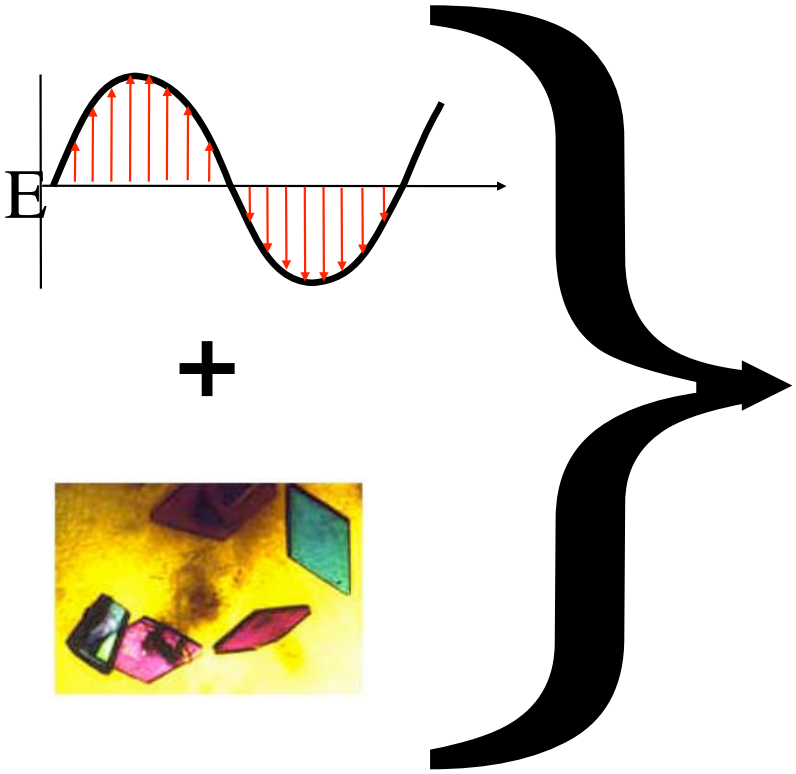
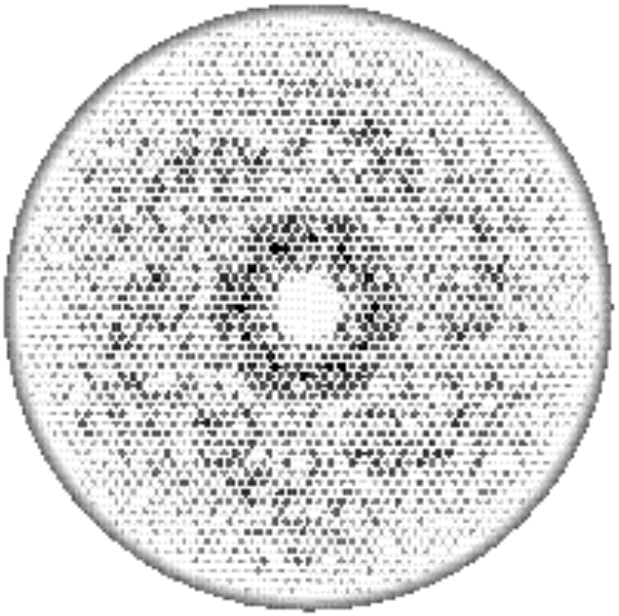
Molecular Modeling 2021 -- lecture 2 -- fri Jan 29

Where do protein structures come from?

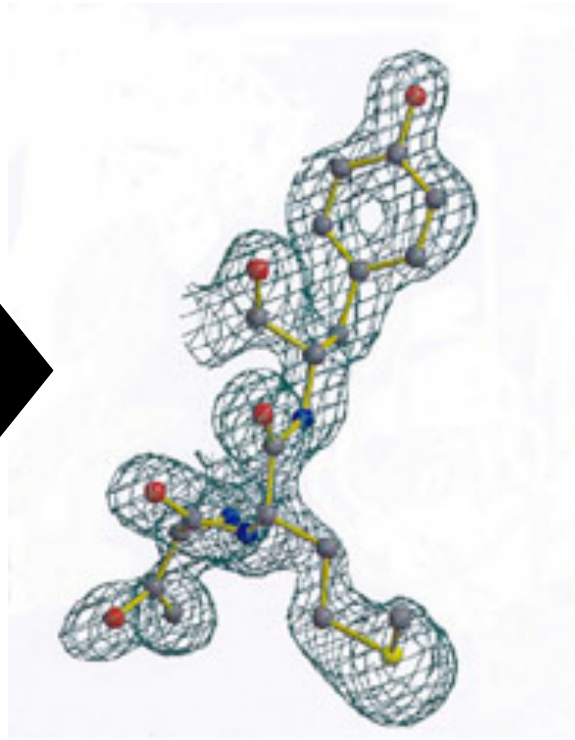
- X-ray crystallography
- Solution NMR

X-ray Crystallography





Diffraction pattern

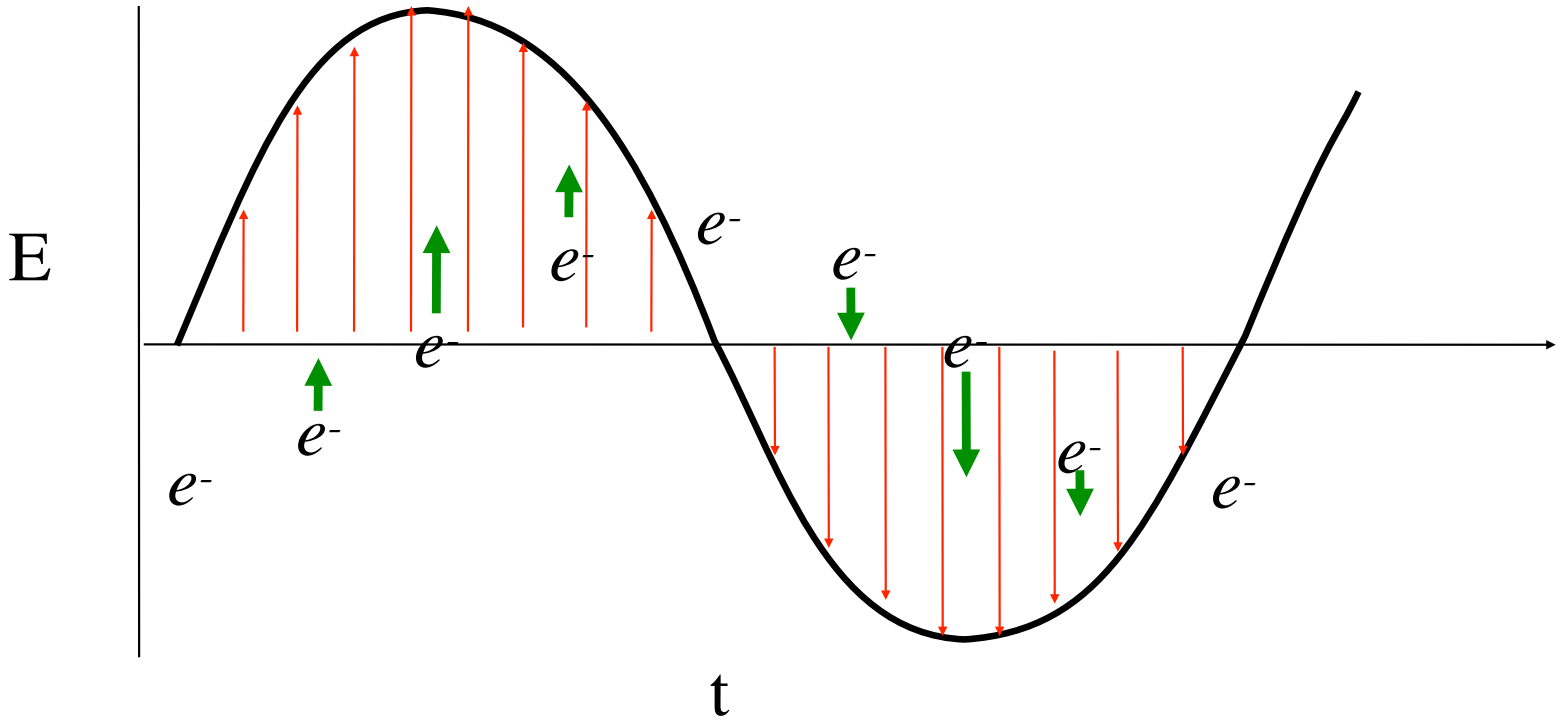


Electron density map

e^- oscillation scatter X-rays

- *e*- has almost zero mass, so it oscillates at the same frequency as the X-rays

