THE UNIVERSITY *of* TEXAS

SCHOOL OF HEALTH INFORMATION
SCIENCES AT HOUSTON

X-Ray Crystallography Pt. II

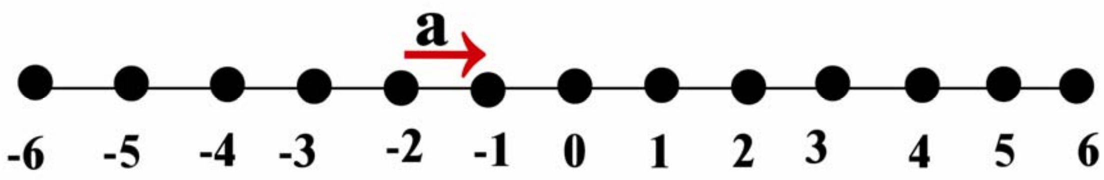
For students of HI 6001-125
“Computational Structural Biology”

Sugoto Chakravarty, Ph.D.
Baylor College of Medicine

http://biomachina.org/courses/structures/03.html

Periodic Array of 2N+1 Identical Atoms

Periodicity a



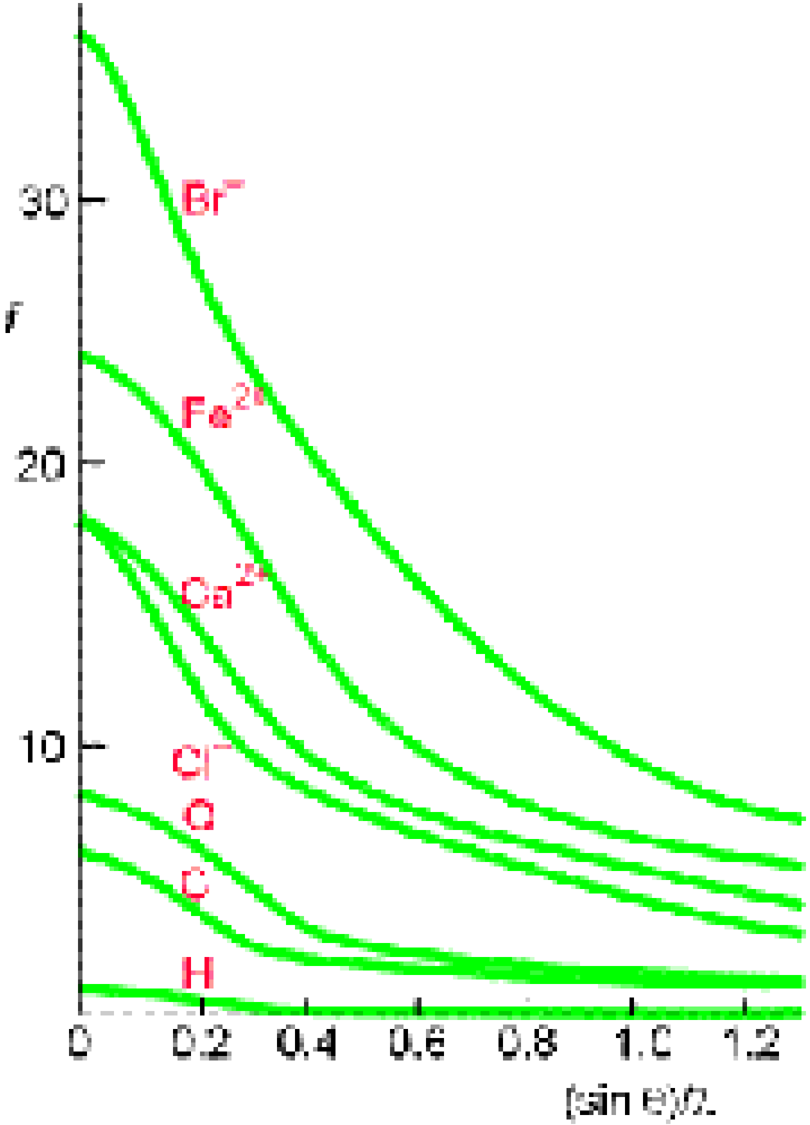
$$F_n(S) = e^{2inS.a}\ F(S)$$

$$\boxed{F_{Tot}(S) = \sum_{n=-N}^{N} F_n(S) = F(S)\sum_{n} e^{2\pi inS.a}}$$

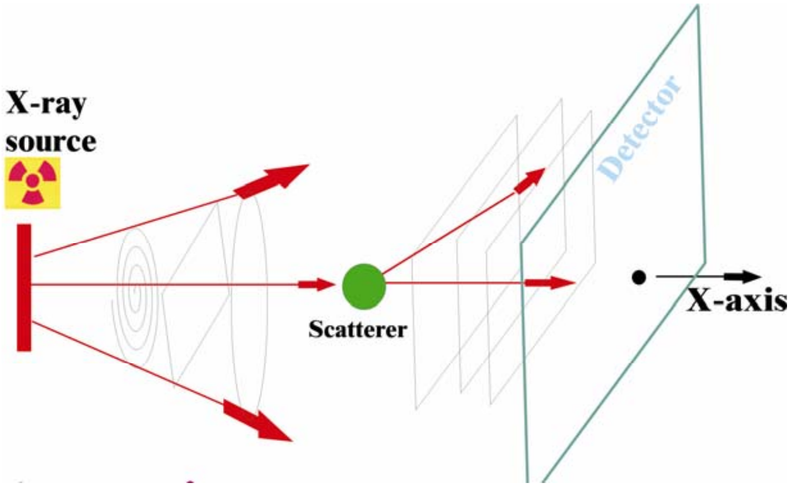
$$\sum_{n=-N}^{N} e^{2\pi inS.a} \longrightarrow \left[f(S)\, e^{-2\pi iNS.a}\frac{\left(1 - e^{2\pi i(2N+1)S.a}\right)}{\left(1 - e^{2\pi iS.a}\right)} \right]$$

Fringe function of the array

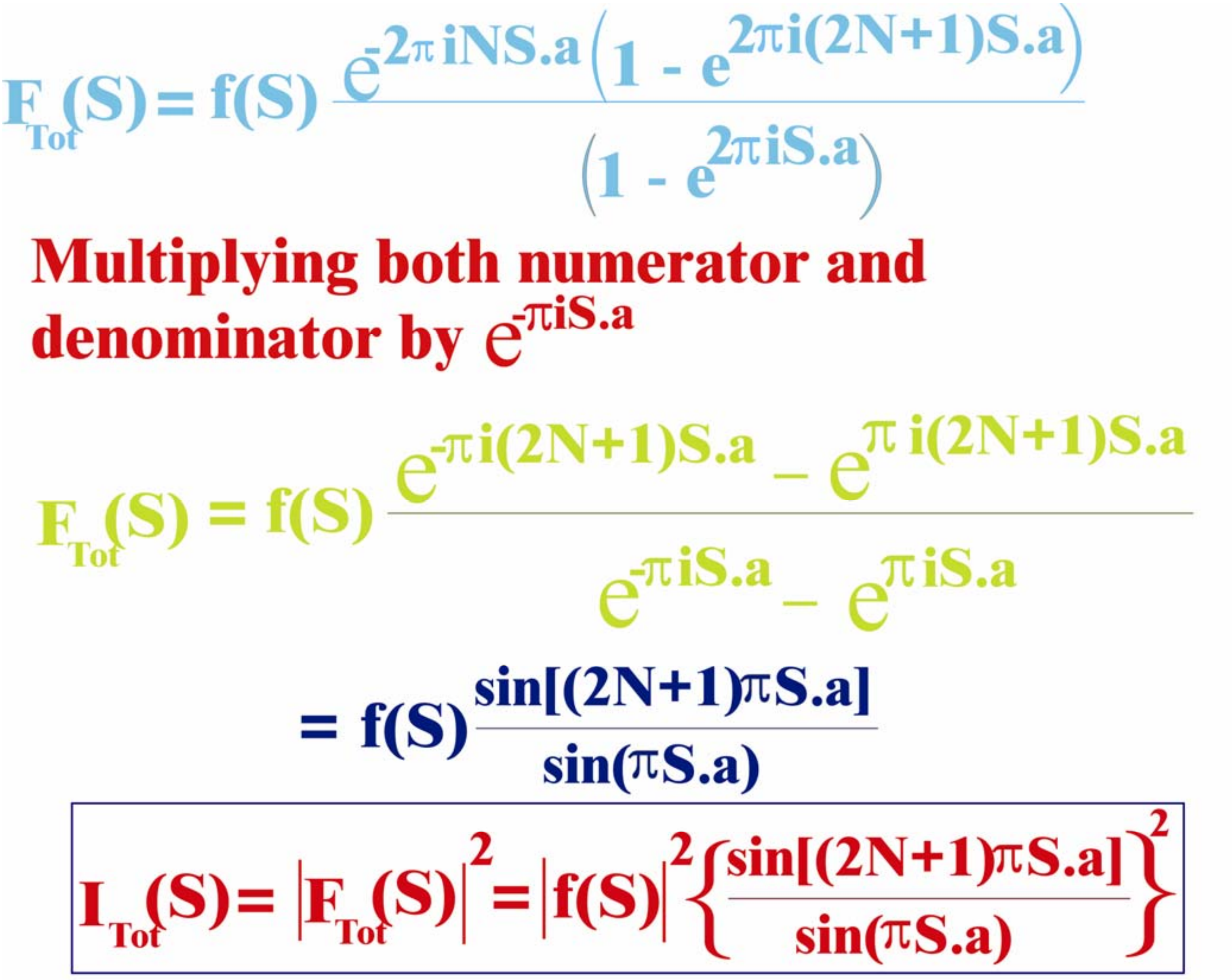
Atomic Form Factors



http://www.public.asu.edu/~awsmith



Periodic Array of 2N+1 Identical Atoms



$$F_{Tot}(S) = f(S)\,\frac{e^{-2\pi iNS.a}\left(1 - e^{2\pi i(2N+1)S.a}\right)}{\left(1 - e^{2\pi iS.a}\right)}$$

Multiplying both numerator and denominator by $e^{-\pi iS.a}$

$$F_{Tot}(S) = f(S)\,\frac{e^{-\pi i(2N+1)S.a} - e^{\pi i(2N+1)S.a}}{e^{-\pi iS.a} - e^{\pi iS.a}}$$

$$= f(S)\,\frac{\sin[(2N+1)\pi S.a]}{\sin(\pi S.a)}$$

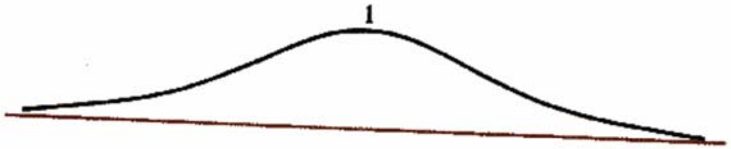
$$\boxed{I_{Tot}(S) = \left|F_{Tot}(S)\right|^2 = \left|f(S)\right|^2 \left\{\frac{\sin[(2N+1)\pi S.a]}{\sin(\pi S.a)}\right\}^2}$$

Fringe Function and Diffraction

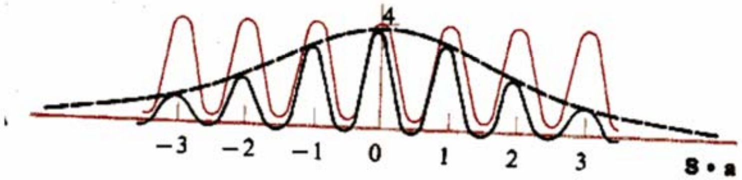
$$I_{Tot}(S) = \left|F_{Tot}(S)\right|^2 = \left|f(S)\right|^2 \boxed{\left\{\frac{\sin[(2N+1)\pi S.a]}{\sin(\pi S.a)}\right\}^2} \longleftarrow$$

