processing: Review

Three Phases:

1. **Indexing** – determining the unit cell dimensions and symmetry, as well as the orientation of the crystal with respect to the beam
2. **Integration** – Integrate the intensities for each reflection.
   - ❖The product of this step is a file with Miller indices (h. k, l) and intensity for each measured reflection.
3. **Scaling** – Scaling attempts to accounts for error in the intensity measurements.
   - ❖The product of this step is a file with **unique** Miller indices (h. k, l), intensity, and error for each measured reflection.

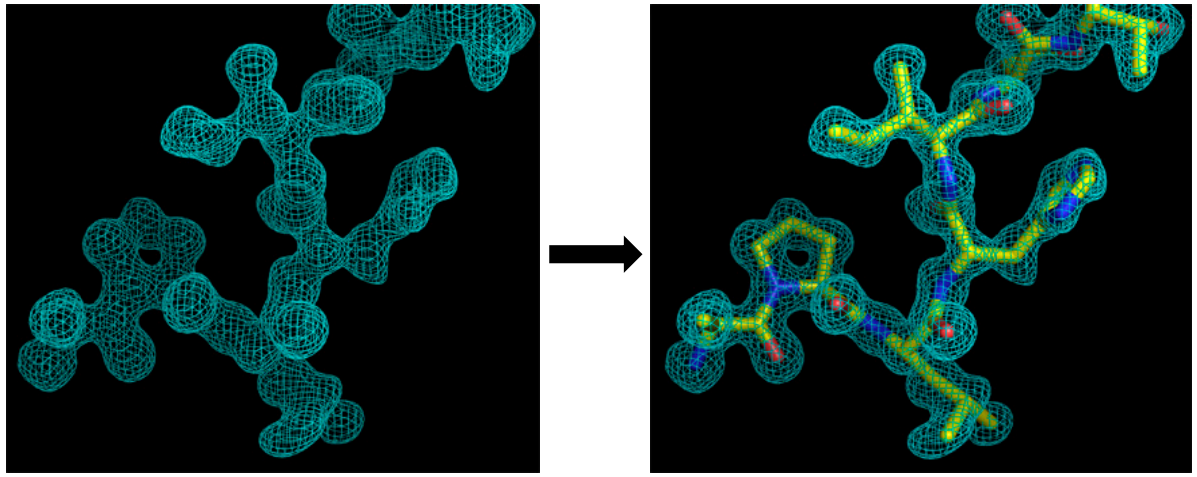
Phase Problem: Review



- From diffraction to electron density map requires Fourier Transform.
- But it's impossible (hard) to extract phase angle from diffraction pattern directly, which is known as **phase problem**.
- To address this:
  - Isomorphous replacement: use the location of metal ion to find the phase angle
  - Anomalous scattering: heavy atoms absorb X-ray radiation and reemit with altered phase.
    - SeMet replacing Met
  - Molecular replacement: place a model of known protein in the unit cell of the new protein

Model Building: Review

- Determination of initial phases allows us to calculate an **initial electron density map** that is used to build the first rough model of the protein
  - $\mathbf{2|F_o|-|F_c|}$: "real map" that is used to add atoms to the model.
  - $\mathbf{|F_o|-|F_c|}$: difference map that highlights sections where the model differs greatly from the experimental data.



- Either by hand or **automatically**, depending on the resolution.
- Refine the model to make the final model agree closely with the experimental data and known stereochemical data.

Resolution

Crystallographers express the resolution of a structure in terms of distance.

–If the resolution is *smaller* than the atoms distance ⟹ atoms' positions can be obtained with high precision

–If the resolution is *bigger* than the atoms distance ⟹ fused electron density

- **Low resolution:**

  ~ 3-5 Angstrom - overall shape, side chains not resolved anymore

- **Medium resolution:**

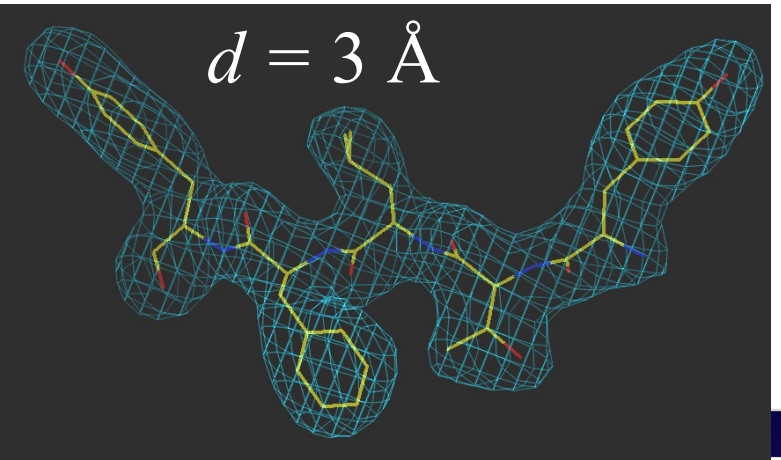
  ~ 2.5-3 Angstrom - side chains can be distinguished