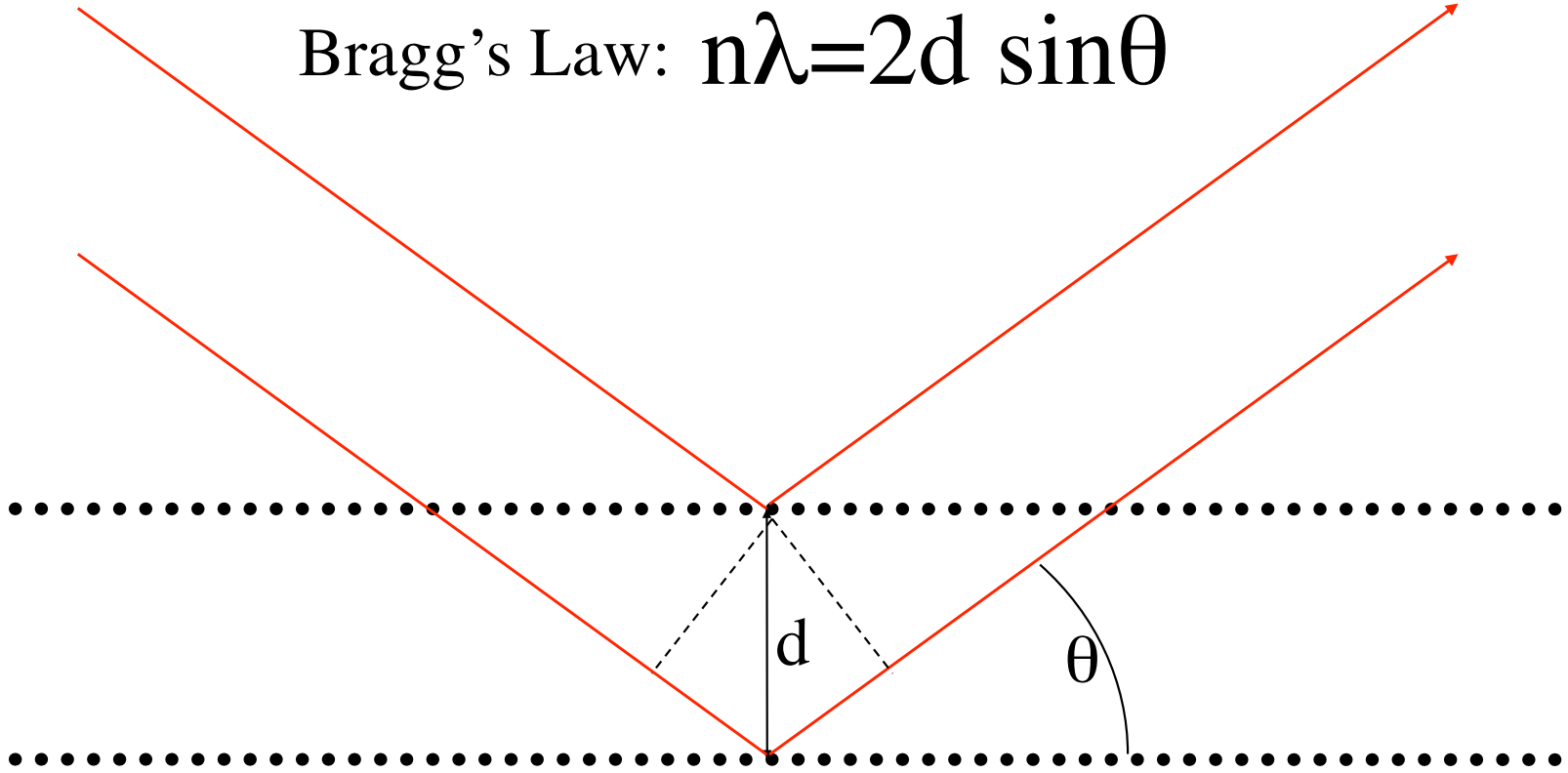
- Oscillating e^- emit light in all directions.



Wavelength: $\lambda \approx 1.54 Å$ Frequency = $c/\lambda \approx 2 \times 10^{18}\ s^{-1}$

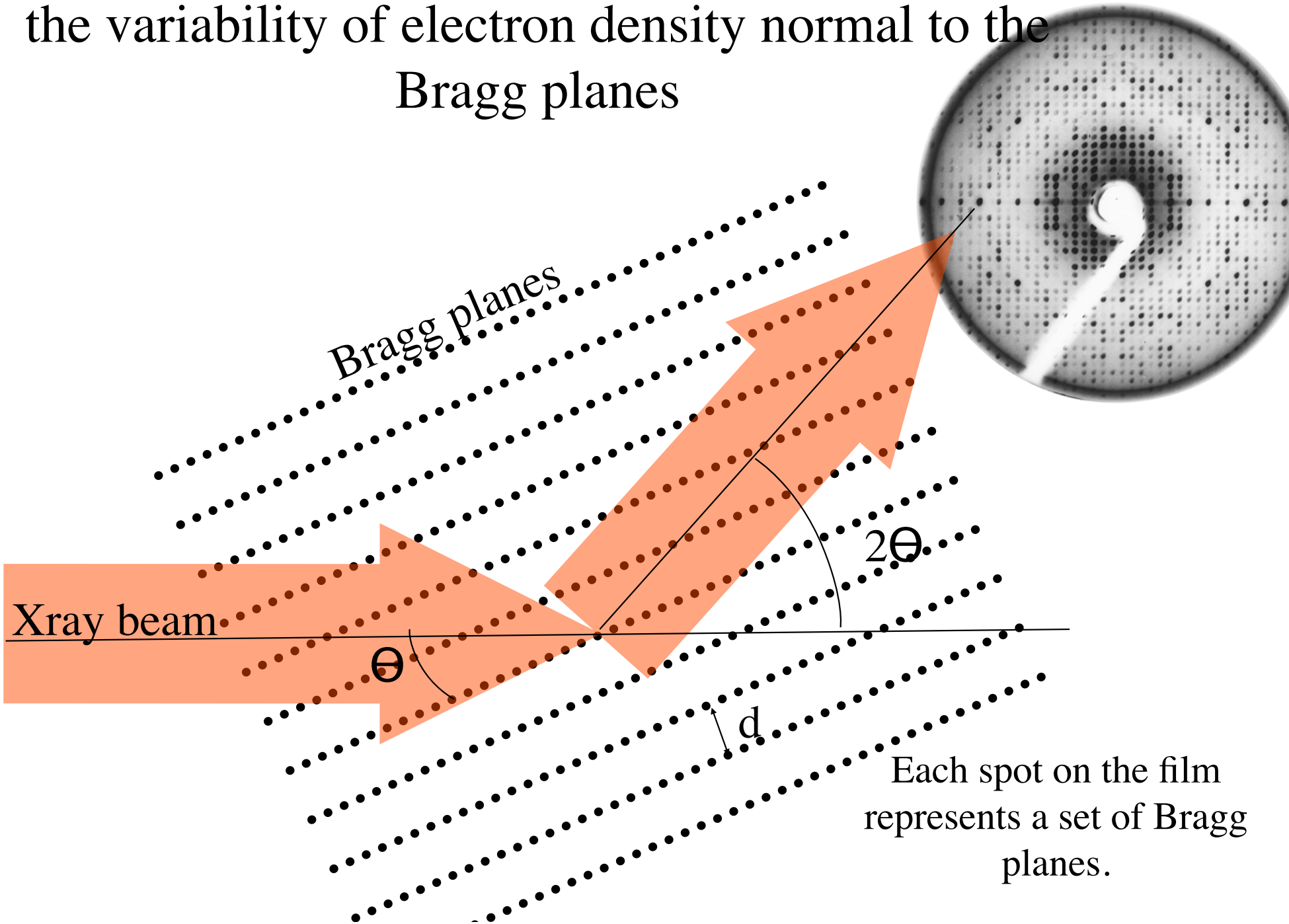
Bragg Planes are
Parallel mirrors separated by *d*

Bragg's Law: $n\lambda=2d\sin\theta$



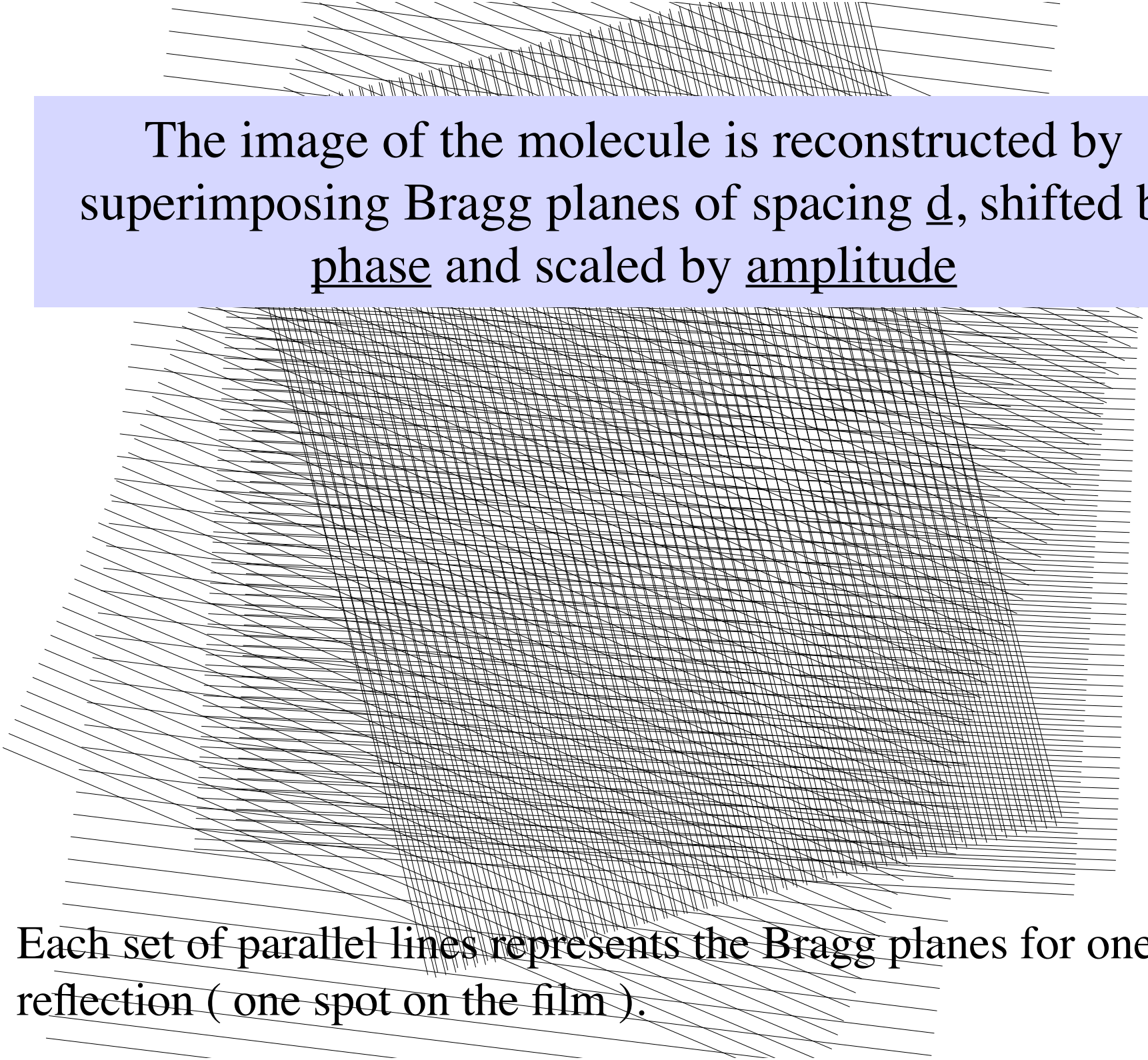
d is the *resolution*, which depends on Θ and λ