Fringe function

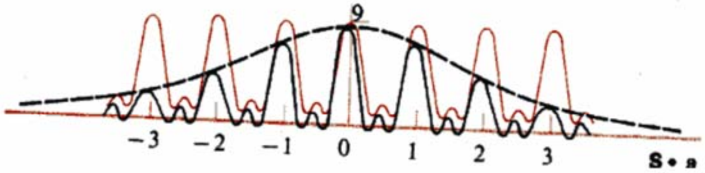
N=1



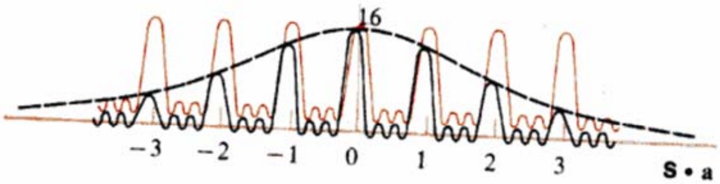
N=4



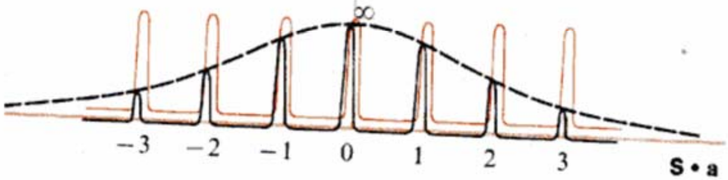
N=9



N=16



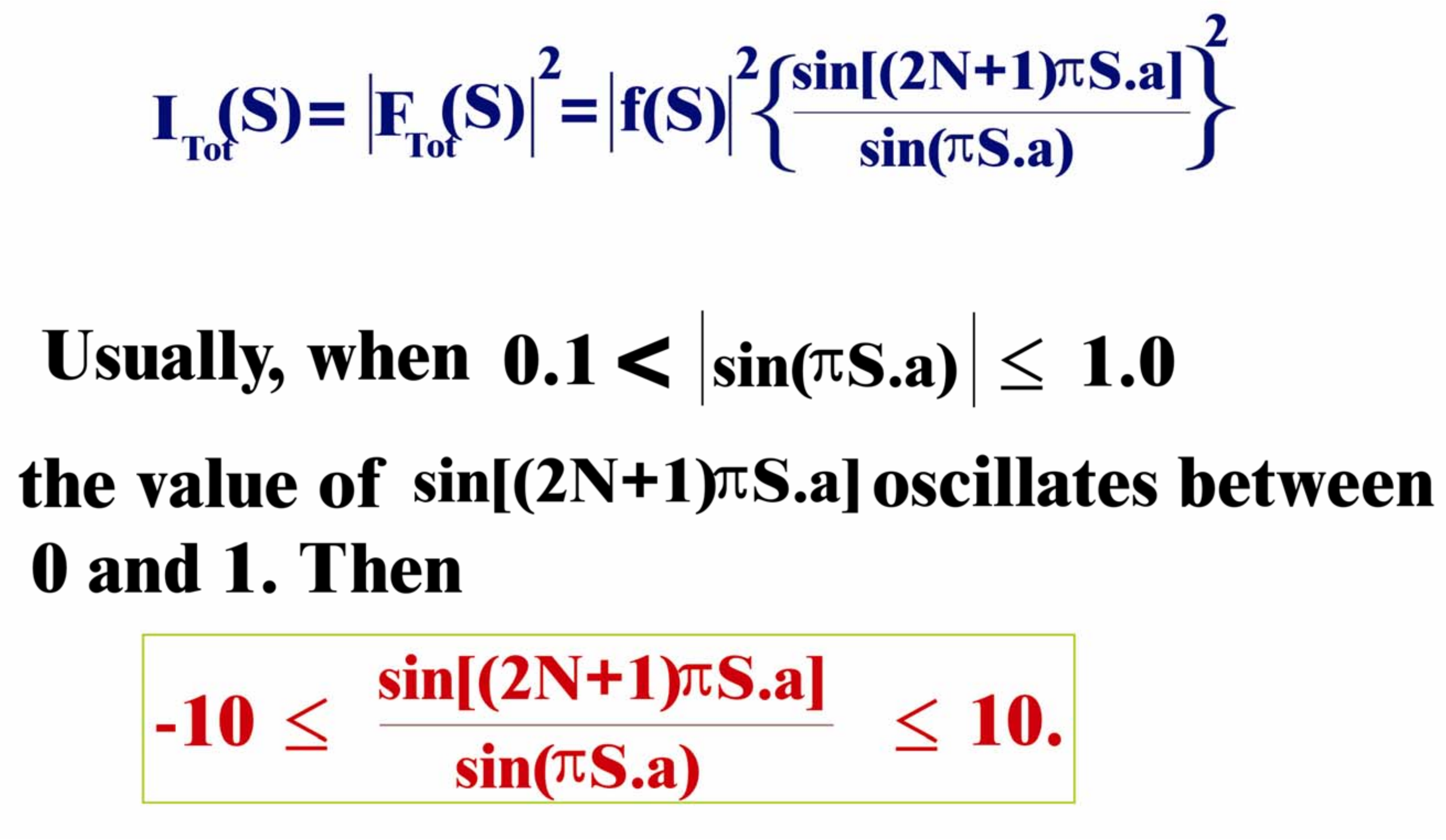
N=∞



Plots:

1. **I = 0 everywhere except integral S.a**
2. **I_{max} only when S.a = 0 i.e. S is in a plane ⊥ to the long atomic axis**

Condition for Diffraction Maximum



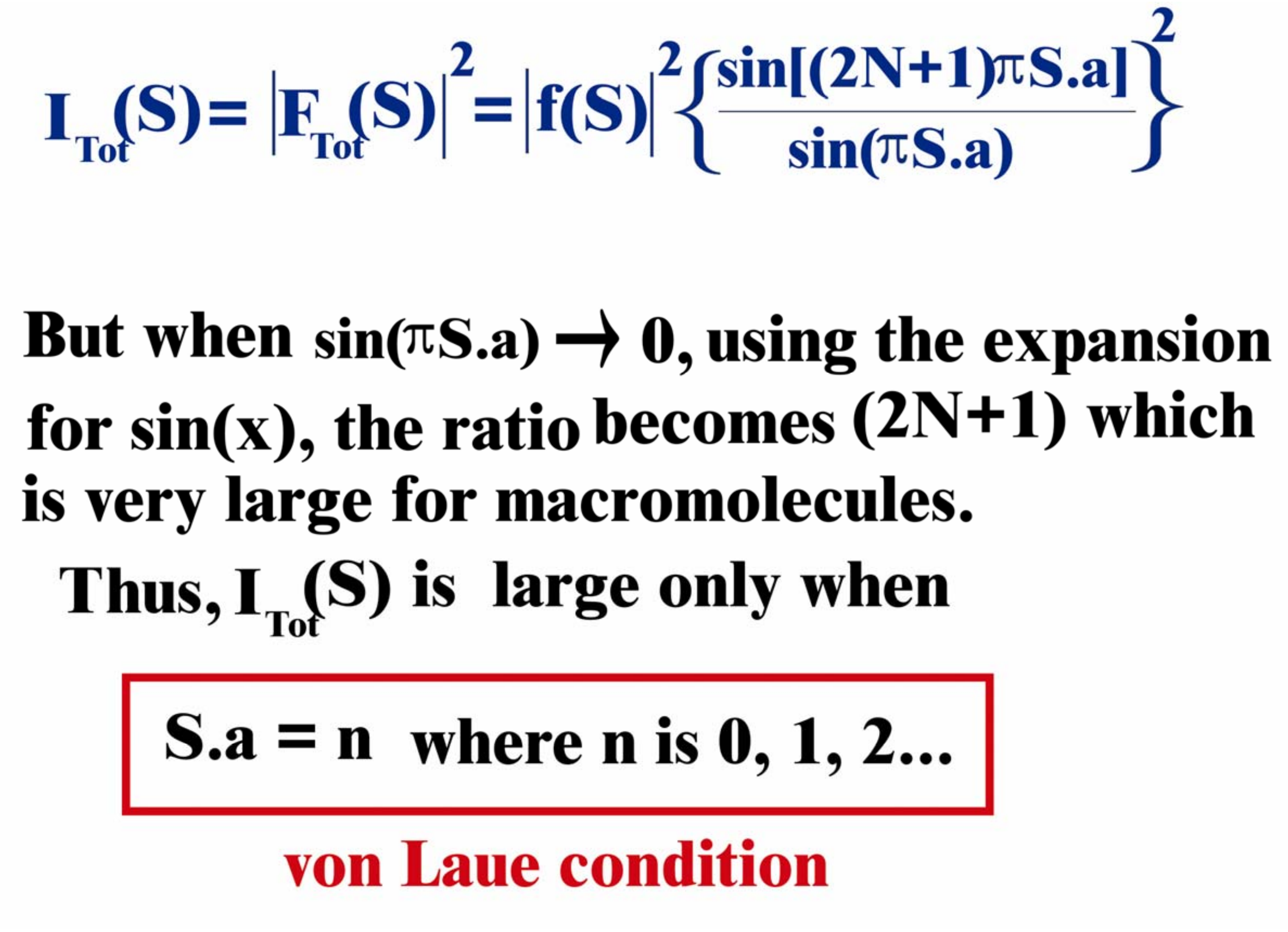
$$I_{Tot}(S) = |F_{Tot}(S)|^2 = |f(S)|^2 \left\{ \frac{\sin[(2N+1)\pi S.a]}{\sin(\pi S.a)} \right\}^2$$

Usually, when $0.1 < |\sin(\pi S.a)| \leq 1.0$

the value of $\sin[(2N+1)\pi S.a]$ **oscillates between 0 and 1. Then**

$$\boxed{-10 \leq \frac{\sin[(2N+1)\pi S.a]}{\sin(\pi S.a)} \leq 10.}$$

Condition for Diffraction Maximum



$$I_{Tot}(S) = \left|F_{Tot}(S)\right|^2 = \left|f(S)\right|^2 \left\{\frac{\sin[(2N+1)\pi S.a]}{\sin(\pi S.a)}\right\}^2$$

But when $\sin(\pi S.a) \rightarrow 0$, using the expansion for sin(x), the ratio becomes (2N+1) which is very large for macromolecules.

Thus, $I_{Tot}(S)$ is large only when

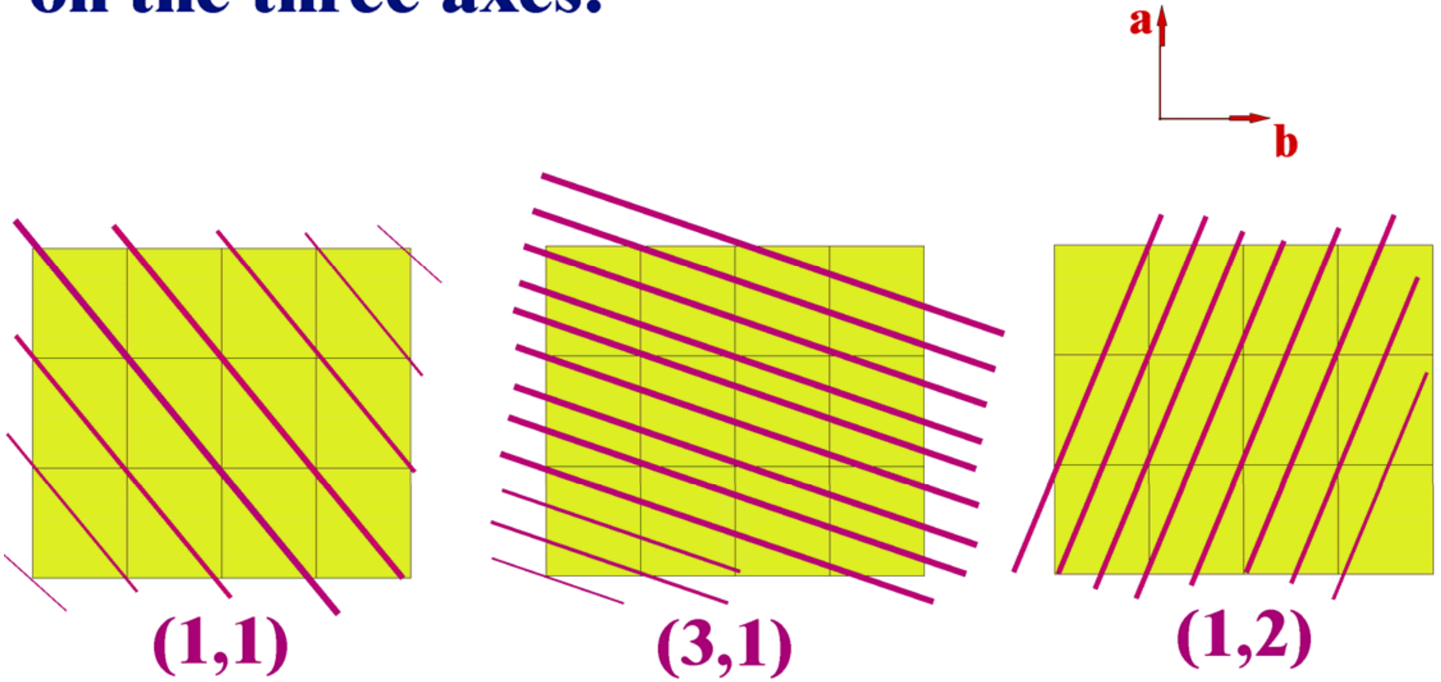
$$\boxed{S.a = n \quad \text{where } n \text{ is } 0, 1, 2\ldots}$$

von Laue condition

Miller Indices in a Lattice

Miller indices (h,k,l)

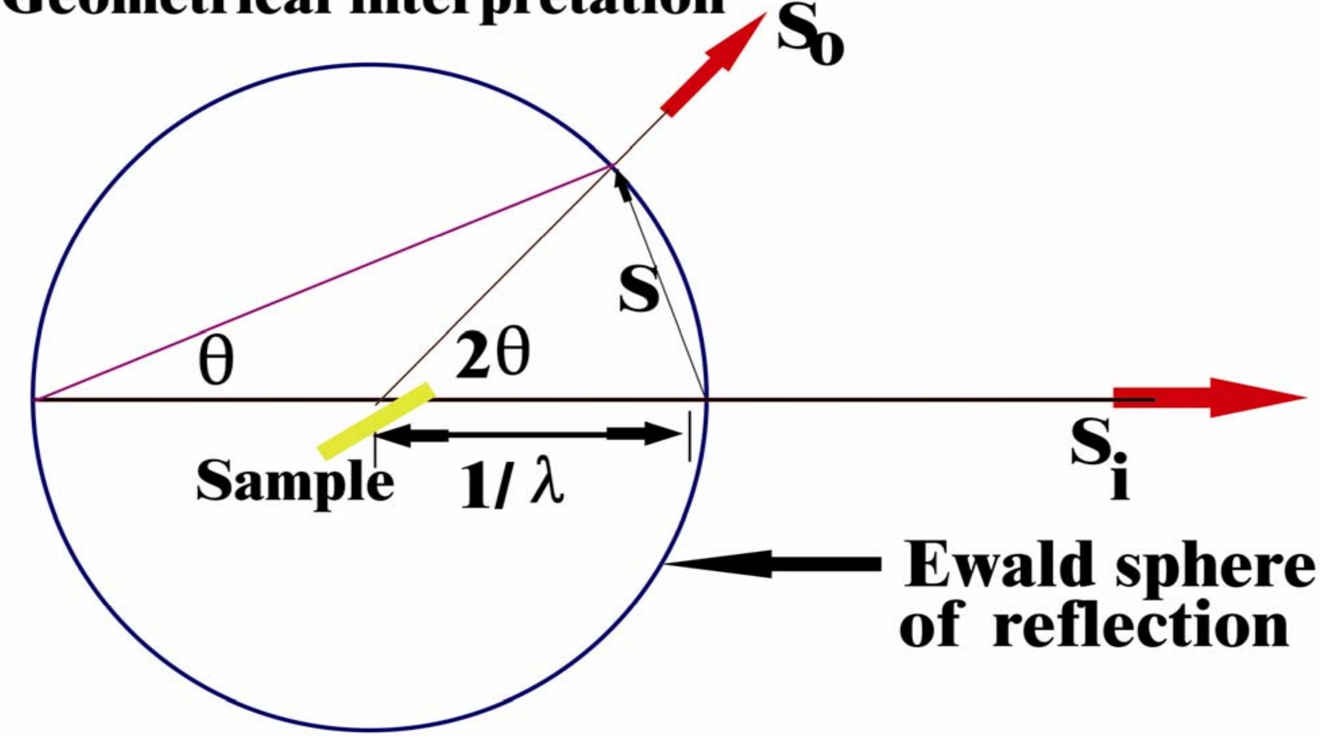
Indices that characterize a set of parallel planes having intercepts a/h, b/k and c/l on the three axes.