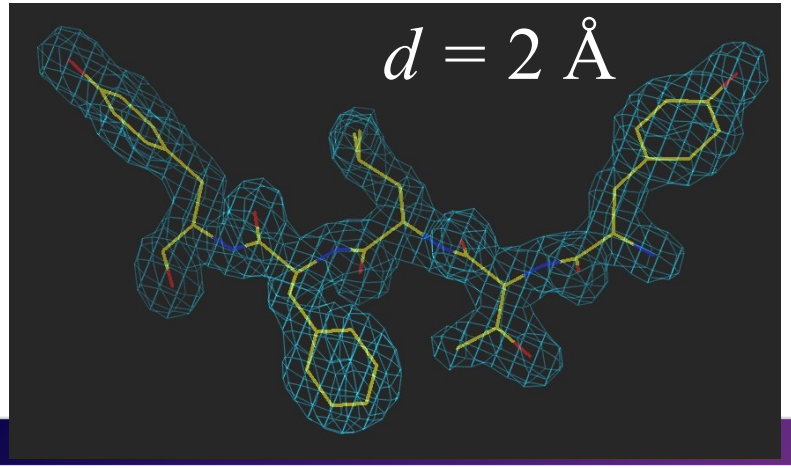
- **High resolution:**

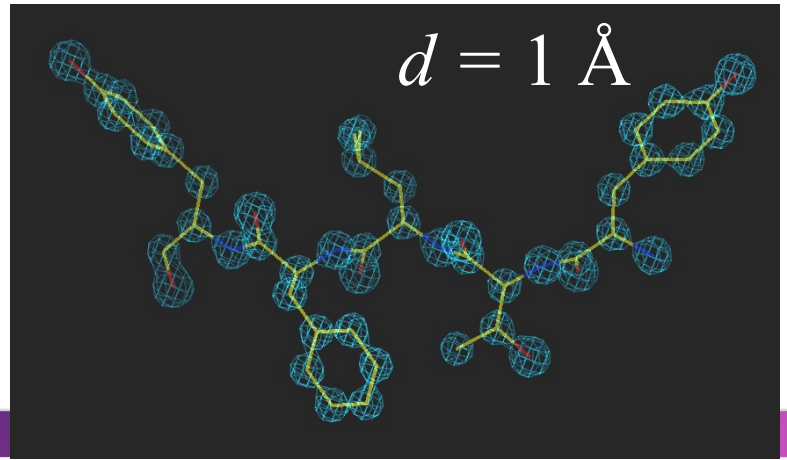
  ~ 2 Angstrom - side chains, waters, ions, small ligands

  < 2 Angstrom - alternate side chain conformations (rotamers), holes in aromatic residues

  < 1.1 Angstrom - some hydrogens







Judging The Quality of Crystallographic Structures$_1$

Data statistics:

1. Completeness – How much of the diffraction pattern was actually recorded. Both the overall completeness and the completeness for high resolution data are important. (overall: ~100%, high resolution shell: ~80%)
2. Multiplicity (redundancy): average number of independent measurements of each reflection in a crystallographic data set. (~ 3.8 multiplicity)
3. R_{merge} or R_{symm} – These "R factors" are a measure of the error in the data – literally the difference between any measurement of a particular reflection and the average of multiple measurements of that reflection (the same reflections are measured multiple times). (Lower R_{merge} is better)
4. Signal to noise ratio (I/σ) – ratio of the average intensity of the diffraction spots and the background intensity
   - At least 2.0 for the highest resolution data.

Judging The Quality Of Crystallographic Structures$_2$

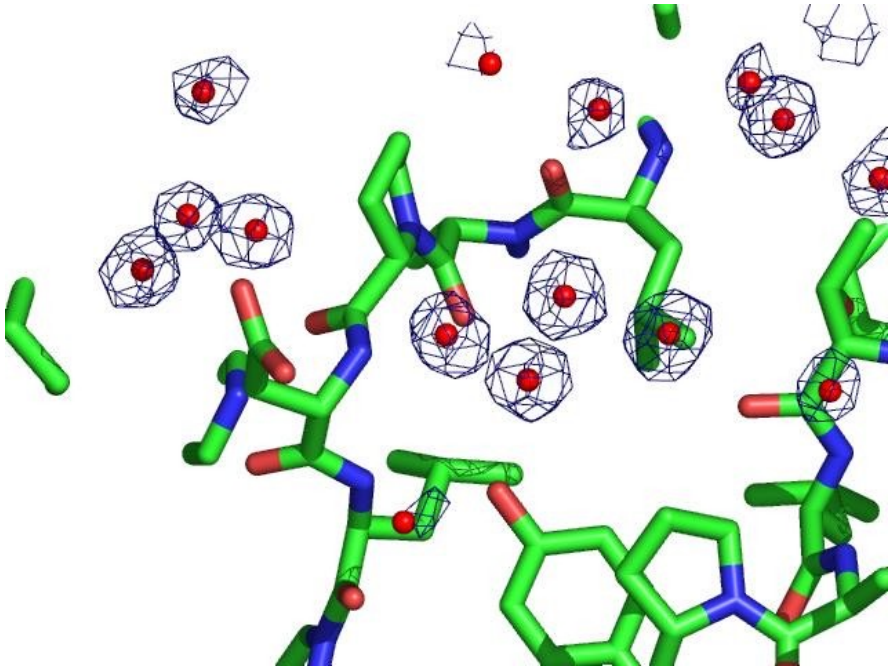
Model quality statistics:

1. Crystallographic R factors (R_{cryst} or R_{free}) – measure the difference between the model and the experimental data. The difference between should be less than 5% at any resolution.
2. Deviations from ideality – The differences between the bond parameters and standard values should be < 0.01 Å and 1.5^o for angles.
3. B factors (temperature factors) – measure the degree of thermal motion of each atom. Show mobile and/or disordered parts of the model.

Things you - as a potential user of crystallographic data - should know about crystals and crystal structures

Two types of solvent: Ordered and Disordered

- Ordered water molecules show up as discrete blobs of electron density in contact with the protein or with other ordered water molecules
- Disordered water regions show up as featureless (flat) electron density



PDB files:

- Basically, just simple text files
- At the top: information about the crystal:
  - Which proteins/ligands etc
  - Crystallization conditions
  - How was the structure solved
  - The resolution
  - Some useful statistics to judge the quality of the crystal
  - How to get from the structure to the biological unit
  - Remarks about missing bits etc.
- Crystal parameters: cell dimensions/space group
- A list of all atoms in the structure

A crystal structure according to the protein data bank (PDB)

occupancy

x,y,z coordinates (Å)

```
ATOM     25  N   ASP A 928      19.062   9.157  35.067  1.00  4.73           N
ATOM     26  CA  ASP A 928      19.770  10.123  34.232  1.00  4.58           C
ATOM     27  C   ASP A 928      19.075   9.938  32.899  1.00  4.56           C
ATOM     28  O   ASP A 928      19.074   8.824  32.351  1.00  5.39           O
ATOM     29  CB  ASP A 928      21.259   9.776  34.071  1.00  3.13           C
ATOM     30  CG  ASP A 928      22.112  10.245  35.233  1.00  5.52           C
ATOM     31  OD1 ASP A 928      21.693  11.114  36.025  1.00  5.42           O
ATOM     32  OD2 ASP A 928      23.239   9.742  35.349  1.00  7.93           O
ATOM     33  N   VAL A 929      18.417  10.985  32.405  1.00  3.68           N
ATOM     34  CA  VAL A 929      17.726  10.864  31.125  1.00  4.63           C
```

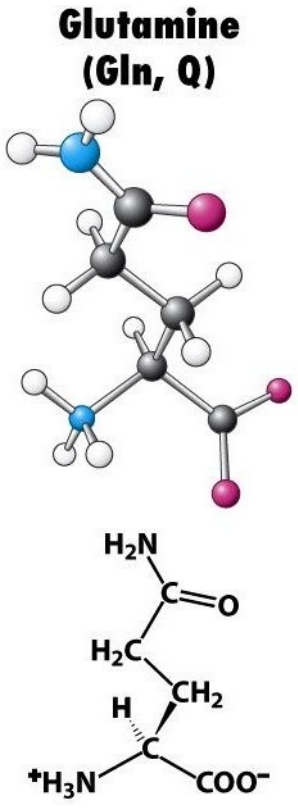
Isotropic B-factor or temperature factor is a measure of the mobility of an atom

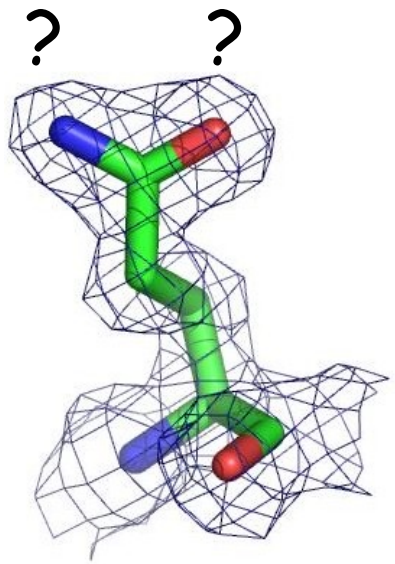
B ($Å^2$) = $8\pi^2\langle u^2\rangle$, where $\langle u^2\rangle$ is the mean square atomic displacement