The amplitude of each reflection is proportional to the variability of electron density normal to the Bragg planes



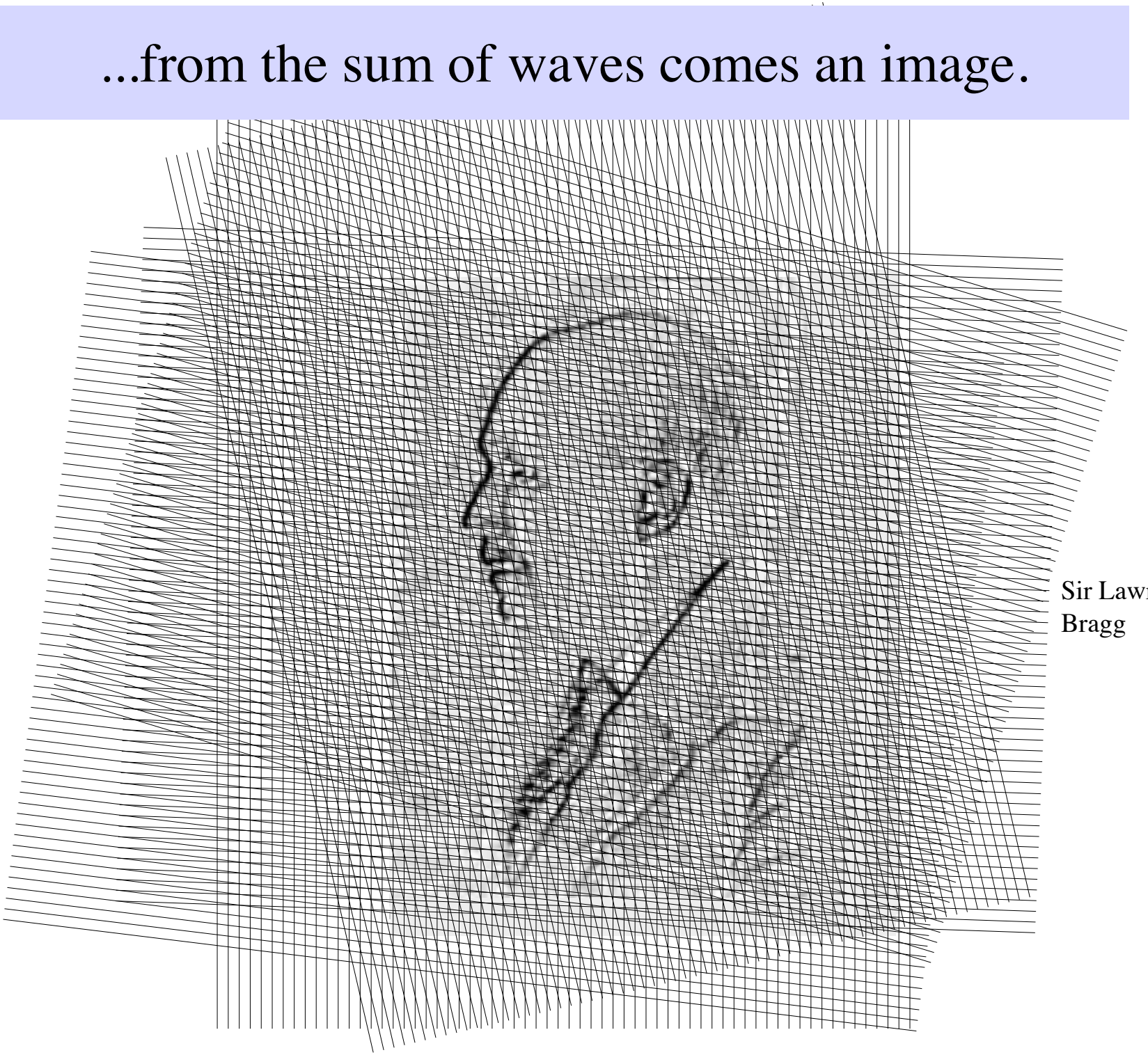
Each spot on the film represents a set of Bragg planes.

The image of the molecule is reconstructed by superimposing Bragg planes of spacing d, shifted by phase and scaled by amplitude



Each set of parallel lines represents the Bragg planes for one reflection (one spot on the film).

...from the sum of waves comes an image.



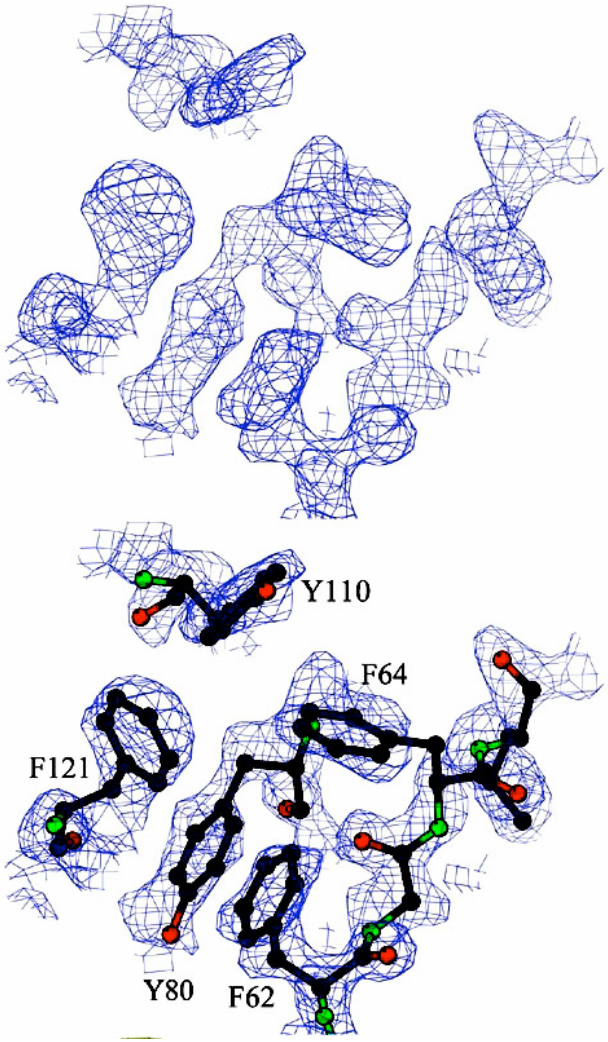
Sir Lawrence Bragg

Fitting the model to the density

3D electron density map = electron density at every point in space.

Visualized by drawing 3D contours.

Since we know the amino acid sequence and we know what the amino acids should look like, we can "fit" a model to the density.



Coordinate refinement

Each atom is moved in X,Y and Z until:

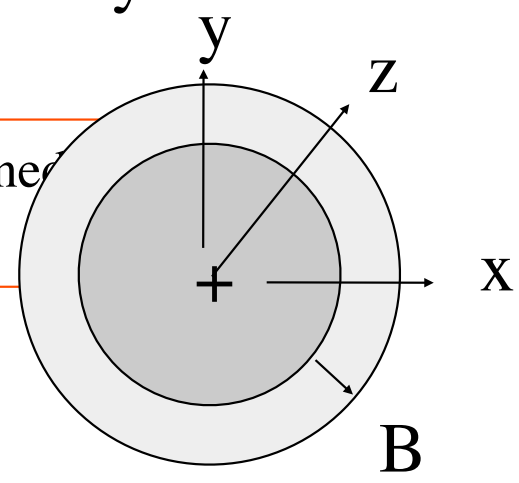
(1) good stereochemistry is achieved,

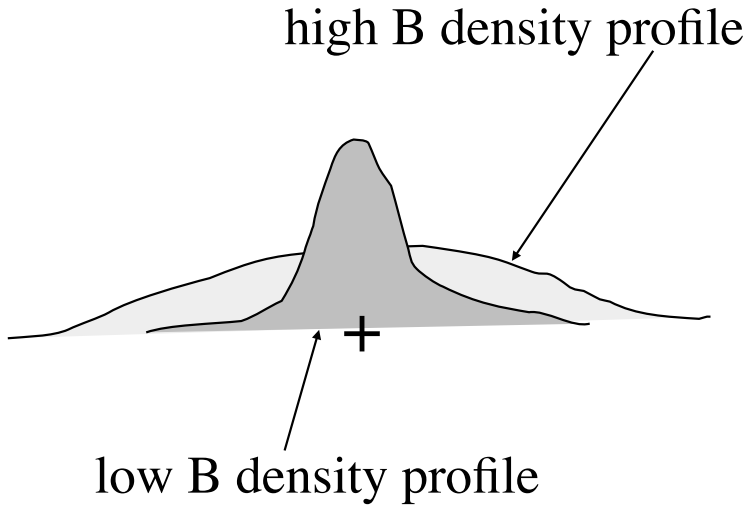
the R-factor

(2) there is a good match between the atoms and the density.

Each atom is assigned a B-factor or "*temperature-factor*", to better fit the density.

4 parameters are refined for each atom





Refined coordinates are deposited in the Protein Data Bank: **www.rcsb.org**

Structure quality: R-factor

- R-factor = $\sum(F_c - F_o)/\sum(F_o)$
- Free R-factor = R-factor calculated on data not used for refinement. Free-R is not biased by overfitting.

free R-factor	**quality**
below 20%	very good
20-30%	typical
30-35%	barely acceptable
above 35%	junk!

Structure quality: resolution

- Resolution = *d* in Bragg's Law. $n\lambda = 2d \sin\theta$. Lower *d* is higher resolution.
- "Resolution" = resolution limit = the lowest *d* observed = the highest scattering angle observed.

Resolution	**quality**
> 4Å	nearly worthless, shows blobs of density
3-4Å	medium. Shows backbone and some sidechains.
2-3Å	typical good structure, all sidechains visible
1.5-2Å	high resolution. Atom positions known within 0.1Å rmsd.
< 1.5Å	ultra high resolution! Hydrogens sometimes visible.

Steps in Xray crystallography

Overview…

1. Purify and concentrate protein.
2. Crystallize.
3. Collect Xray data (1000's of reflections)
4. Solve for the phases*
5. Model atoms into density
6. Refine.