Two Conditions for Diffraction

1. $|\mathbf{S}| = 2\,|\sin\theta|/\lambda$

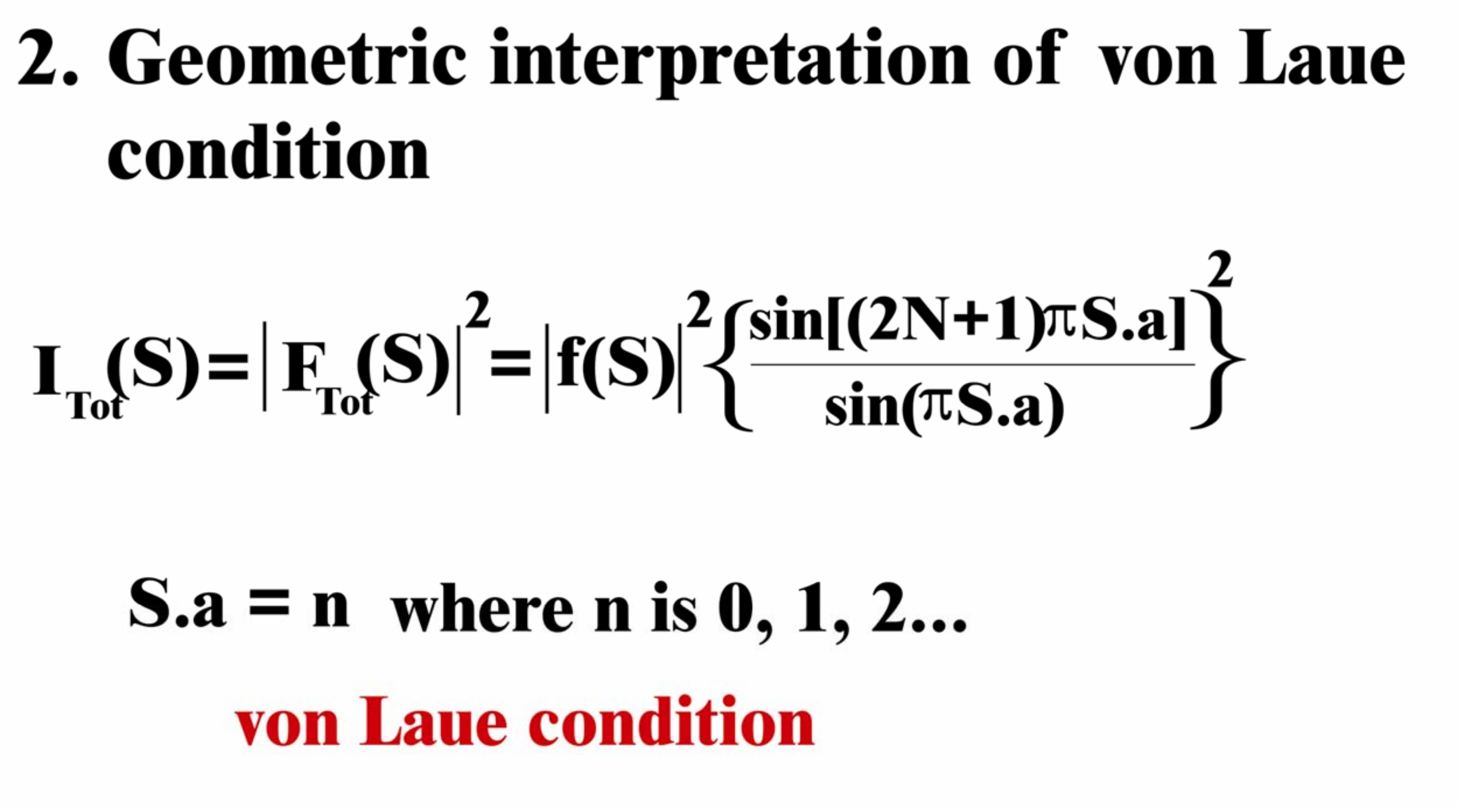
Geometrical interpretation



$$\boxed{\sin\theta = \frac{|\mathbf{S}|}{2/\lambda}}$$

Scattering vectors that satisfy this condition diffract.

Two Conditions for Diffraction



2. Geometric interpretation of von Laue condition

$$I_{Tot}(S)=|F_{Tot}(S)|^2=|f(S)|^2\left\{\frac{\sin[(2N+1)\pi S.a]}{\sin(\pi S.a)}\right\}^2$$

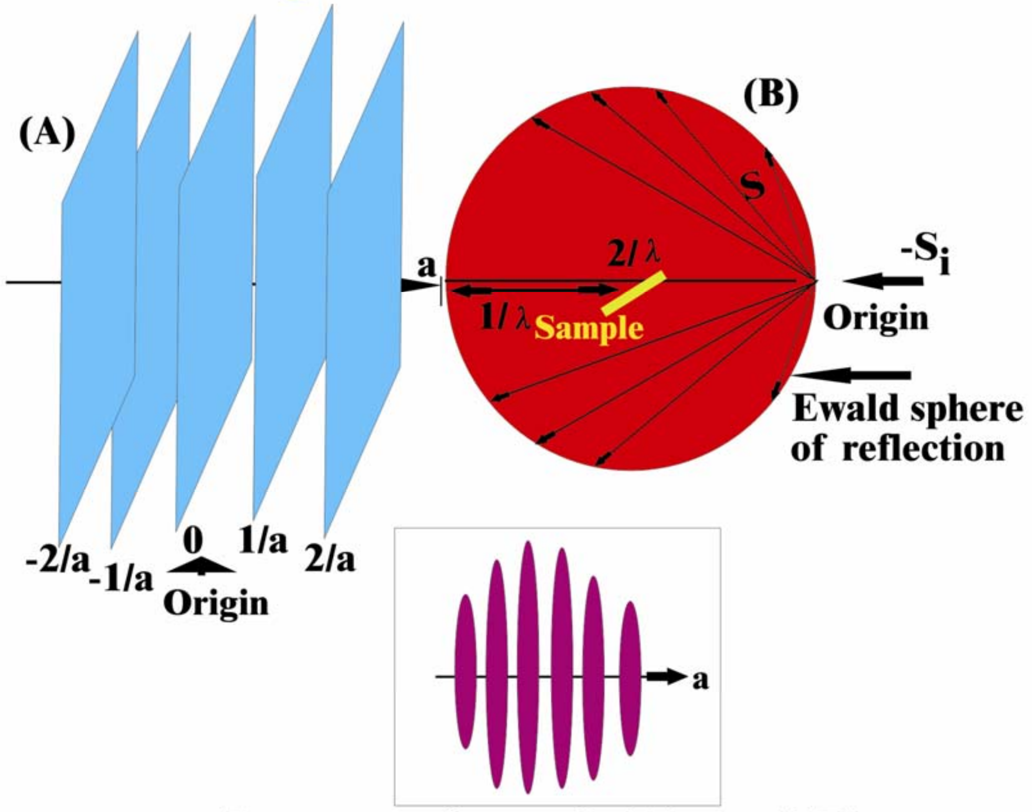
$S.a = n$ where n is 0, 1, 2...

von Laue condition

Visualizing the Two Conditions for Diffraction

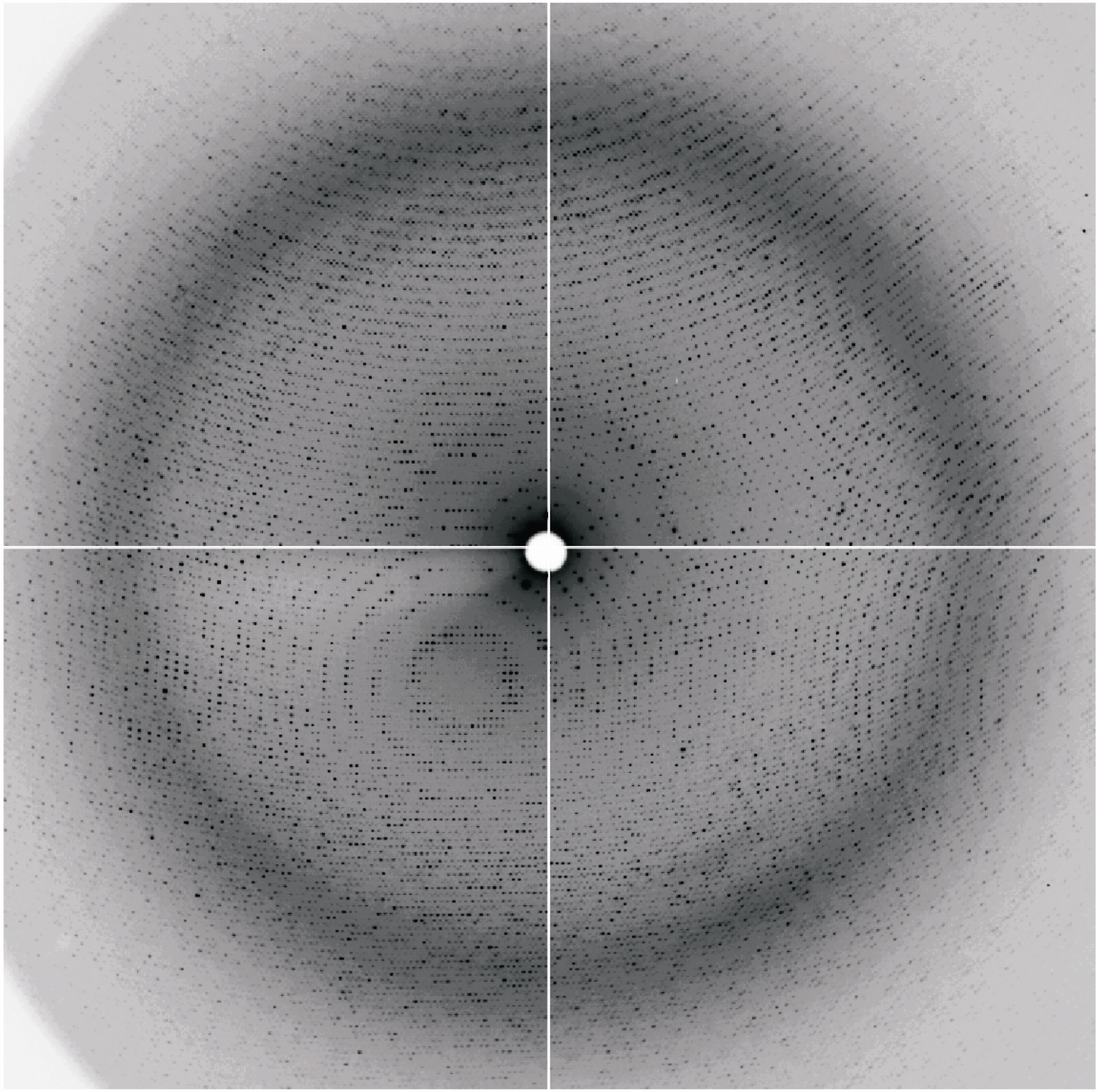
Diffracted intensities are observed only when both the sphere of reflection and the von Laue conditions are satisfied together

Geometric interpretation



Intersection of (A) and(B)

Diffraction Pattern

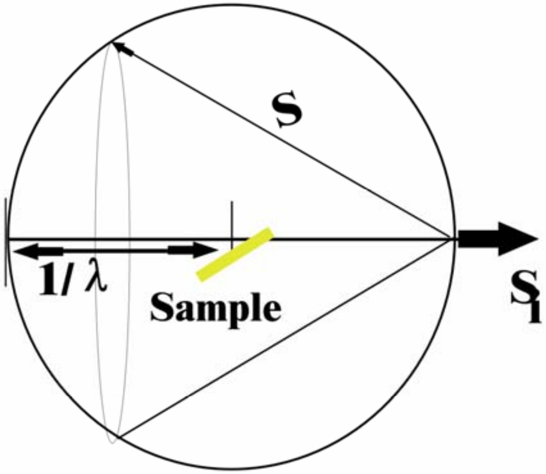


Observable Part of Ewald Sphere

Satisfy both sphere of reflection and von Laue conditions

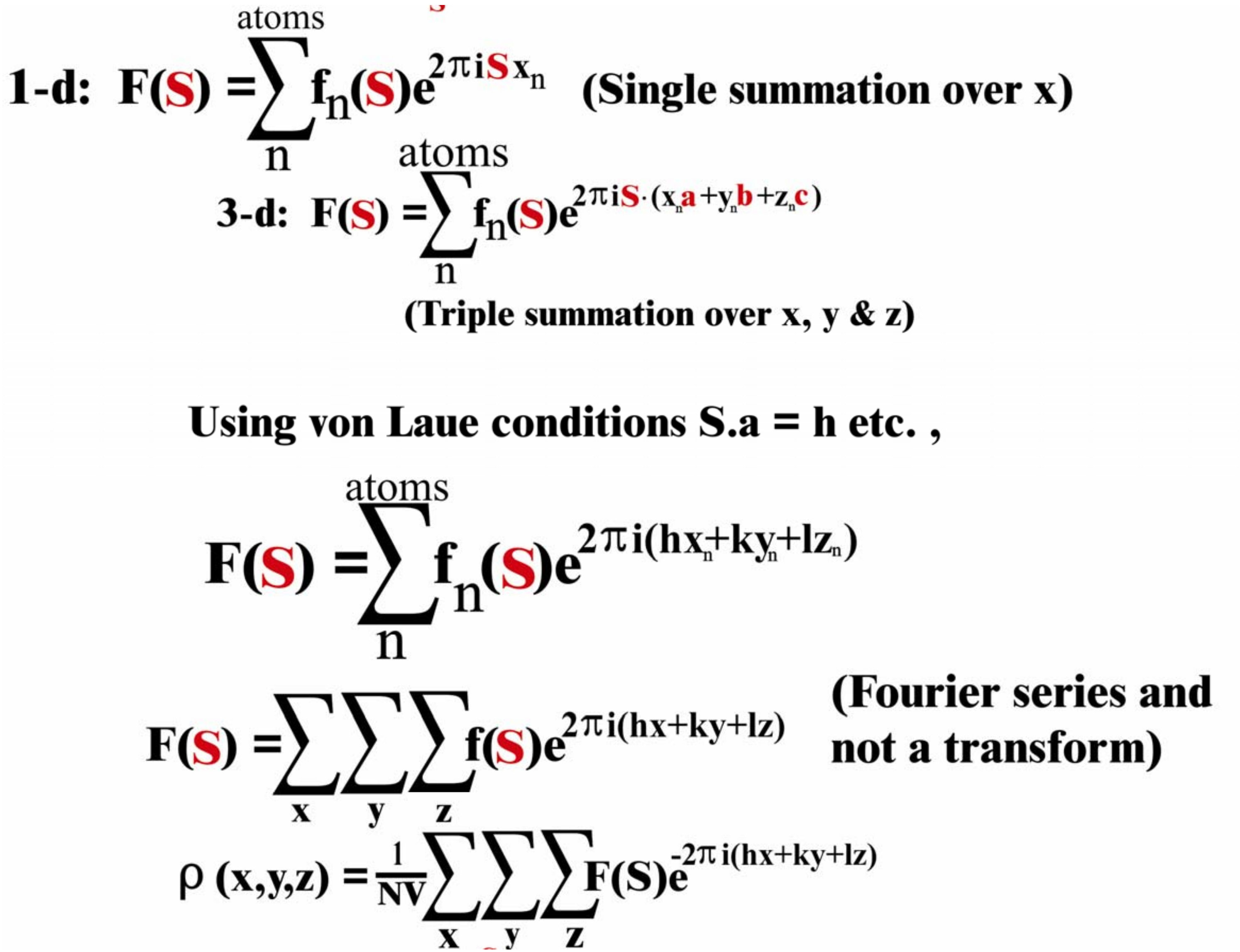
If the wavelength λ and the incident direction S_1 are fixed, only a limited portion of the Ewald sphere can diffract. Limiting sphere: $2/\lambda$

To increase the diffracting region, the incident direction and/or the wavelength needs to be changed.