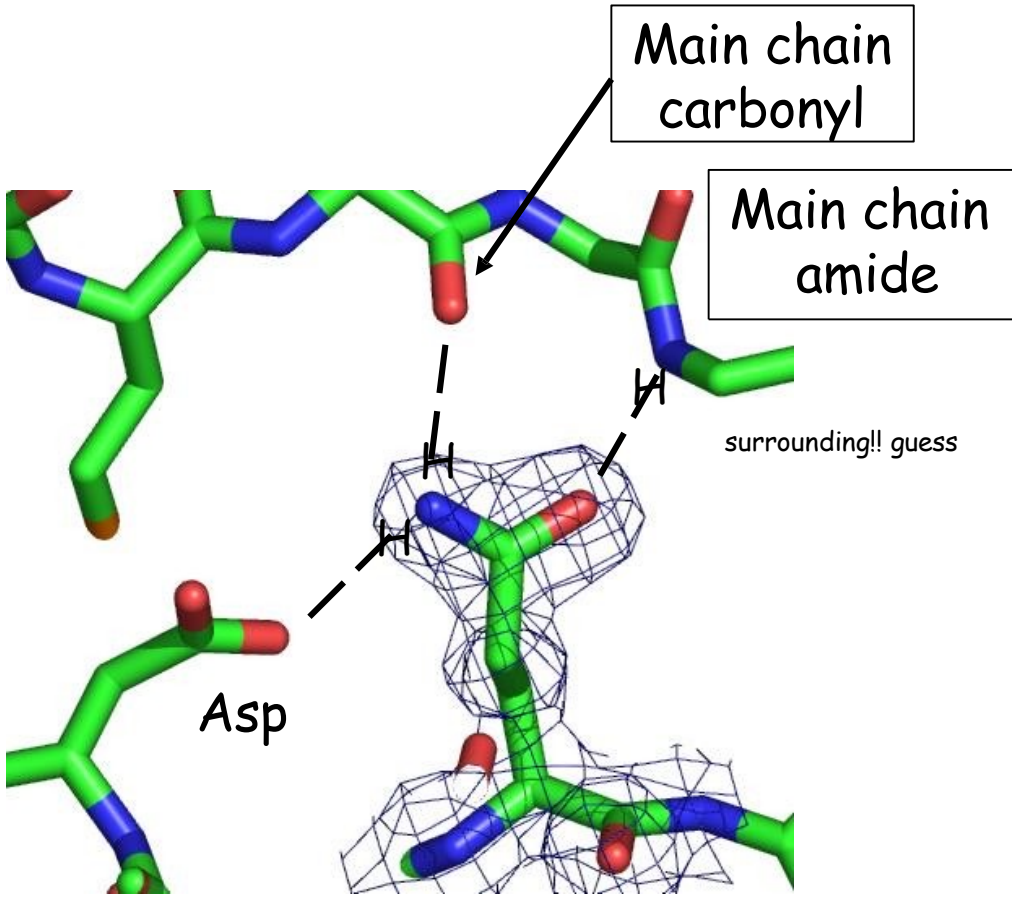
“Occupancy”

- The occupancy “n_j” of atom “j”: is a measurement of the fraction of molecules in the crystal in which atom j occupies the position specified in the model.
- If all molecules in the crystal are precisely identical, then occupancies for all atoms are 1.00.
- Occupancy is necessary because occasionally two or more distinct conformations are observed for a small region like a surface side chain. For example, if the **two conformations occur with equal frequency, then atoms involved receive occupancies of 0.5 in each of their two possible positions.**
- So, “occupancies” is an estimates of the frequency of alternative conformations, giving some additional information about the dynamics of the protein molecule.

Position of N and O atoms in Gln (and Asn) side chain must be inferred from hydrogen bonding network







The same holds for the orientation and protonation of the imidazole ring of histidines