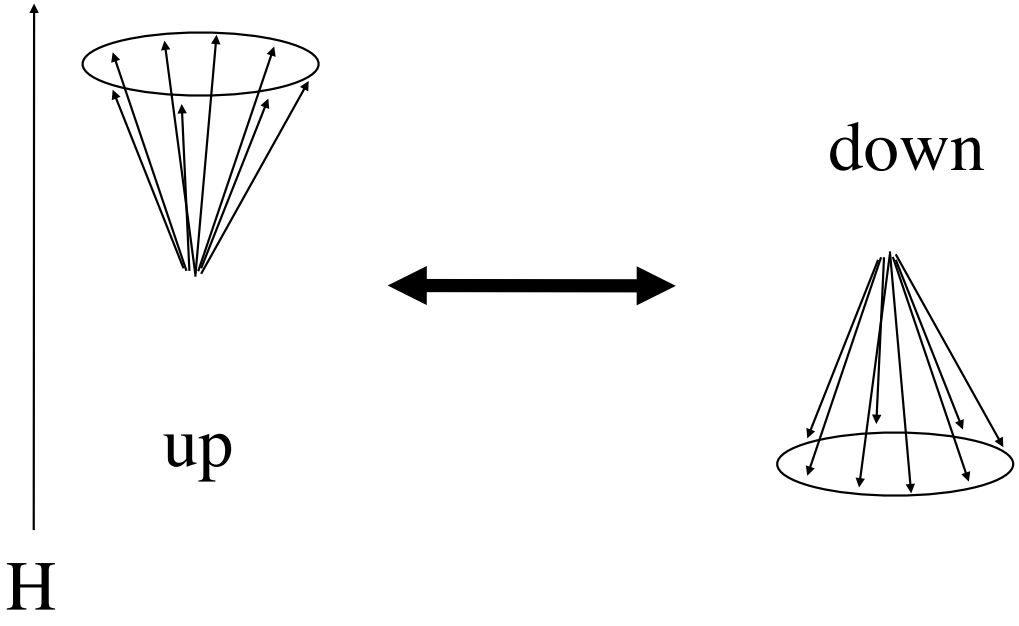
*using molecular replacement or heavy atom methods.

Nuclear Magnetic Resonance

Isotopes that have nuclear spin = 1/2

^{1}H, ^{13}C, ^{15}N and ^{31}P

..can adopt two orientations in a **magnetic field (H)**.
At equilibrium slightly more spins are aligned with the field than against it.



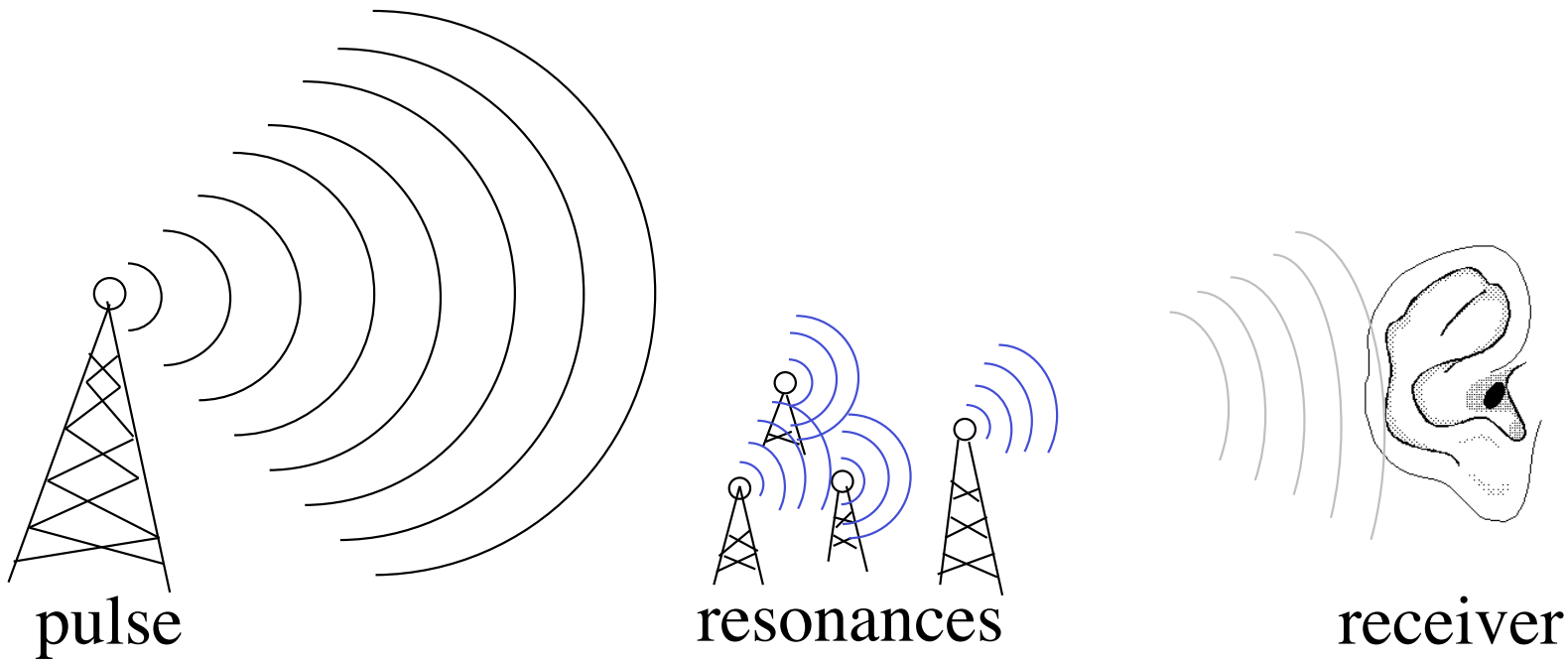
The difference in energy between up and down states lies in the **radio frequency** range.

Radio pulses perturb equilibrium, which relaxes back, emitting radio freq.

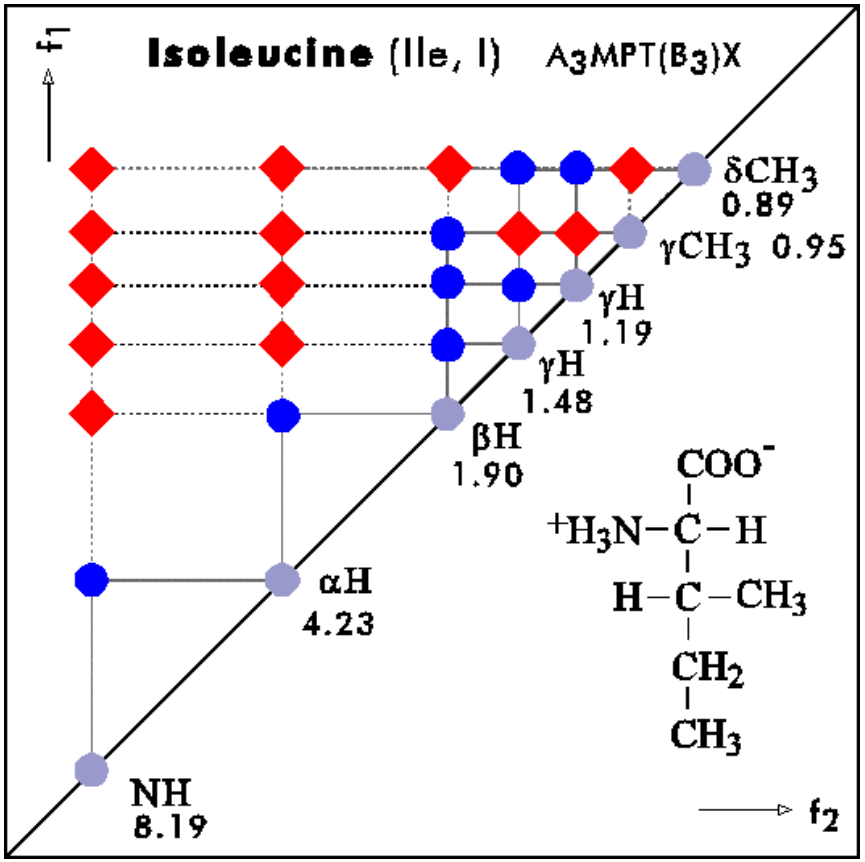
Short pulses "ring" through bonds --> **TOCSY**

A **TOCSY** experiment finds cross-talk between ^{1}H in a "spin system." Characteristic sets of resonances allow the easy identification of amino acids.
A **COSY** experiment finds cross-peaks between ^{1}H that are separated by 2 or 3 bonds.