Leads to many experimental data collection strategies.

Scattering in 3 Dimensions



1-d: $F(\mathbf{S}) = \sum_{n}^{\text{atoms}} f_n(\mathbf{S}) e^{2\pi i \mathbf{S} x_n}$ **(Single summation over x)**

3-d: $F(\mathbf{S}) = \sum_{n}^{\text{atoms}} f_n(\mathbf{S}) e^{2\pi i \mathbf{S}\cdot(x_n \mathbf{a} + y_n \mathbf{b} + z_n \mathbf{c})}$

(Triple summation over x, y & z)

Using von Laue conditions S.a = h etc. ,

$$F(\mathbf{S}) = \sum_{n}^{\text{atoms}} f_n(\mathbf{S}) e^{2\pi i(hx_n+ky_n+lz_n)}$$

$$F(\mathbf{S}) = \sum_{x}\sum_{y}\sum_{z} f(\mathbf{S}) e^{2\pi i(hx+ky+lz)}$$

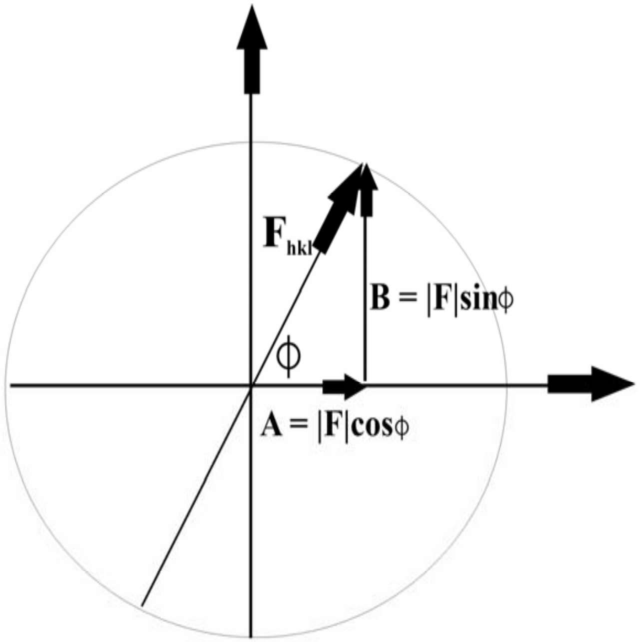
(Fourier series and not a transform)

$$\rho\,(x,y,z) = \frac{1}{NV}\sum_{x}\sum_{y}\sum_{z} F(S) e^{-2\pi i(hx+ky+lz)}$$

Argand Diagram of Structure Factors

$$|F| = \left[A^2 + B^2\right]^{1/2}$$ **: Measured experimentally**

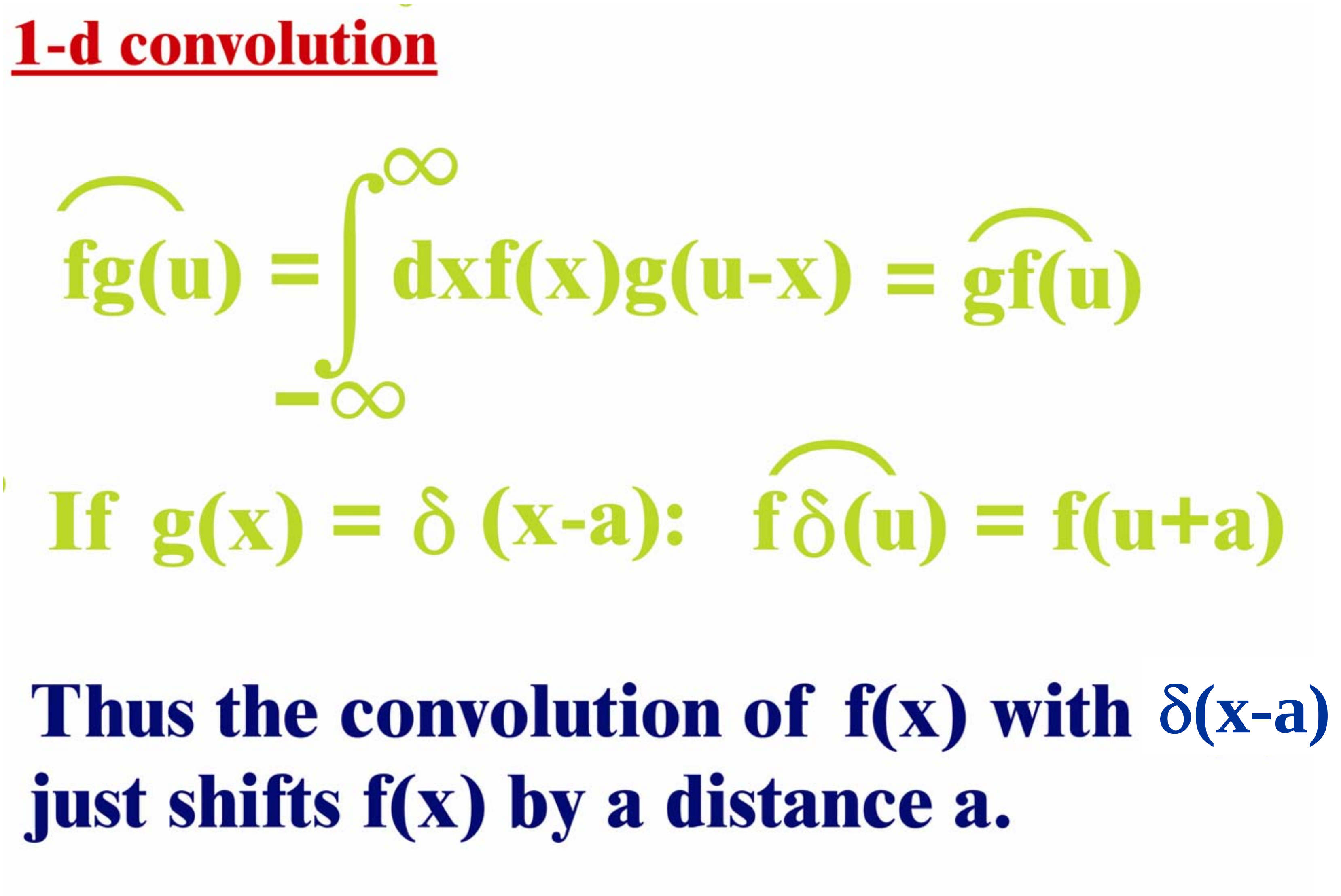
$$\phi = \tan^{-1}(B/A)$$ **: Unknown**



The circle indicates that F is a vector quantity with an amplitude and a phase

Signs of both A and B important in determining the magnitude and quadrant of ϕ

Crystal Lattice and Convolution



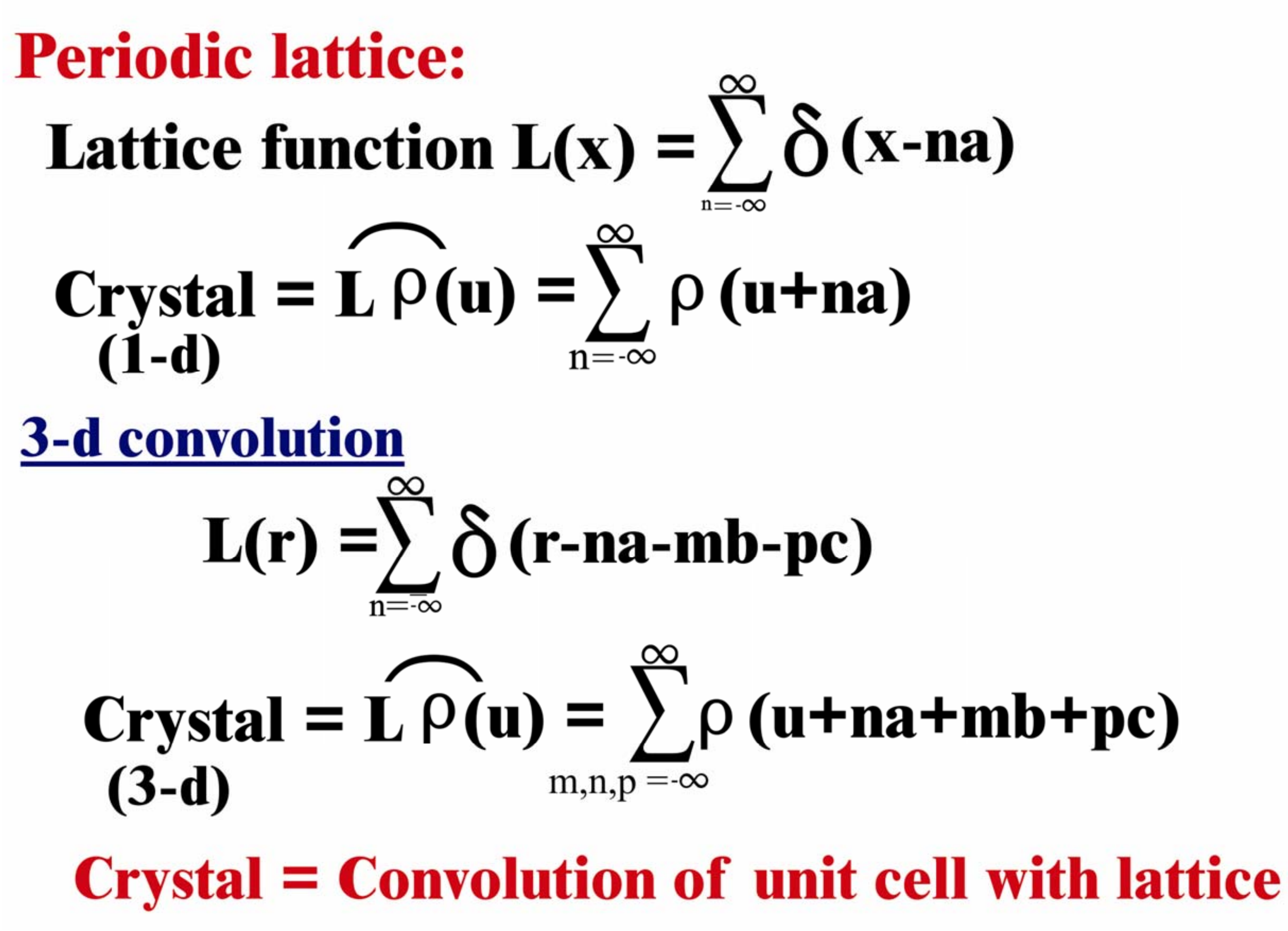
1-d convolution

$$\widehat{fg(u)} = \int_{-\infty}^{\infty} dx f(x) g(u-x) = \widehat{gf(u)}$$

If $g(x) = \delta\,(x\text{-}a)$: $\widehat{f\delta(u)} = f(u+a)$

Thus the convolution of f(x) with $\delta(x\text{-}a)$ **just shifts f(x) by a distance a.**

Crystal Lattice and Convolution



Periodic lattice:

Lattice function $$L(x) = \sum_{n=-\infty}^{\infty} \delta(x-na)$$

Crystal $$= \widehat{L\,\rho}(u) = \sum_{n=-\infty}^{\infty} \rho(u+na)$$

(1-d)

3-d convolution

$$L(r) = \sum_{n=-\infty}^{\infty} \delta(r-na-mb-pc)$$

Crystal $$= \widehat{L\,\rho}(u) = \sum_{m,n,p=-\infty}^{\infty} \rho(u+na+mb+pc)$$

(3-d)

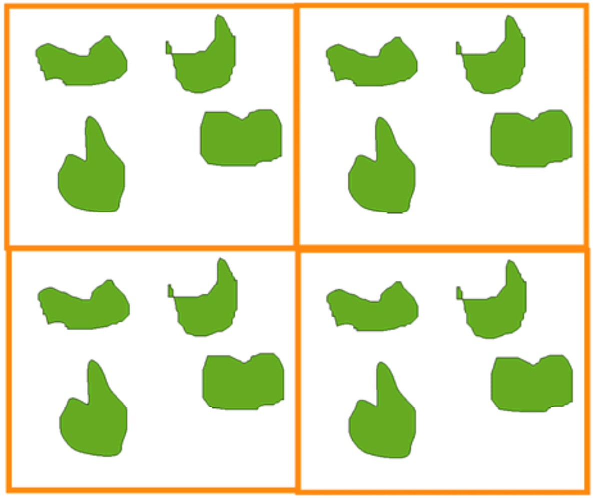
Crystal = Convolution of unit cell with lattice

Unit Cell and Symmetry

Lattice: repetition of unit cells by pure translation

Contents of the unit cell:

Contents cannot be arbitrarily arranged but there must be an asymmetric unit that is rotated (and possibly fractionally translated) to generate the complete unit cell contents



2-d lattice

Possible rotational symmetry among the asymmetric units

1, 2, 3, 4, 6-fold: 5-fold not allowed by translational symmetry

Mirror reflection not allowed in biological molecules

Fractional translations: ½, 1/3, 2/3, 1/4, 3/4, 1/6,.....
of the unit cell dimensions