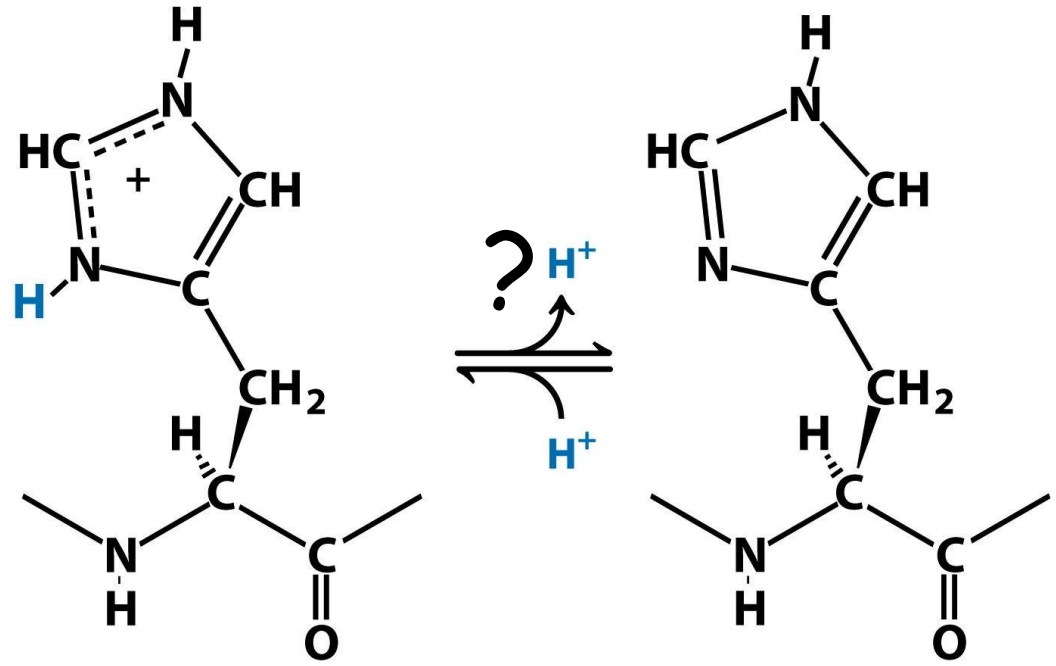
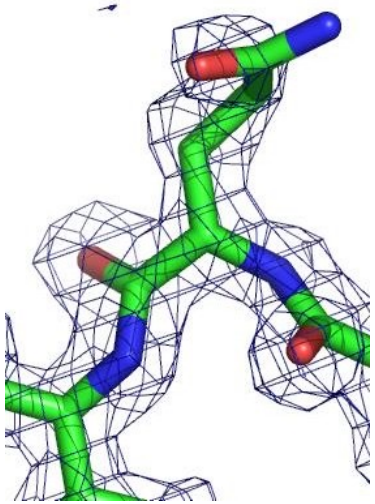
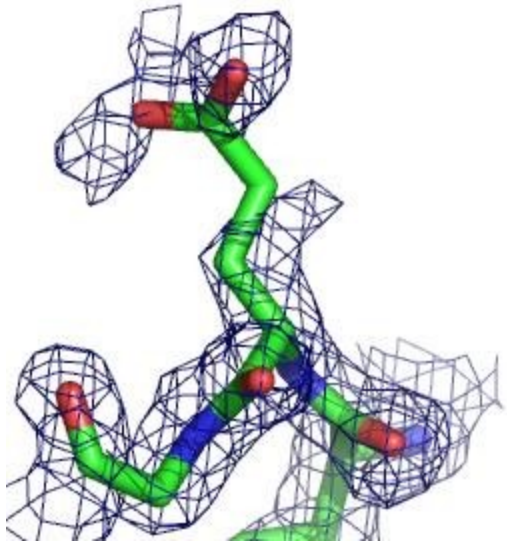


Figure 2-15
Biochemistry, Sixth Edition

A pdb file may contain residues for which no, or only limited electron density is visible

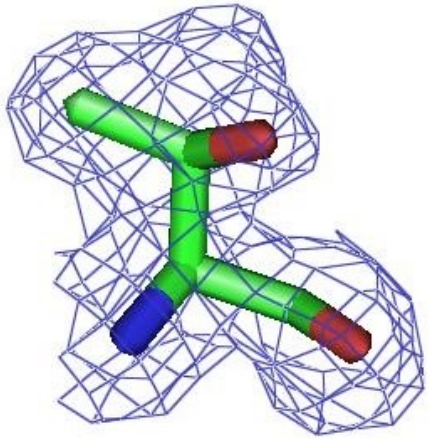


No density for amide N of glutamine

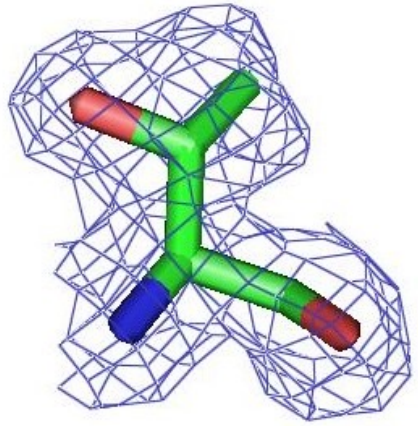


Break in side chain density of glutamate

Sometimes the electron density suggests two side chain conformations but may only one is modeled in the pdb file