Long pulses “resonate” through space --> **NOESY**

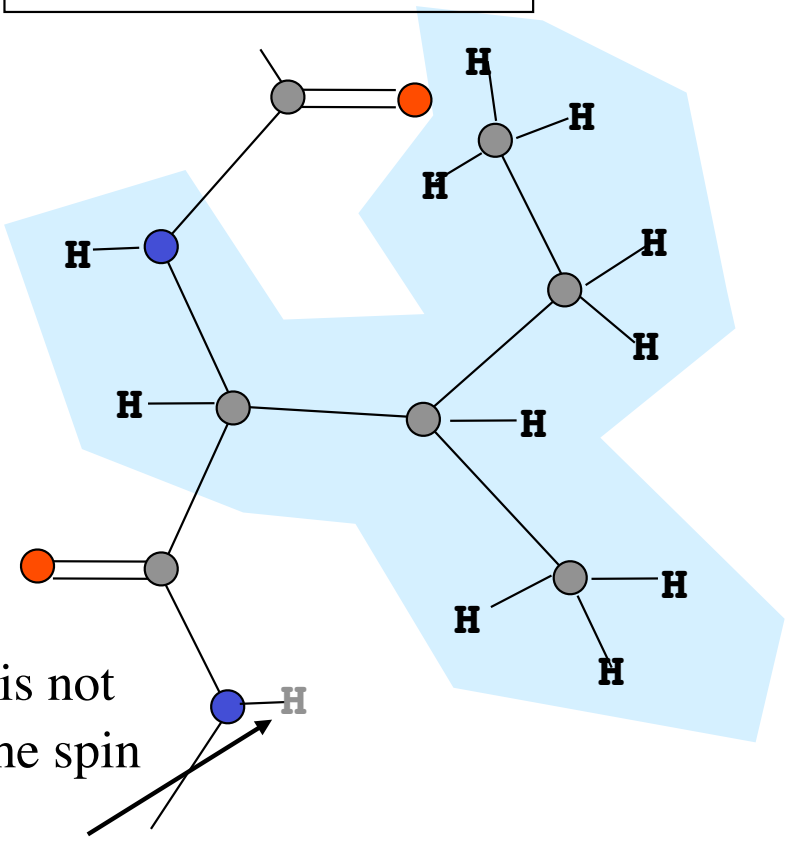


TOCSY/COSY. Characteristic patterns allow assignment of amino acid



TOCSY peaks: red diamond
COSY peaks: blue circle

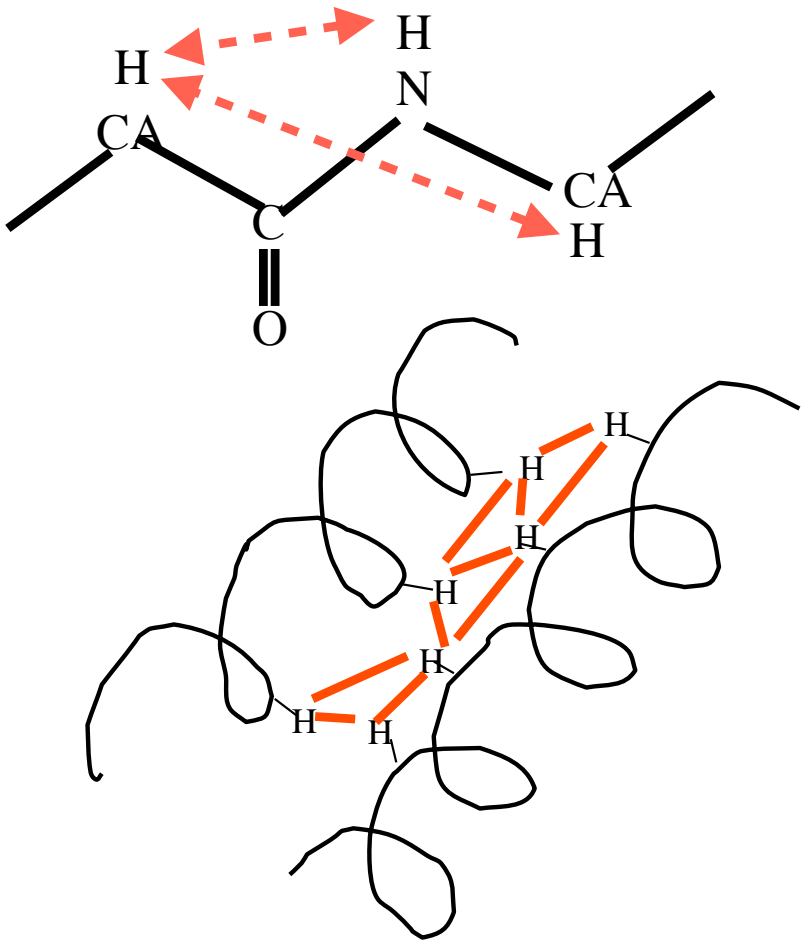
Chemical shifts for ILE:	
NH	8.19
αH	4.23
βH	1.90
γCH3	0.97,0.94
γCH2	1.48,1.19
δCH3	0.89



This ^{1}H is not part of the spin system

NOESY: finds short distances

NOESY spectra tell us which ^{1}H are physically close in space, causing the Nuclear Oberhauser Effect (NOE).



NOE's occurring between sequential residues, allow assignment of a sequence position to a resonance.

The structure is solved by distance geometry calculations. Atomic positions must satisfy the constraint distances and the stereochemistry.

Molecular dynamics is used to refine the solution(s).

Distance geometry problem

- Minimize the distance geometry function $\boldsymbol{f}$ as a function of the atomic positions $\boldsymbol{x}$.

$$f(x) = \sum_{i,j \in \mathcal{S}} p_{i,j}(x_i - x_j)$$

Penalty function *p*

Parabolic square well

$$p_{i,j}(x) = \min^2\left\{\frac{\| x \|^2 - l_{i,j}^2}{l_{i,j}^2},\ 0\right\} + \max^2\left\{\frac{\| x \|^2 - u_{i,j}^2}{u_{i,j}^2}\right\}$$

Hyperbolic smooth well

$$p_{ij} = \sqrt{\lambda^2\left(c - \sqrt{\|x_i - x_j\|^2 + \tau^2}\right)^2 + \tau^2},$$

l_{ij} is lower limit
u_{ij} is upper limit
$\|x\| = \|x_i - x_j\|$ is actual ij distance.

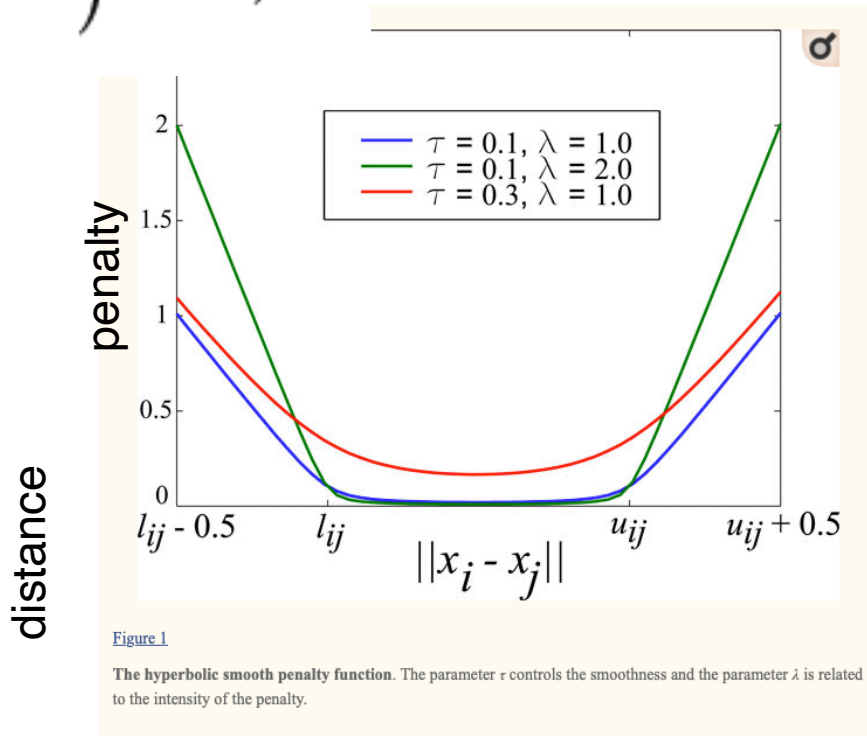


Figure 1

The hyperbolic smooth penalty function. The parameter τ controls the smoothness and the parameter λ is related to the intensity of the penalty.

Moré, J. J., & Wu, Z. (1999). Distance geometry optimization for protein structures. *Journal of Global Optimization*, *15*(3), 219-234

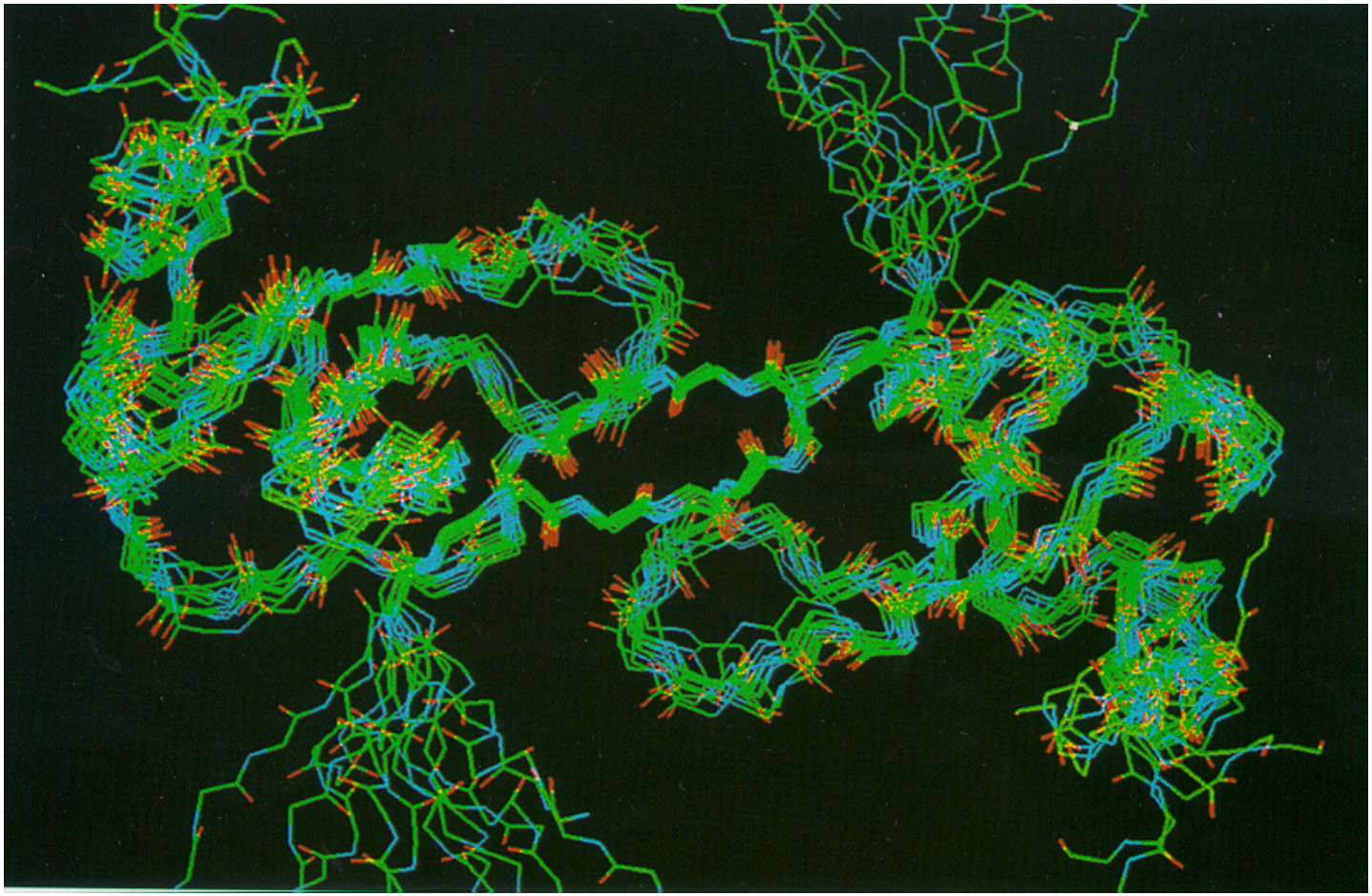
Souza, M., Lavor, C., Muritiba, A., & Maculan, N. (2013). Solving the molecular distance geometry problem with inaccurate distance data. *BMC bioinformatics*, *14*(S9), S7.