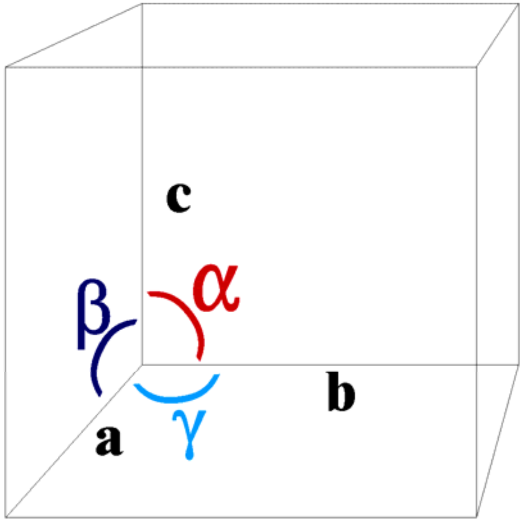
Unit Cell Types Determined by Symmetry

Point groups: All possible rotations and reflections among the asymmetric units

32 Point groups

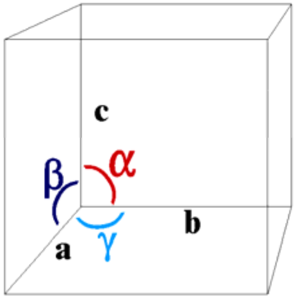
Crystal systems: Point groups + translation symmetry restricts types of the unit cell possible (7 crystal systems)

Restrictions on the unit cell lengths and angles

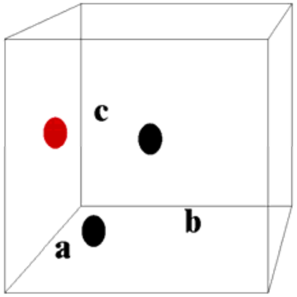


7 Crystal Systems and 14 Bravais Lattices

System	Lattice	Min symmetry	Unit cell
Triclinic	P	None (1-fold)	$a \neq b \neq c$ $\alpha \neq \beta \neq \gamma$
Monoclinic	P	2-fold along b	$a \neq b \neq c$ $\alpha \neq \gamma$ $\beta \neq 90$
Orthorhombic	P, C, I, F	2-folds along a,b,c	$a \neq b \neq c$
Tetragonal	P,I	4-fold along c	$a = b \neq c$
Trigonal/ Rhombohedral	R,P	3-fold along c	$a = b = c$ $\alpha = \beta = \gamma \neq 90$
Hexagonal	P	6-fold along c	$a = b \neq c$ $\alpha = \beta = 90$ $\gamma = 120$
Cubic	P,I,F	3-fold along body diagonals	$a = b = c$

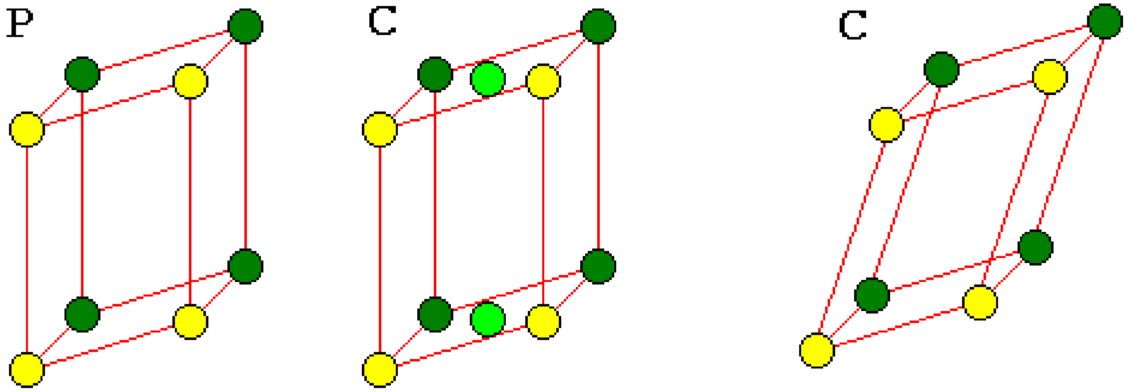


Centering

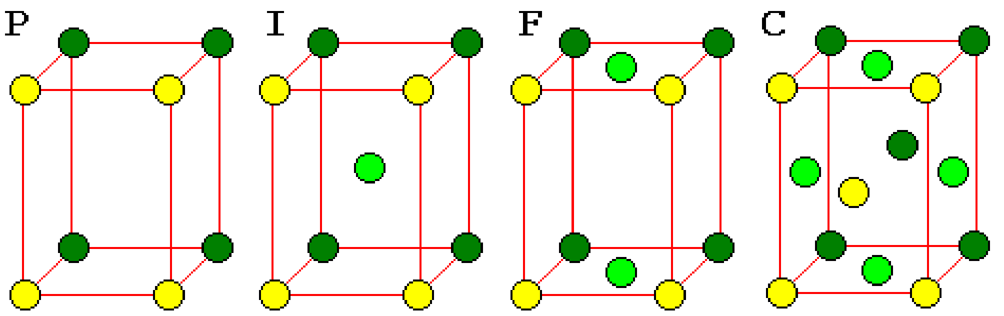


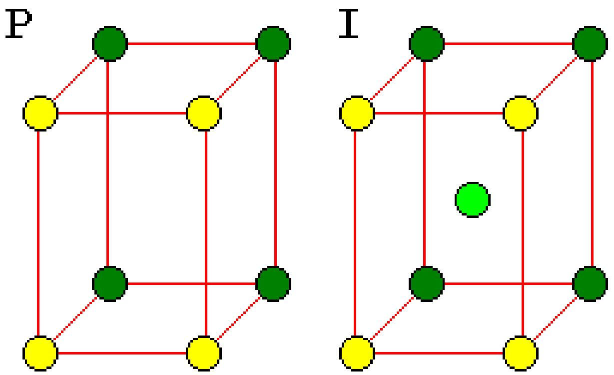
Different Lattices and Centering

Triclinic & Monoclinic



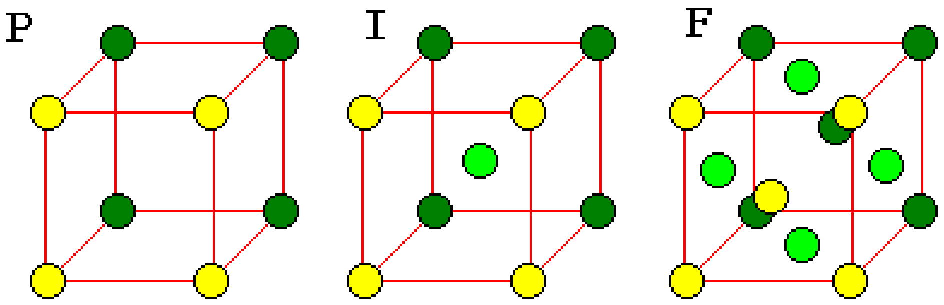
Orthorhombic





Tetragonal

Cubic

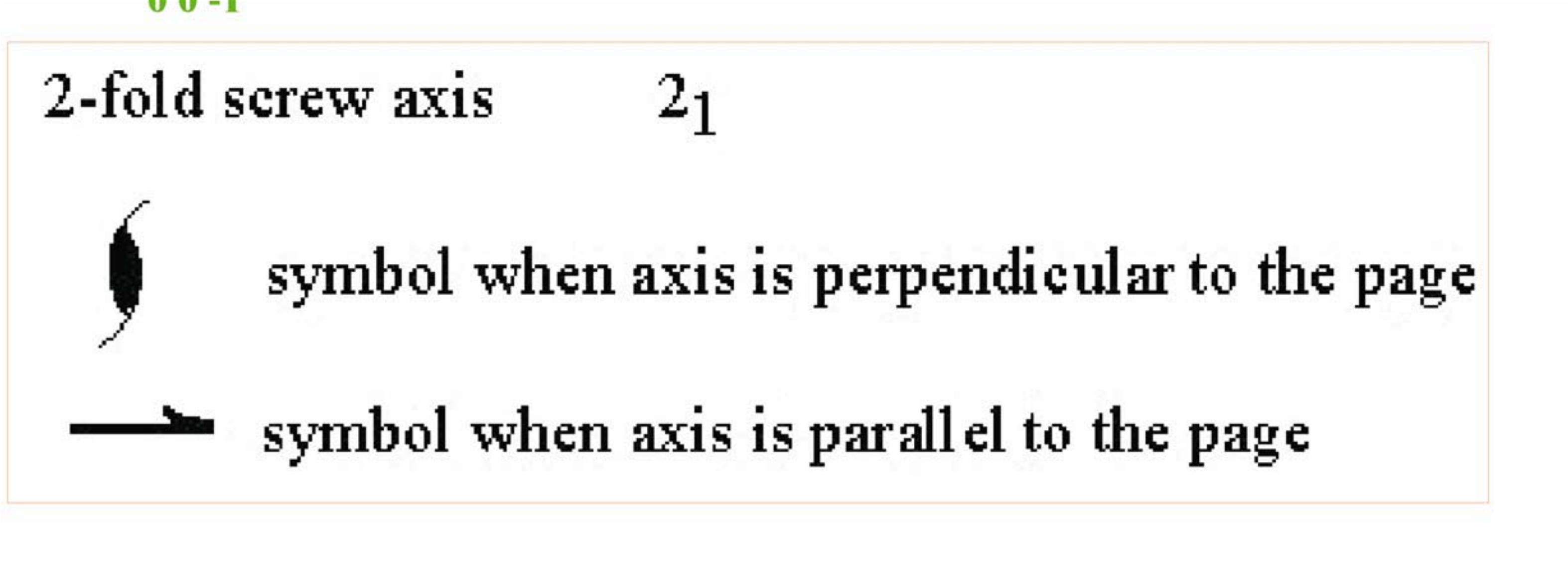


© http://www.uwgb.edu/dutchs/symmetry/bravais.htm

Applying a Screw Operation

2-fold through origin:

$\begin{matrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{matrix}$ x,y,z x, -y, -z



2-fold screw axis 2_1

symbol when axis is perpendicular to the page

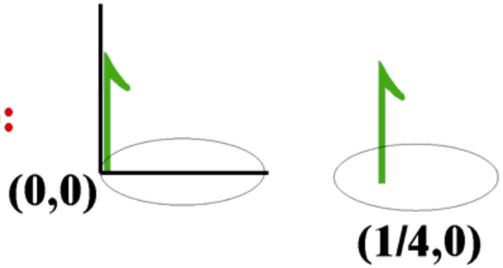
symbol when axis is parallel to the page

2_1 along [001] and passing through origin:

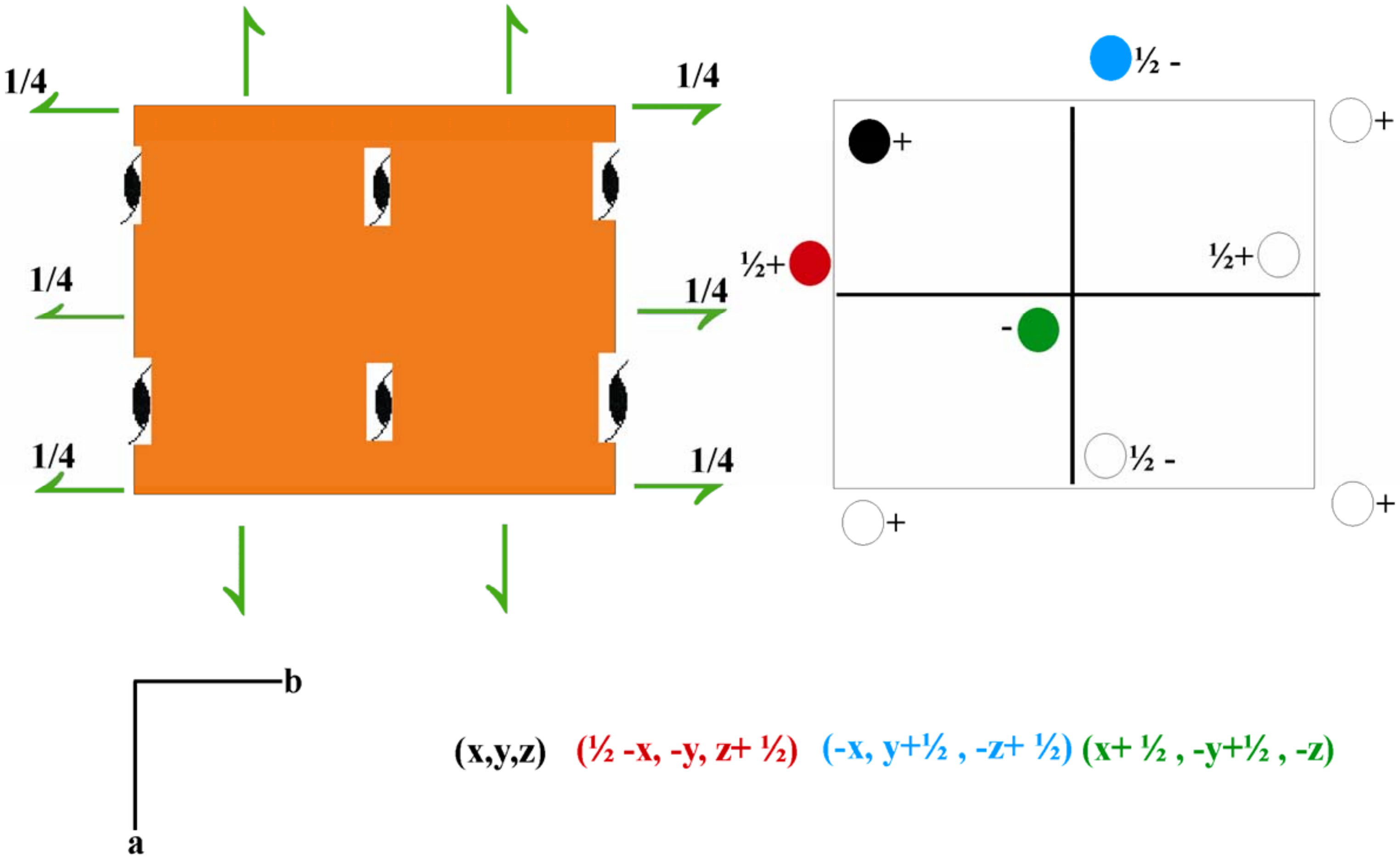
(x, y, z) (-x -y ½ +z)

2_1 along [001] and passing through (1/4, 0, 0):

(x, y, z) (½ -x -y ½ +z)



Example of a Space Group

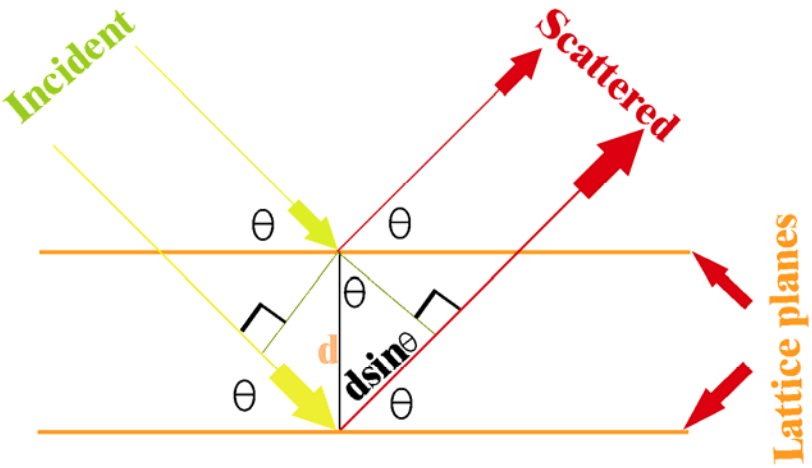


(x,y,z) (½ -x, -y, z+ ½) (-x, y+½ , -z+ ½) (x+ ½ , -y+½ , -z)

$P2_12_12_1$ from International tables (#19)

Bragg’s Law of Diffraction

The atoms in a crystal can be considered as a series of parallel planes.