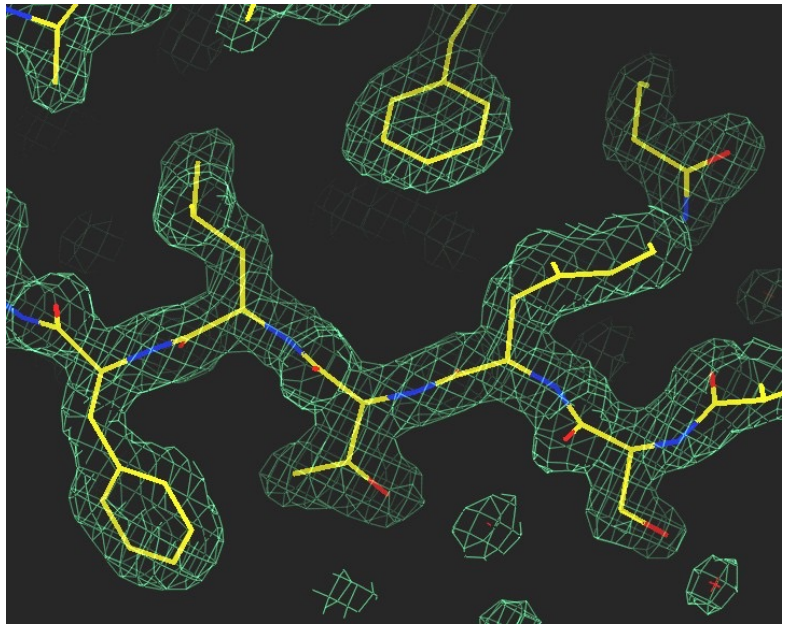


Threonine side chain conformation present in pdb file

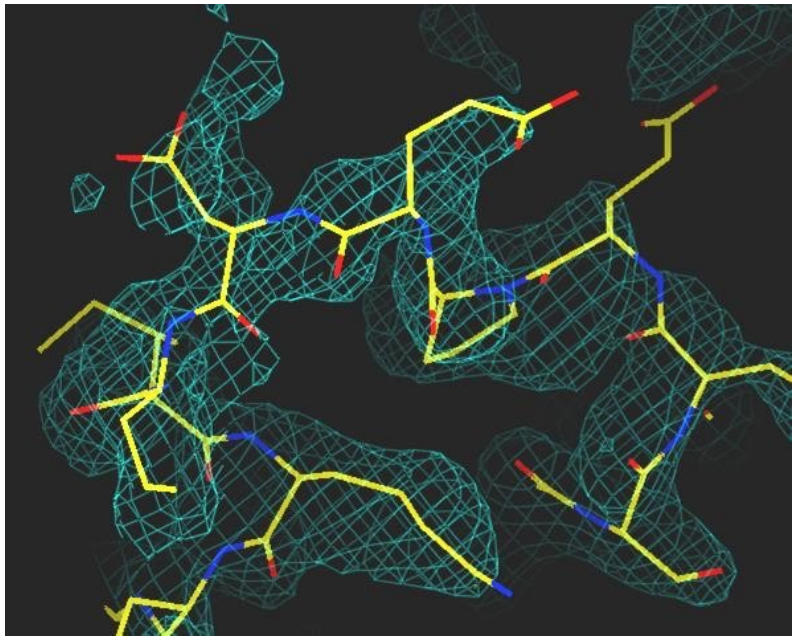


Alternative conformation that is also compatible with electron density

The interpretation of dynamic loops in the pdb file may be tentative

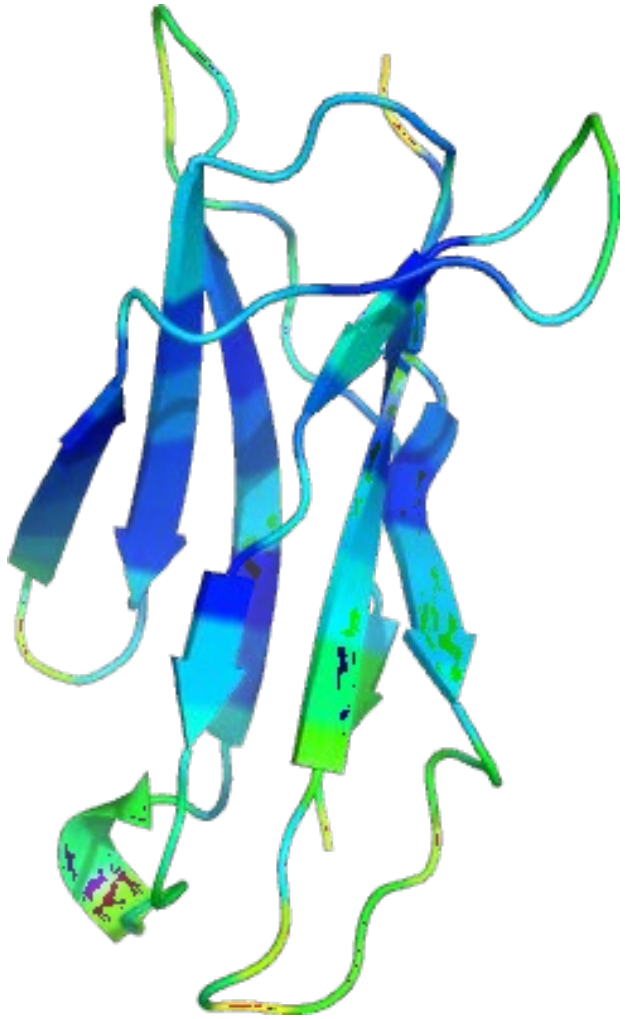


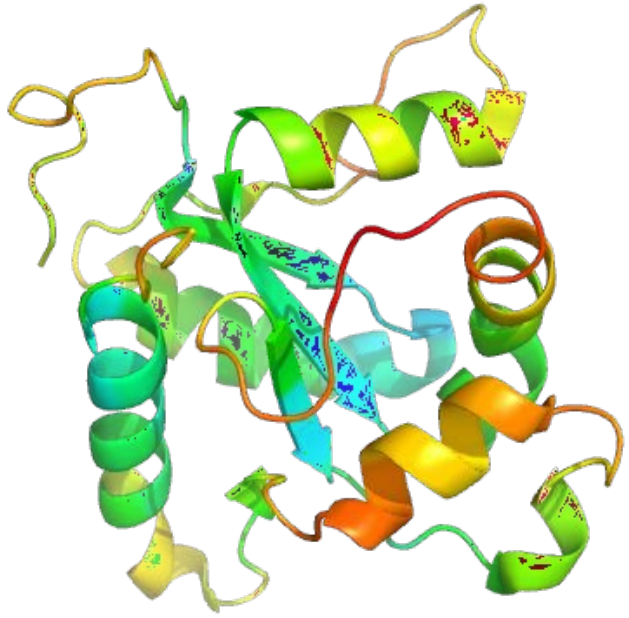
Well defined β-strand in the core of a protein: atomic positions are reliable



Flexible loop at the surface of a protein: atomic positions are not well defined

Look at B-factor distribution!





Protein coloured by B-factor:

Well defined regions have low B-factors (blue/green)

Poorly defined/more mobile regions have high B-factors (yellow/orange/red)

A protein molecule is dynamic

- The electron density is a spatial average over all molecules in the crystal and a time average over the duration of the X-ray data measurement
- Multiple discrete conformations of a residue in different molecules are superimposed.
- Damage caused by X-rays may change the protein (mainly breaking of disulfide bonds)
- A crude description of dynamics is provided in the pdb file as the isotropic “B-factor”