Steps in Protein NMR

Overview…

1. Grow protein in ^{13}C and/or ^{15}N enriched media.
2. Purify and concentrate protein.
3. Collect NMR spectra (2,3 or 4-dimensions).
4. Assign the peaks (TOCSY/COSY).
5. Assign distance constraints (NOESY)
6. Solve the distance geometry problem.

NMR result: an ensemble of structures



Ensemble = the set of structures that satisfy distance geometry and stereochemistry. Shows flexible and poorly modeled regions.

In class exercise 2.1

- Go to www.rcsb.org
- Search for DHFR
- Select and download 2hqp and 3frd
- Open the Xray file 3frd in MOE
- Show all atoms. Make them spacefill. Hide solvent.
- Color all atoms by B-factor.
- How are the B-factors distributed?

In class exercise 2.1

- Upload the NMR file 2hqp. Model limit: all
- Select all 2hqp chains. Hide selected. Ribbon.
- Are the uncertainties/flexibilities you see in 2hqp, in the same place as the high B-factors in its homolog 3frd?

Other NMR experiments

Additional information about the conformation may be gained by

- H/D-exchange

  Deuterium (^{2}H) is invisible to NMR. Disappearing ^{1}H's tell us which ones are exposed to solvent. Especially amide NH's.

- Temperature sensitivity of resonances.

  Chemical shift oh ^{1}H changes with T *less* if H-bonded.

- HSQC

  Direct coupling of ^{15}N to ^{1}H through a single bond.

Compare and contrast Xray and NMR

	Xray	**NMR**
Wavelength	10^{-10} m (Å)	1 m
Emitters	electrons	nuclei
Preparation	crystals	isotope labeling
Coordinates	Cartesian	Internal

Review questions

- What causes Xrays to scatter?
- What causes diffraction?
- What are the results of Xray crystallography?
- What is a temperature factor?
- What wavelength light is used in Xray crystallography?
- What does “resolution” mean in Xray crystallography?
- What is a crystal?
- What wavelength of light is used in NMR?
- What kind of atom resonates with light?
- What is an ensemble in NMR?
- What measure in NMR is the analog of resolution in Xray?
- What type of NMR experiement assigns resonances to amino acid types?
- What type of NMR experiment provides distances between different parts of the protein chain?
- Which method produces Cartesian coordinates? Internal coordinates?

Supplementary slides

2.3 Coordinate systems

- Cartesian versus Internal

Before Cartesian cartography

- Internal coordinates versus global coordinates

A. Vespucci's map of the world, made before J. Harrison's clock (1735), using internal coordinates.

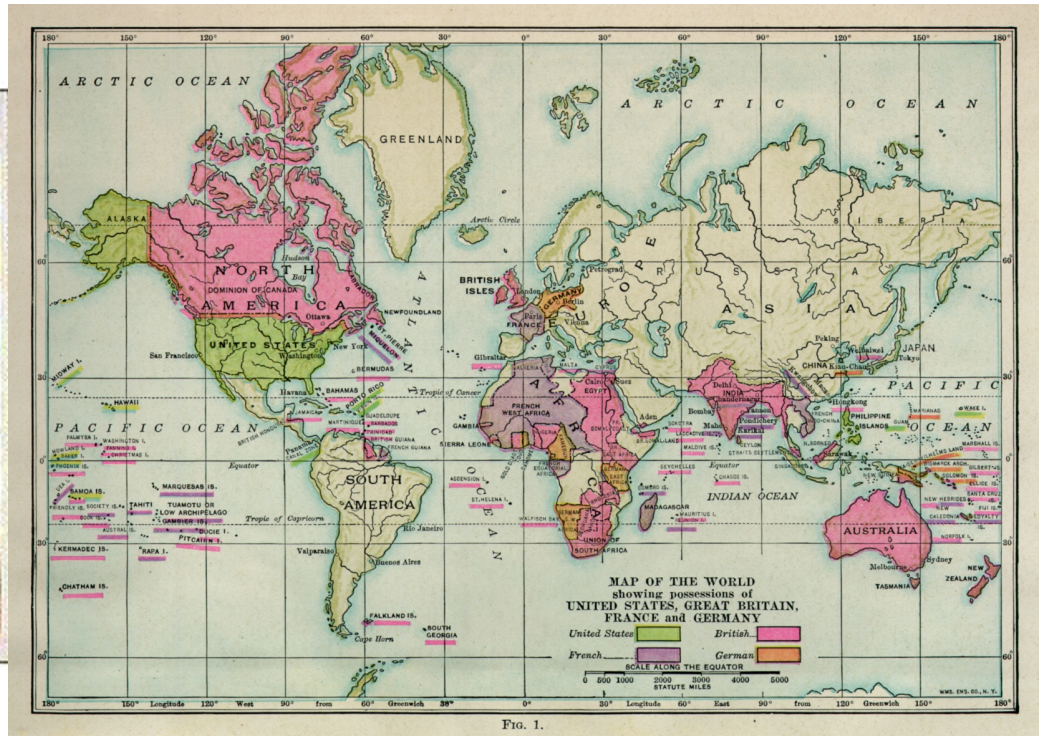


FIG. 1.

...and after, using global coordinates.

Internal coordinates versus global coordinates

Internal coordinates to my house:

From the walking bridge, take a right, go five blocks, then take a left and a right, then bear left and go half a block. It's on the left.

Global coordinates of my house:

N42° 37′ 04″

W73° 44′ 24″.

Global or internal? 110 8th St, Troy NY 12180

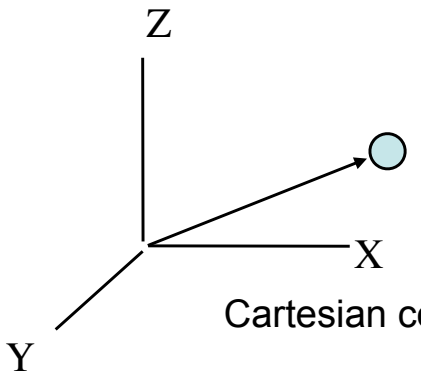
Global or internal? directions from a GPS

Two ways to express structure: Cartesian coordinates

					X	Y	Z				
ATOM	1	N	VAL	1	0.616	-1.613	20.826	1.00	68.81		8DFR 152
ATOM	2	CA	VAL	1	0.737	-1.197	19.414	1.00	65.36		8DFR 153
ATOM	3	C	VAL	1	0.597	-2.511	18.644	1.00	62.65		8DFR 154
ATOM	4	O	VAL	1	1.207	-3.526	18.989	1.00	65.13		8DFR 155
ATOM	5	CB	VAL	1	1.994	-0.410	19.048	1.00	67.55		8DFR 156
ATOM	6	CG1	VAL	1	2.452	0.572	20.132	1.00	68.01		8DFR 157
ATOM	7	CG2	VAL	1	3.154	-1.279	18.586	1.00	66.94		8DFR 158

Biologists use *angstroms*, physicists use *nanometers*.

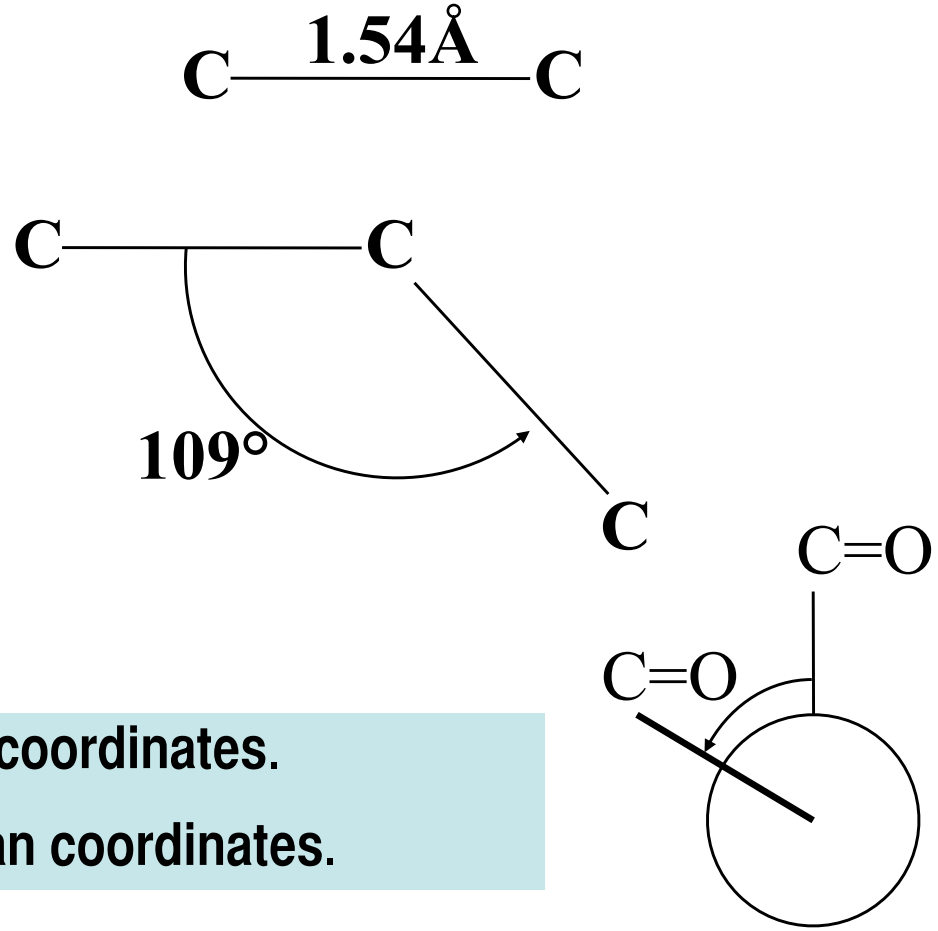
Å = angstroms = 10^{-10} m 1Å = 0.1 nm



Cartesian coordinates have a reference frame

Internal coordinates are independent of reference frame

- Internal coordinates model the covalent structure of the molecule.
- Components:
  - bond lengths
  - bond angles
  - torsion (dihedral) angles
  - planar groups
  - pairwise distances