To observe diffraction, path difference between reflected beams from adjacent planes must be an integral number of wavelengths

$$\boxed{2d\sin\theta = n\lambda}$$

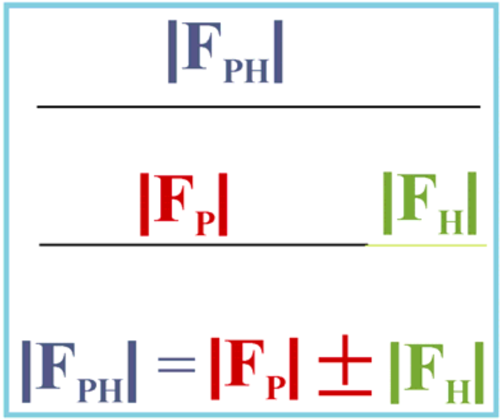
Phase Solution

Aim: Determine ϕ_p for each reflection

- **Isomorphous replacement**
- **Anomalous scattering**
- **Molecular replacement**

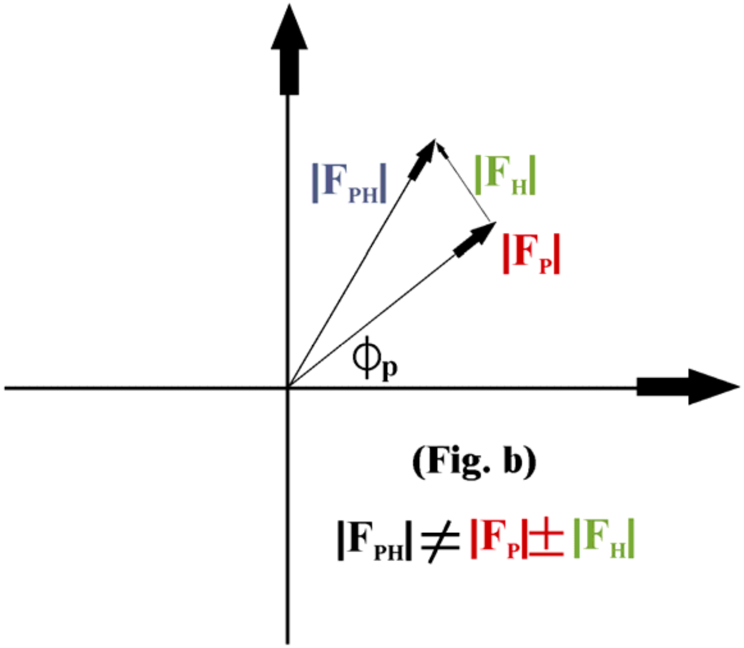
(Note: A.s. not explained here, similar to I.r., but using special wavelengths to break Friedel symmetry, see http://www.bmsc.washington.edu/scatter/AS_tutorial.html)

Isomorphous Replacement



$$|F_{PH}| = |F_P| \pm |F_H|$$

(Fig. a)



(Fig. b)

$$|F_{PH}| \neq |F_P| \pm |F_H|$$

Soak in heavy atoms into the crystal

If the derivatized structure remains similar, hope to get vector relationships between F_P, F_H and F_{PH}

Simple collinear relationship between the structure factors holds for only centro-symmetric reflections whose phases are 0 or π (Fig. a)

No simple relationship between the structure factors for the general non-centrosymmetric reflections (Fig. b)

Patterson Functions to Locate Heavy Atoms

$$\rho(\mathbf{r}) = \int \mathbf{F(S)}e^{-2\pi i\mathbf{S}.\mathbf{r}}\, d\mathbf{r}$$

$$\boxed{\mathbf{P} = \int \mathbf{I(S)}\, e^{-2\pi i\mathbf{S}.\mathbf{r}}\, d\mathbf{r}}$$

$$= \int \mathbf{F^{*}(S)F(S)}e^{-2\pi i\mathbf{S}.\mathbf{r}}\, d\mathbf{r}$$

$$= \overparen{\rho(\mathbf{r})\rho(-\mathbf{r})}$$

The Patterson is a convolution of the electron density with its image that is inverted through the origin

Patterson are Inter-Atomic Distance Maps

$$\mathbf{I}_h \propto |\mathbf{F}_h|^2 = \sum_{j}^{N} f_j e^{2\pi i h \cdot r_j} \sum_{k}^{N} f_k e^{-2\pi i h \cdot r_k}$$

$$= \sum_{j}^{N} f_j f_k e^{2\pi i h \cdot (r_j - r_k)}$$

Thus, the Patterson functions, computed using the intensities as coefficients, map the inter-atomic vectors

In real space: $\mathbf{P} = \int_{\text{unit cell}} \rho(\mathbf{r})\,\rho(\mathbf{r}+\mathbf{u})\,d^3\mathbf{r}$

Pattersons Determine Heavy Atom Vectors

For large N, Patterson maps become uninterpretable

Orthorhombic $P2_12_12_1$:

(x,y,z); (½-x,-y,1/2+z); (½+x,1/2-y,-z); (-x,1/2+y, ½-z)

Difference Patterson:
Coefficients: $|F_D|^2 - |F_N|^2$

Peaks at inter-atomic vectors U,V,W:

U	V	W	Peak type
0.404	0.453	0.5	Harker
0.5	0.048	0.199	Harker
0.42	0.108	0.986	General (NH)

Patterson peaks:

½- 2x	-2y	½ (H)
½	½-2y	-2z (H)
2x	½	-½-2z (H)
-2x	½	½-2z (H)
-½	½+2y	-2z (H)
-½ -2x	2y	½ (H)

H: Harker section peak

Solve for x,y & z using the Harker peaks:

x = 0.048 ; y = 0.774 ; z = -0.099 = 0.900 (translate by 1.0)

These are the heavy atom coordinates

Patterson Heavy Atom Peaks May Solve Structure

Hypothetical 2 molecule structure:
(10 Carbon atoms + 1 Br atom) / molecule

Patterson peak ht $\propto Z_A Z_B$

Br-Br peaks will be the highest

Determine the coordinate of Br

$$\boxed{F_h = F_h^{known} + F_h^{unknown}} \longleftarrow \text{(Due to rest of structure)}$$

(due to the heavy atoms in the unit cell)

Compute partial electron density map, fill up the missing density and iterate

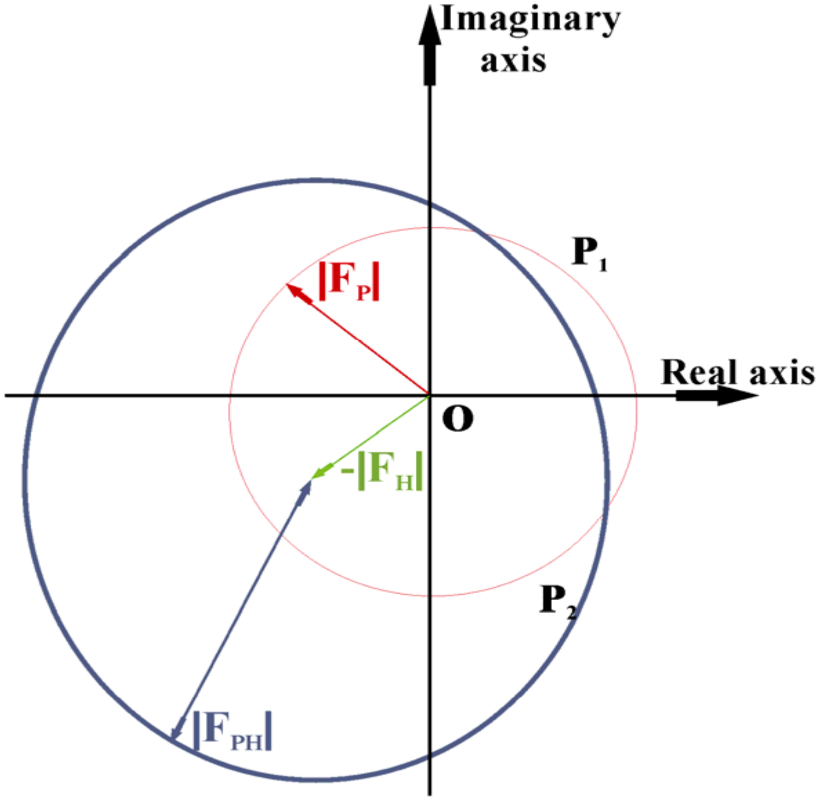
Single Isomorphous Replacement (SIR)

For a given reflection hkl (Vector addition of Sfs)

Known: $|F_P|$, $|F_H|$ and $|F_{PH}|$

$|F_P|$ circle center: Origin O

$|F_H|$: contribution of heavy atom to structure factor amplitude

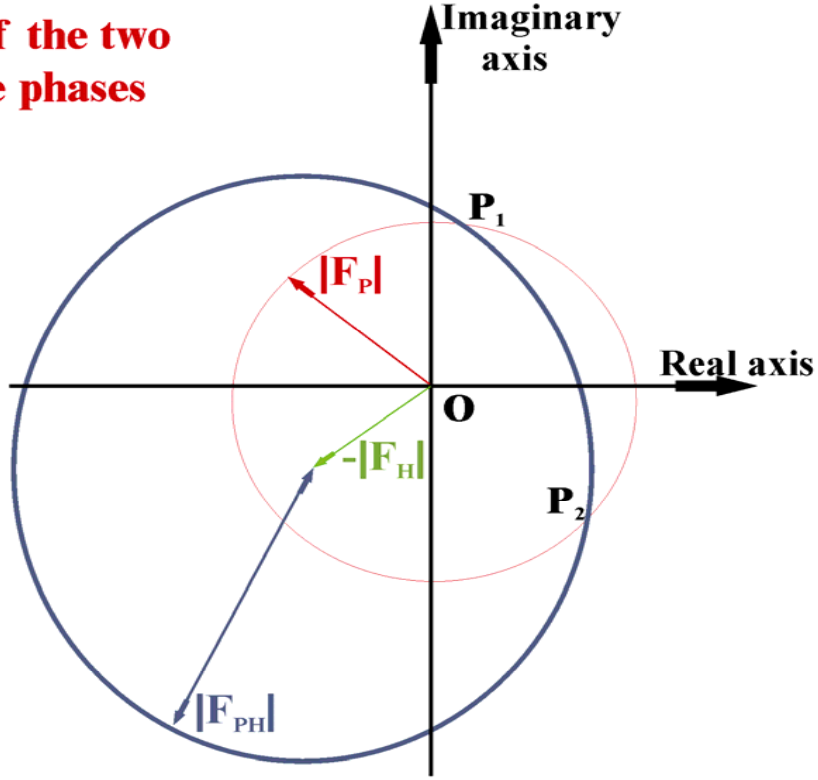


P$_1$ and P$_2$, the points of intersection of the two circles, represent the two possible values of the phases for the given reflection hkl

$|F_{PH}|$ circle centered about the end of the vector $-|F_H|$

Resolving Phase Ambiguities in SIR

P_1 and P_2, the points of intersection of the two circles, represent the two values of the phases for the given reflection hkl



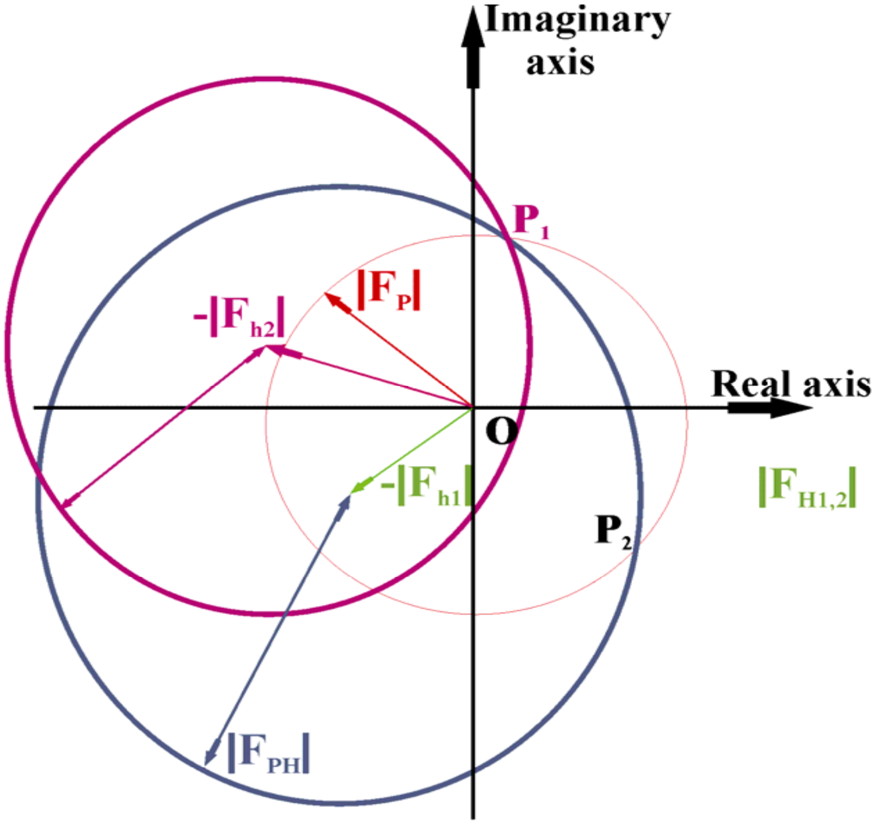
How to choose the best phase?