- Some dynamical aspects evident in the electron density are lost in the pdb file

Reading a crystallography paper:

Table 1. Crystallographic Data Collection and Refinement Statistics

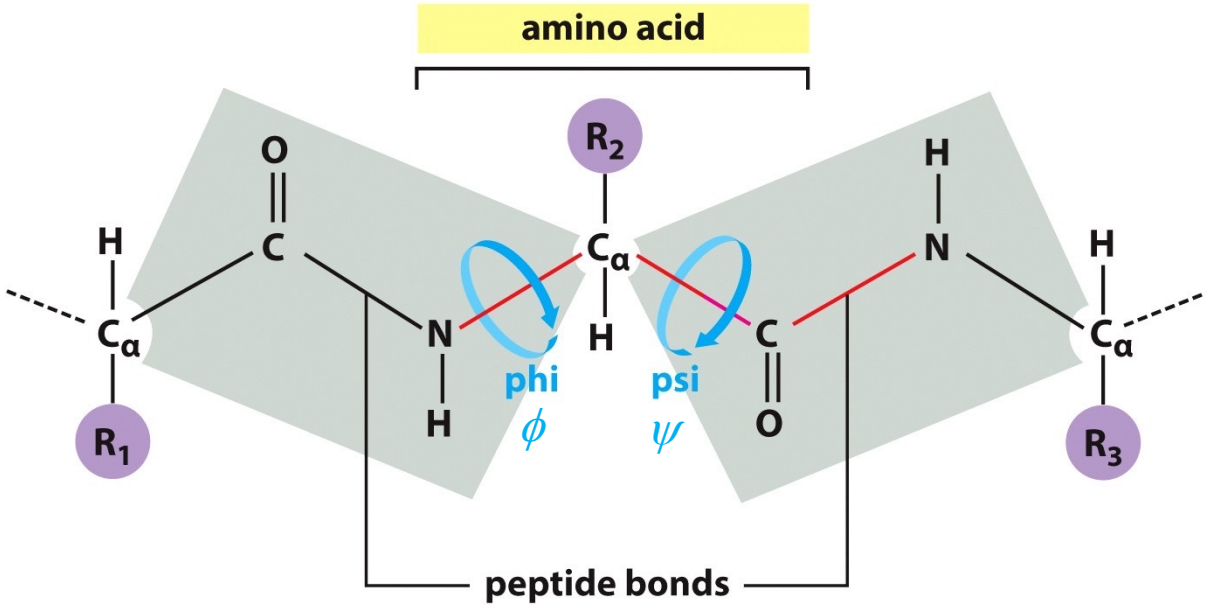
	SwMppP	SwMppP·D-Arg
resolution (Å) (last shell)[a]	41.45–2.10 (2.14–2.10)	44.53–2.25 (2.29–2.25)
wavelength (Å)	0.97852	0.97856
no. of reflections		
observed	1532933 (60186)	256371 (12494)
unique	106188 (5236)	79816 (3984)
completeness (%)[a]	100.0 (100.0)	90.2 (91.2)
R_{merge} (%)[a,b]	0.106 (0.734)	0.106 (0.726)
multiplicity	14.4 (11.5)	3.2 (3.1)
$\langle I/\sigma(I)\rangle$[a]	30.7 (5.2)	10.0 (1.9)
Model Refinement		
no. of reflections in the working set	100876	76747
no. of reflections in the test set	5256	3015
R_{cryst} (R_{free})	0.148 (0.177)	0.162 (0.197)
no. of residues	1417	1404
no. of solvent atoms	903	693
no. of TLS groups	29	33
average B factor (Å^2)[c]		
protein atoms	32.0	34.1
ligands	30.9[d]	43.9[d]
solvent	35.6	36.6
root-mean-square deviation		
bond lengths (Å)	0.013	0.015
bond angles (deg)	1.395	1.565
coordinate error (Å)	0.17	0.22
Ramachandran statistics (favored/allowed/outliers) (%)	98.3/1.7/0	98.4/1.6/0