NMR structures are solved in **Internal coordinates.**

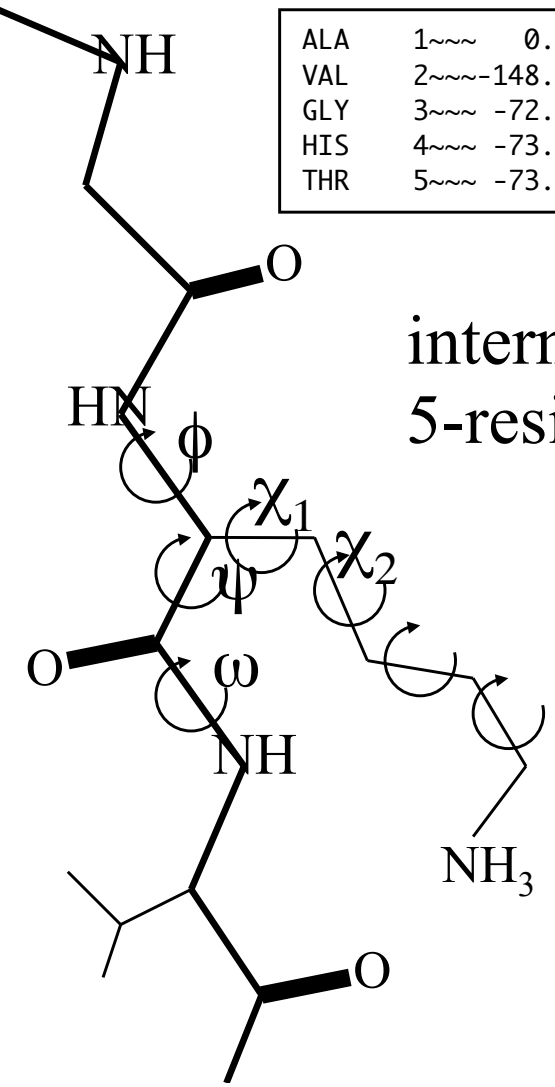
X-ray structures are solved in **Cartesian coordinates.**

Short peptides can be expressed as a set of torsion angles

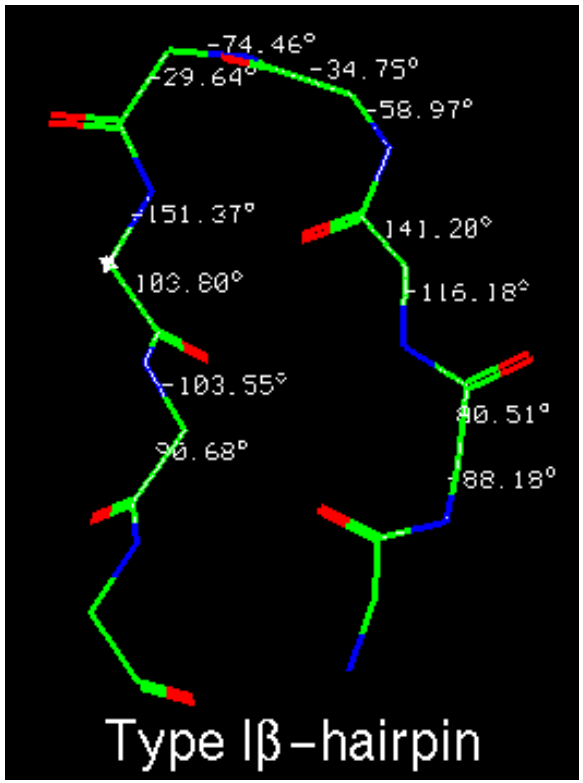
ϕ ψ ω χ_1 χ_2

ALA	1~~~	0.000	127.140	180.000		
VAL	2~~~	-148.378	111.409	180.000	-179.551	
GLY	3~~~	-72.763	39.684	180.000		
HIS	4~~~	-73.084	122.882	180.000	-87.256	-62.962
THR	5~~~	-73.735	116.210	180.000	49.292	

internal coordinates of a 5-residue peptide



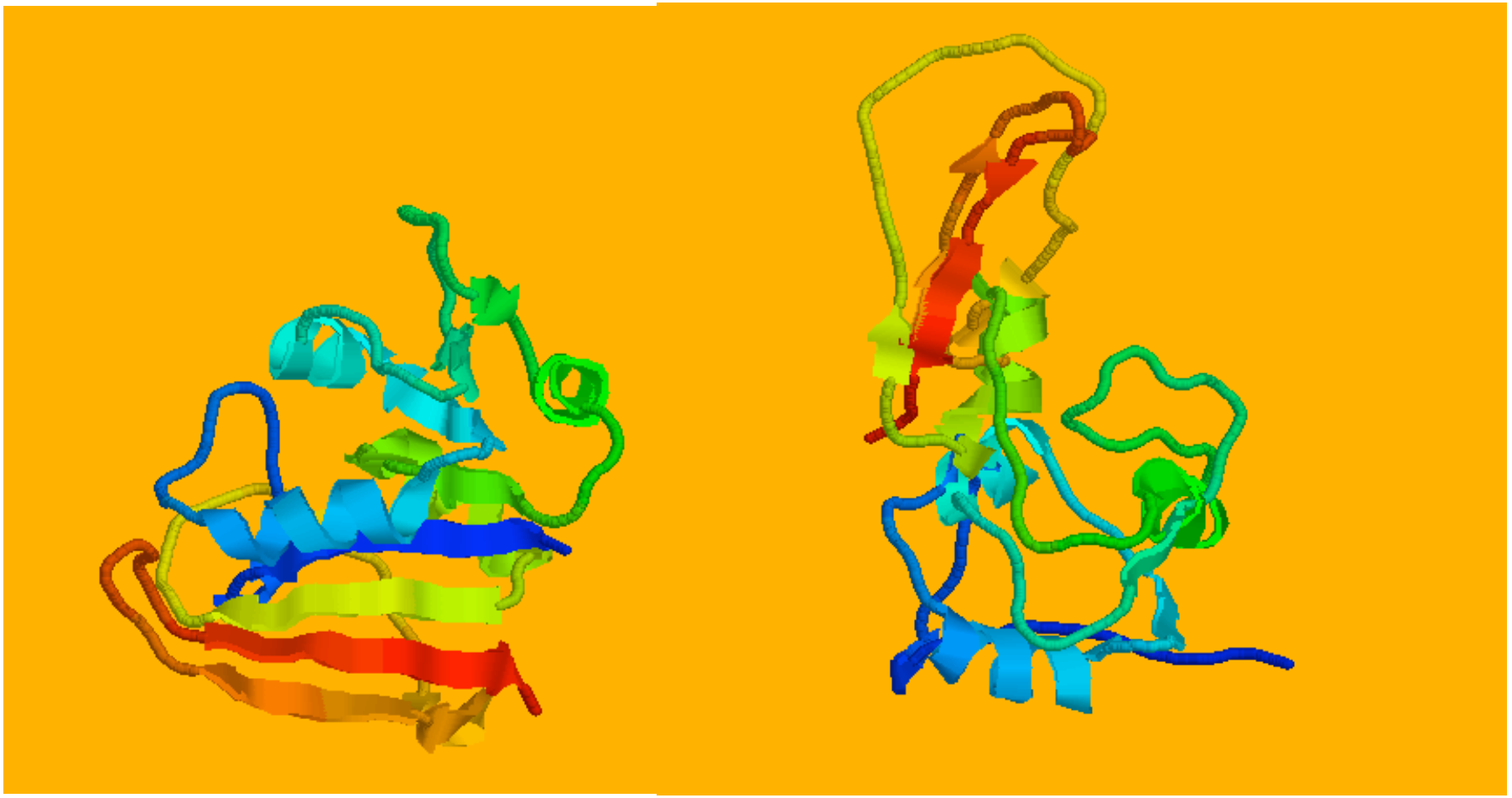
Cartesian coordinates



If there are sufficient constraints, then internals coordinates may be converted to Cartesian coordinates.

If Amerigo Vespucci had mapped a protein...

These two molecules have *identical torsion angles*, and only *slight differences* in backbone bond lengths and bond angles.



Protein structure if it were solved by *John Harrison*

Protein structure if it were solved by *Amerigo Vespucci*. Note local similarity.