- $\overline{P_1 + P_2}$

(Large phase error if F_h is relatively small)

Prepare multiple heavy-atom derviatives

Multiple Isomorphous Replacement (MIR)



Known: $|F_P|$, $|F_{H1,2}|$ and $|F_{PH1,2}|$

Contribution of 2 separate heavy atoms to the structure factor amplitude

P$_1$ represents the only possible phase for the given reflection hkl

$|F_{PH1,2}|$ circles centered about the ends of the vectors $-|F_{H1,2}|$

Errors in MIR Phases

Assumptions:

1. **Ideal isomorphism**
2. **Exact heavy atom posititions**

But

1. **Random and systematic errors in measurement of intensities**
2. **Errors in estimation of heavy atom positions**
3. **Errors due to lack of isomorphism**

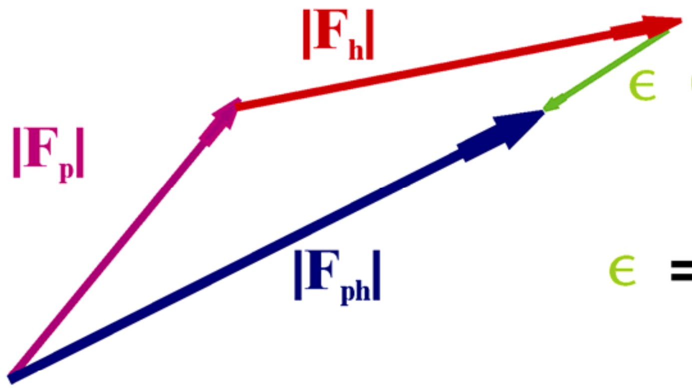
How to treat these errors?

Errors in MIR Phases

Assumptions:

1. All errors are Gaussian

2. Errors in heavy atom position and that due to lack of isomorphism can be considered together



ϵ (Lack of closure error)

$\epsilon = |F_{ph}| - \{|F_p| + |F_h|\}$

σ_{exp}: Errors in experimental observations

σ_{fh} : Combined errors in F_h and lack of isomorphism

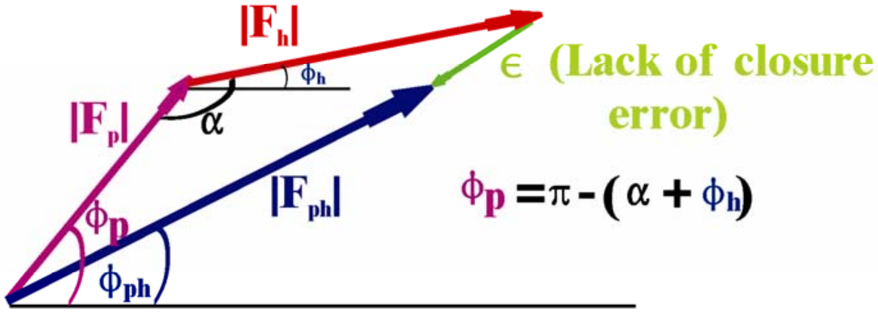
Errors in MIR Phases

Total error in F_{PH}: $\sigma = [\sigma_{exp}^2 + \sigma_{fh}^2]^{1/2}$

σ_{exp} : Errors in experimental observations

σ_{fh} : Combined heavy atom position and lack of isomorphism error

Errors in MIR Phases



Probability distribution $P(\phi) = C\exp\left(-\epsilon^2/2\sigma^2\right)$

where C is a normalization constant such that

$$\int^{2\pi} P(\phi)\,d\phi = 1$$

Errors in MIR Phases

Probability distribution $P(\phi) = C\exp\left(-\epsilon^2/2\sigma^2\right)$

Using

$$\cos\phi = \frac{F_h^2 + F_p^2 - (F_{ph}+\epsilon)^2}{2F_hF_{ph}}$$

it is possible to calculate the probability distribution for each phase of each derivative

For all derivatives

$$\boxed{P(\phi) = \prod_j P_j(\phi) = C\exp\sum_j\left(-\epsilon_j^2/\sigma_j^2\right)}$$

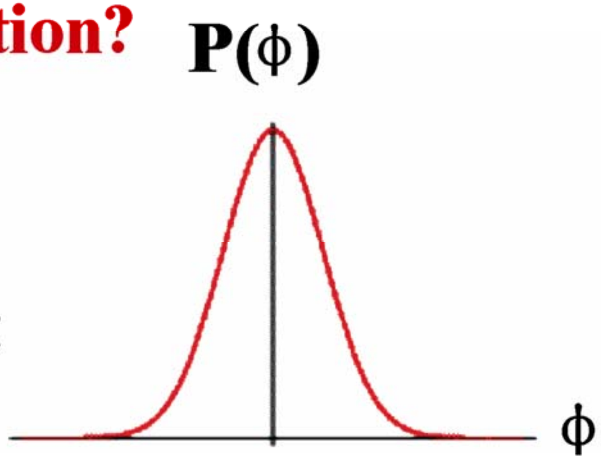
Best MIR Phase

Probability distribution of the MIR phases

$$P(\phi) = \prod_j P_j(\phi) = C\exp\sum_j (-\epsilon_j^2/\sigma_j^2)$$

Best phase for a given reflection?

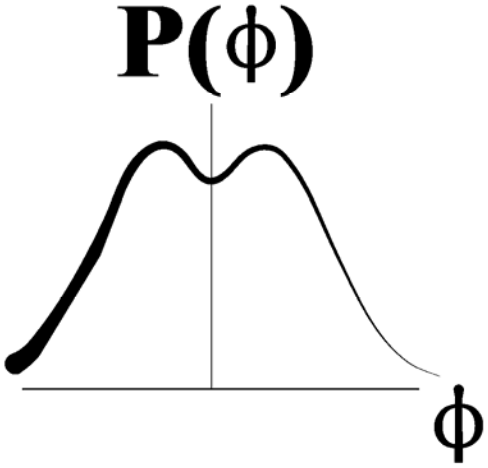
For a unimodal distribution, the phase angle ϕ at which $P(\phi)$ is maximum, is the best phase angle.



contd

Best MIR Phase

Usually $P(\phi)$ is bimodal, and the centroid of the phase probability gives the best phase



$$\phi_{best} = \frac{\int_0^{2\pi} P(\phi)\exp(i\phi)d\phi}{\int_0^{2\pi} P(\phi)d\phi}$$

Figure of Merit of a Phase

$$\phi_{best} = \frac{\int_0^{2\pi} P(\phi)\exp(i\phi)d\phi}{\int_0^{2\pi} P(\phi)d\phi}$$

The Figure-of-merit (m) of a phase is defined as the mean of the cosine of the phase error

$$m = \frac{\sum P(\phi_i)\cos(\phi_i)}{\sum P(\phi_i)} = \langle \cos \Delta\phi_i \rangle$$

where $\Delta\phi_i = \phi_{best} - \phi_i$

The FOM weighted Fourier coefficient is $mF \exp i\phi_{best}$

Molecular Replacement

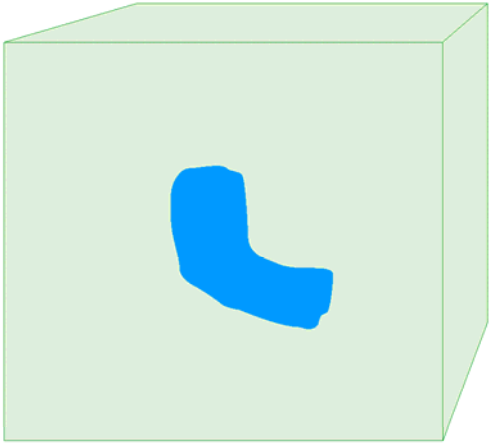
- Similar structure exists (sequence identity) MIR not required

- Orient the known structure as closely as possible to the unknown structure

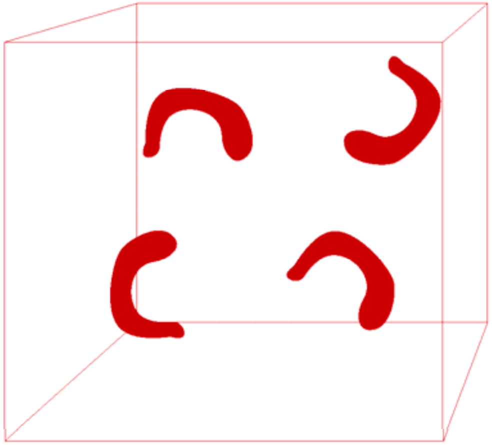
- Place the known structure as correctly as possible

- Rotational and translational parameters will give a good set of starting phases

Molecular Replacement

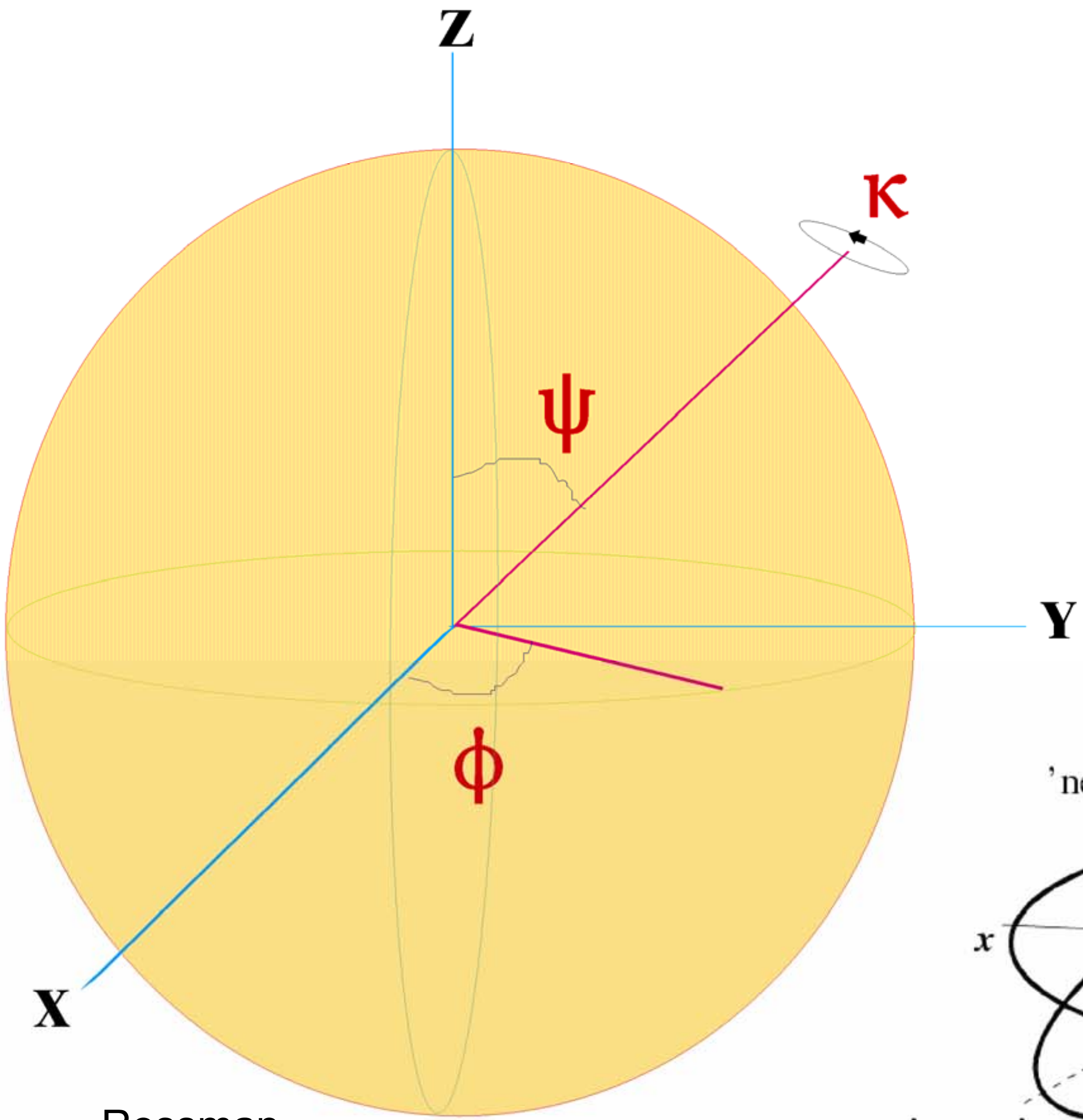


Model
Human lysozyme



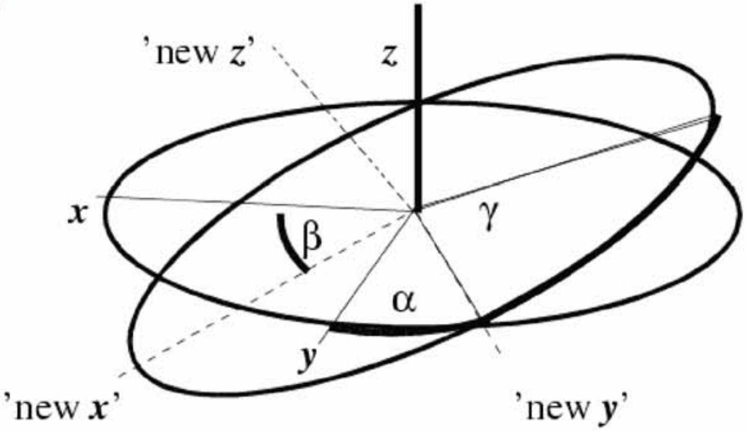
Data
Dinosaur lysozyme

Rotation Conventions



Rossman

Euler



Patterson Rotation Functions

$$\mathbf{R}(\kappa,\phi,\psi) = \int_{r_{min}}^{r_{max}} \mathbf{P}_{data}(\mathbf{u})\ \mathbf{P}_{model}(\kappa,\phi,\psi,\mathbf{u})\,d\mathbf{u}$$

model: Atomic model of known structure (Human lysozyme)

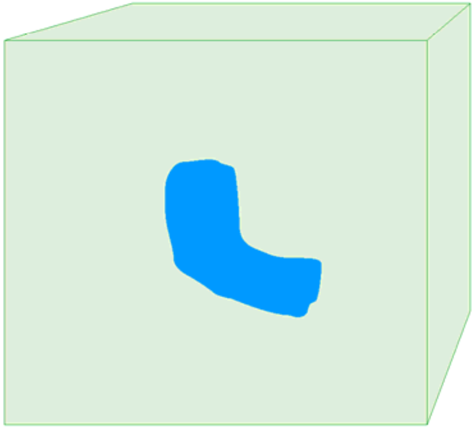
data: Unknown structure (Dinosaurus lysozyme)

Real space: Rotation of inter-atomic vectors
Reciprocal space: Convolution of SF **2

Determine the best(κ ϕ ψ) at which R shows a maximum

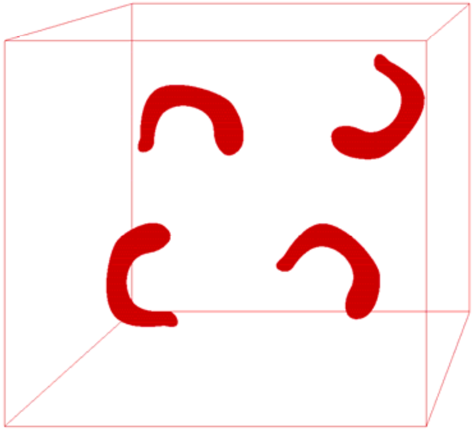
Orient the model through these angles

Orientation of Unknown Molecule



Model
Human lysozyme

One asymmetric unit in unit cell (P1 symmetry)



Data
Dinosaur lysozyme
Crystallographic symmetry

Position of unknown molecule wrt symmetry axes?