[a]Values in parentheses apply to the high-resolution shell indicated in the resolution row. [b]$R = \sum(||F_{obs}| - \text{scale} \times |F_{calc}||)/\sum|F_{obs}|$. [c]Isotropic equivalent B factors, including the contribution from TLS refinement. [d]In the unliganded SwMppP structure, "ligands" refers to the bound Cl ions, while in the D-Arg complex structure, it refers to the D-Arg-PLP unit.

Judge the quality of the data:

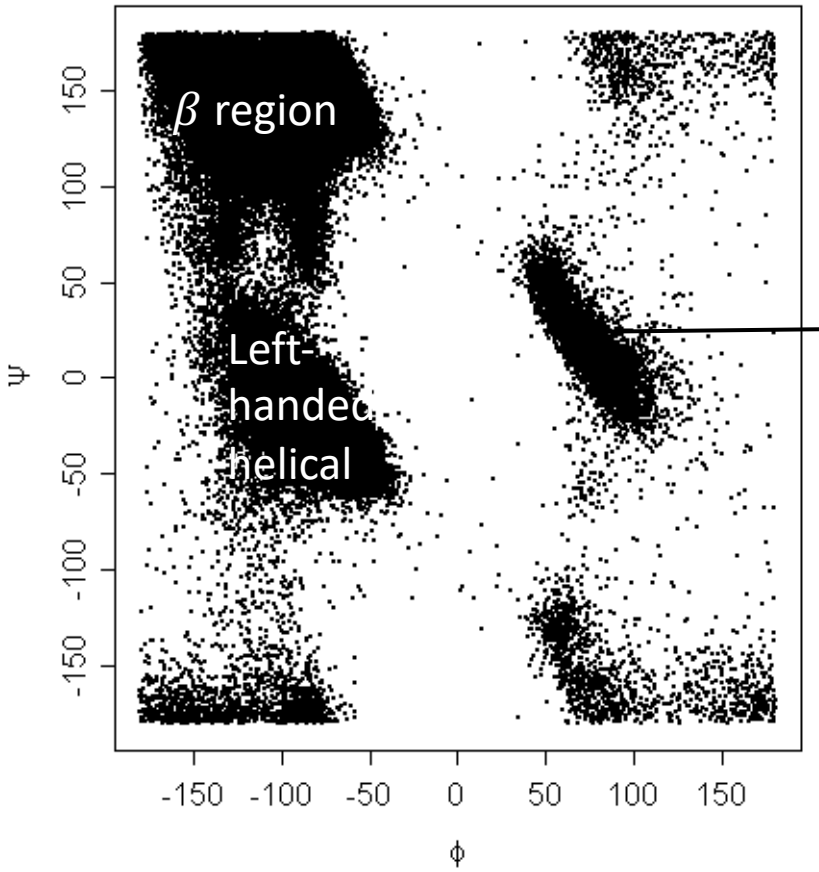
- R_{merge}: 0.05-0.10 good, 0.1-0.15 acceptable
- I/σ = signal/noise >2.0
- Completeness
- Redundancy
- R_{work}/R_{free}:
  - difference < 0.05,
  - R_{work}≈ resolution/10
- Deviations of known geometry

Ramachandran Plot



The torsional angles of each residue define the geometry of its attachment to its two adjacent residues, so the torsional angles determine the conformation of the residues and the peptide.

- shows frequency of (ϕ, ψ) observed for residues in folded proteins



Each dot represents an amino acid

The **green/yellow regions** correspond to conformations where there are no steric clashes

White regions: sterically disallowed for all amino acids except glycine