The Protein Data Bank (PDB)

Go to **www.pdb.org**

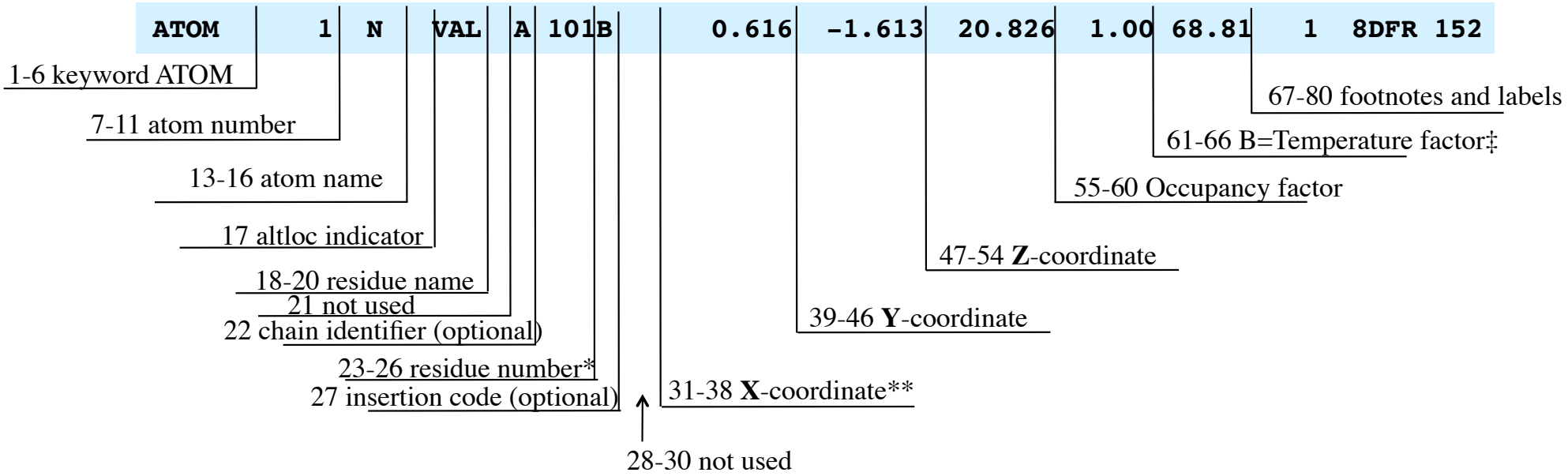
Search for "**1CA2**"

Display PDB file (appears as plain text file)

- HEADER, CMPND, REMARK : reference information.
- HET, FORMUL, HETNAM : ligands, non-standard groups.
- HELIX, SHEET, TURN : secondary structure elements.
- ATOM : coordinates, names, numbering.
- HETATM : coordinates, names, numbering, for HET groups

- There is no *explicit* information about what atoms are bonded to what. (This is determined by distances and atom names.)
- No direct information about the formal or partial charges on atoms. (These are calculated by the force field.)

PDB ATOM lines

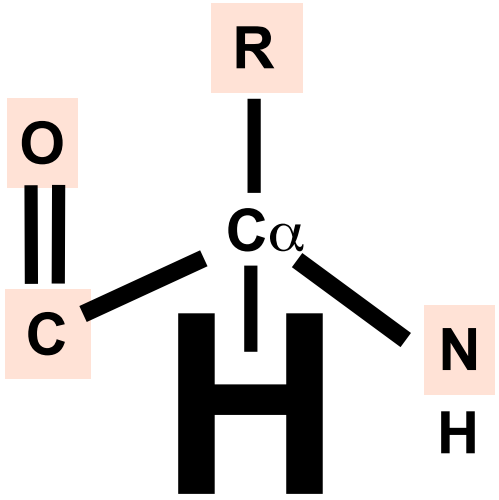


* Usually, but not always, residues are numbered sequentially 1,2,3 etc. Often the numbering starts from a number other than 1.

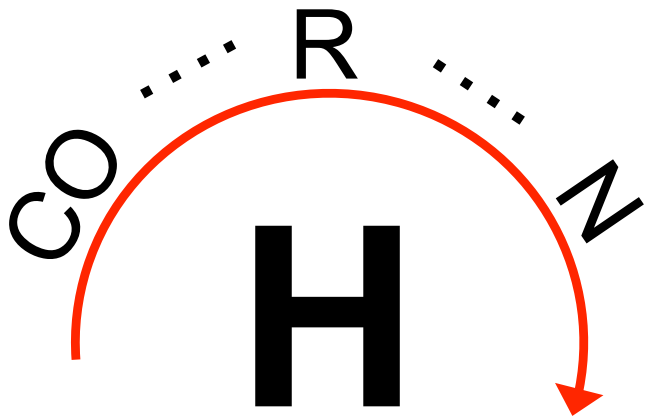
** Coordinates are in orthogonal angstroms by convention. May be converted to crystallographic coordinates using CRYST lines.

‡Mean square displacement $\langle u^2 \rangle$ is proportional to B: $\langle u^2 \rangle = B/(8\pi^2)$

How to check L-amino acid chirality



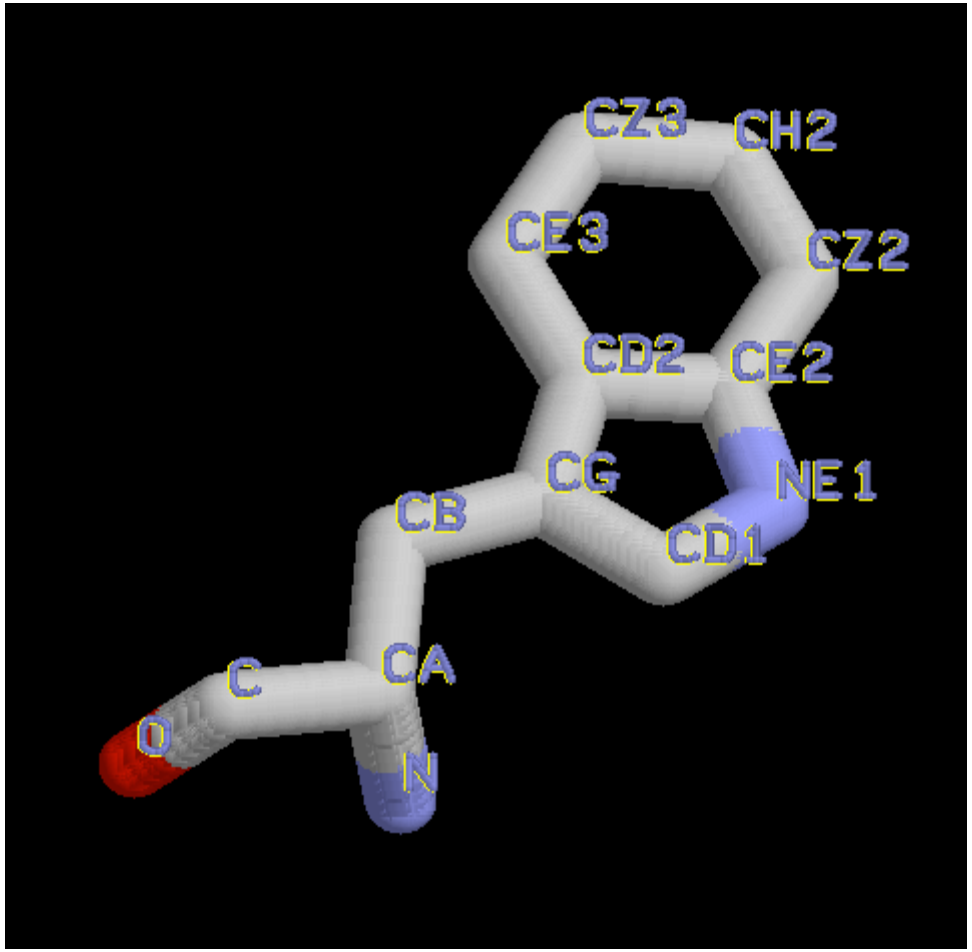
The H is toward you.



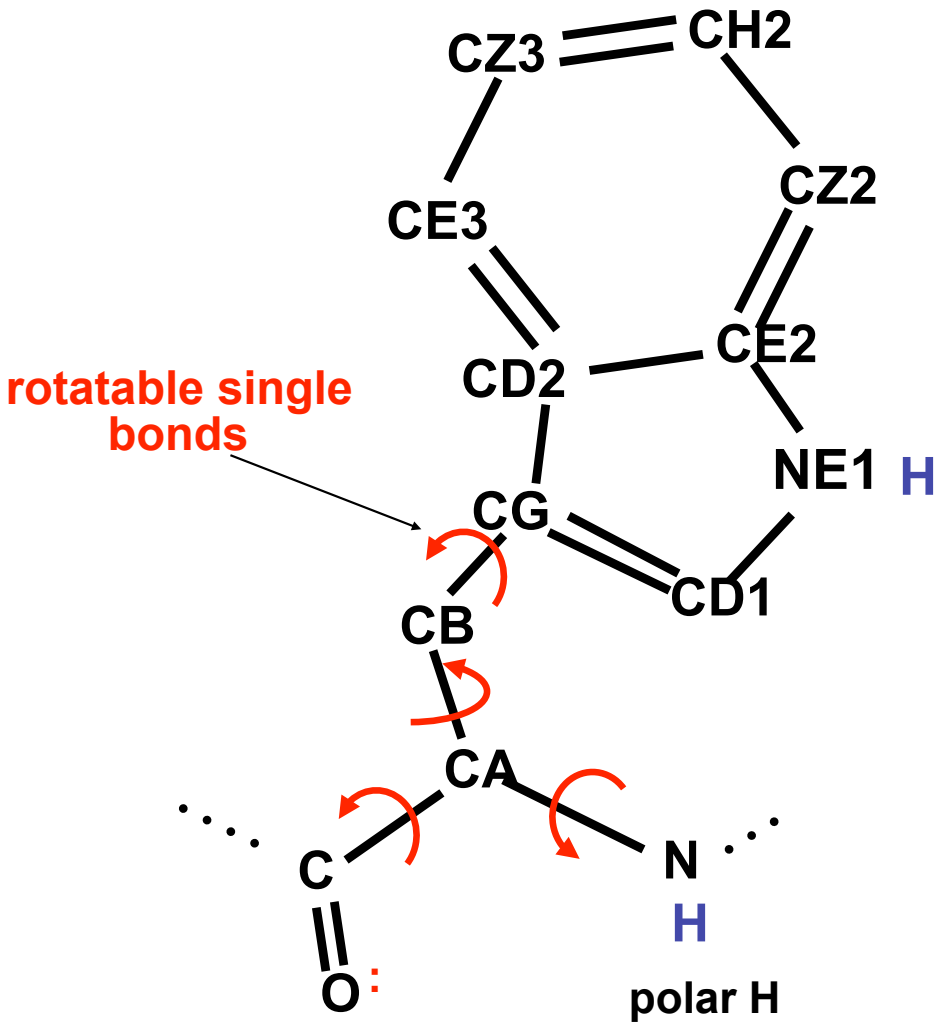
the CORN crib

When an L-amino acid is drawn with the <u>alpha-H forward</u> and the R-group in the back, the letters read clockwise spell "CORN". The "Corn Crib" is a good way to remember which side the R-group (i.e. sidechain) goes on.

atom names: tryptophan



PDB convention atom names follow the formula: <element><greek letter><alt posit>



CA = Alpha Carbon, CB = Beta Carbon, OG = Gamma Oxygen, Delta.., Epsilon.., Zeta..,