Position of Molecule: Translation Function

$$T(u) = \int P_{data}(u)\, P_{model}(u+r)du$$

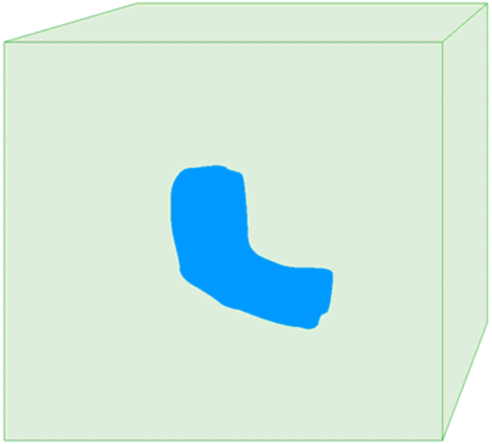
model: Atomic model of known structure (Human lysozyme)

data: Unknown structure (Dinosaurus lysozyme)

Real space: Rotation of inter-atomic vectors
Reciprocal space: Convolution of SF **2

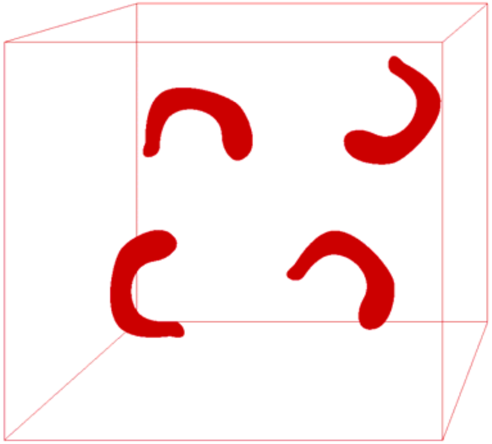
Determine the best position r at which T shows a maximum for the oriented model

Intra- and Inter-Molecular Vectors



Model
Human lysozyme

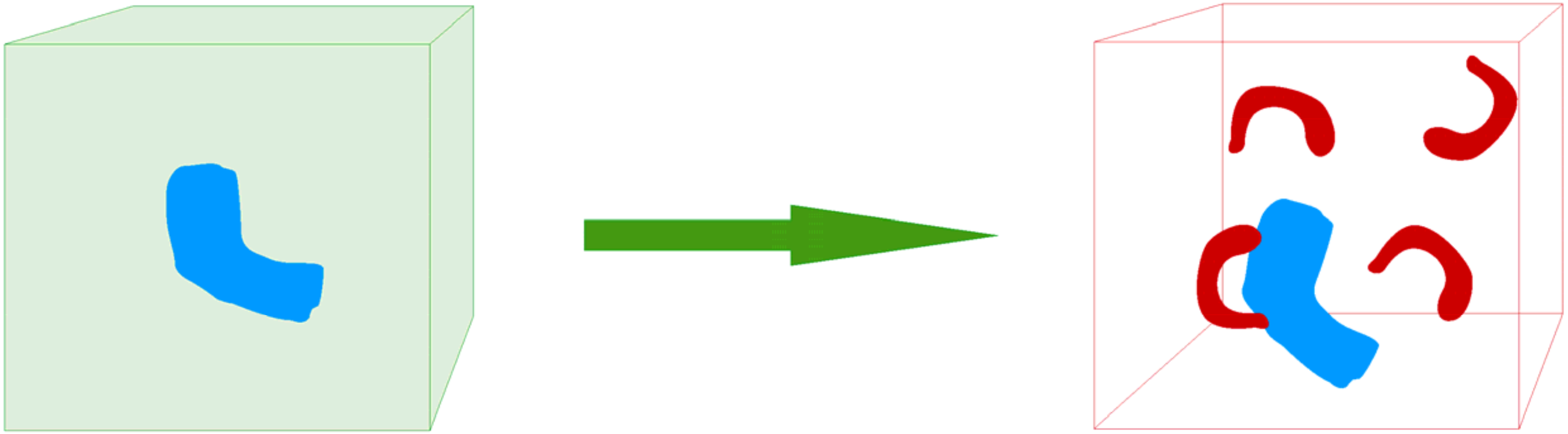
One asymmetric unit in unit cell (P1 symmetry)



Data
Dinosaur lysozyme
Crystallographic symmetry

Rotation fn: Match intra-molecular vectors
Translation fn: Match inter-molecular vectors

Another Translation Function



Model
Human lysozyme

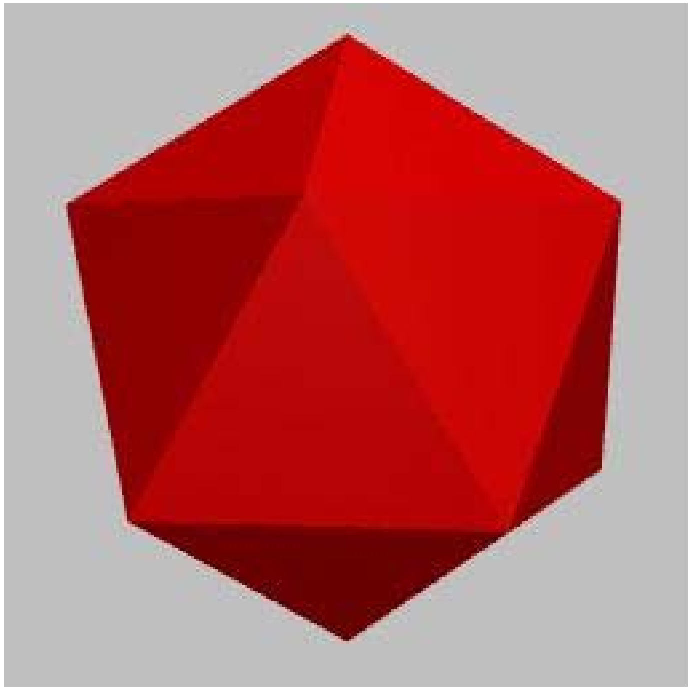
One asymmetric unit in unit cell (P1 symmetry)

Data
Dinosaur lysozyme
Crystallographic symmetry

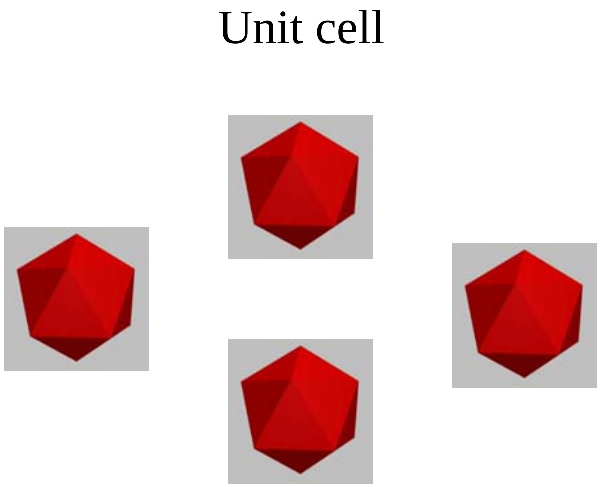
Packing function: Discrepancy between the calculated and observed SFs

Position which gives the minimum discrepancy factor is the best position of the molecule in the unit cell

Averaging Using Inherent Symmetry

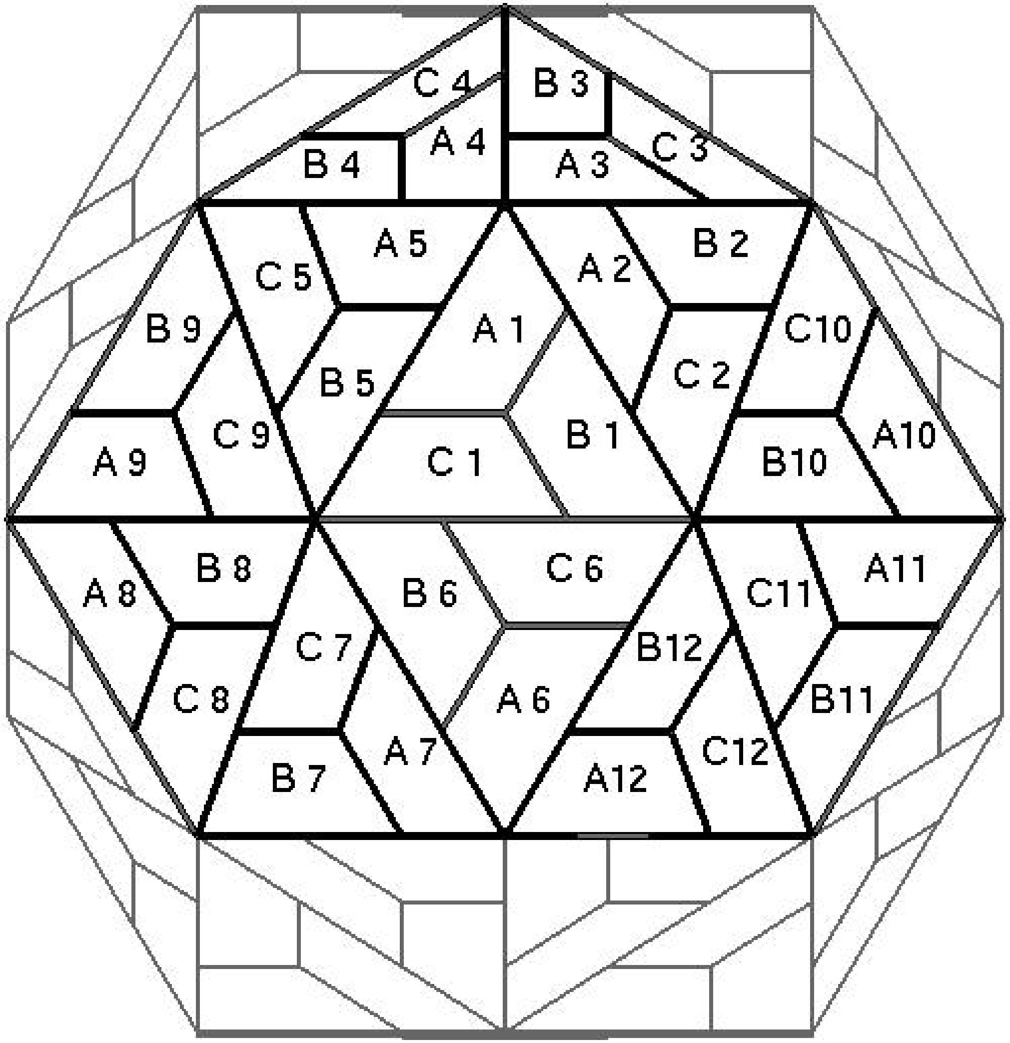


No of faces : 12 (5-folds)
No of edges : 30 (2-folds)
No of vertices : 20 (3-folds)

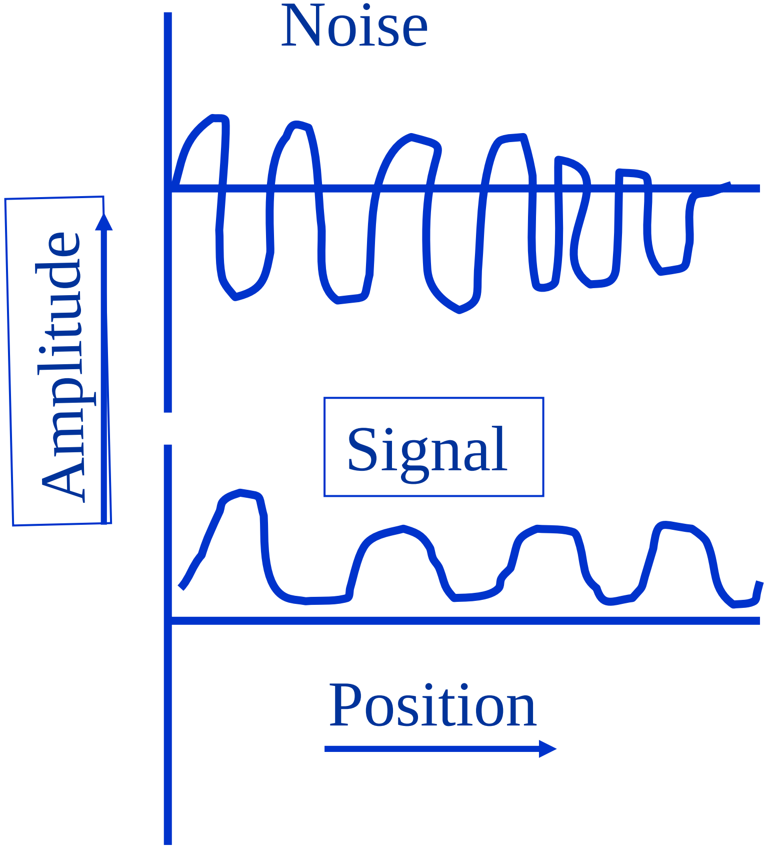


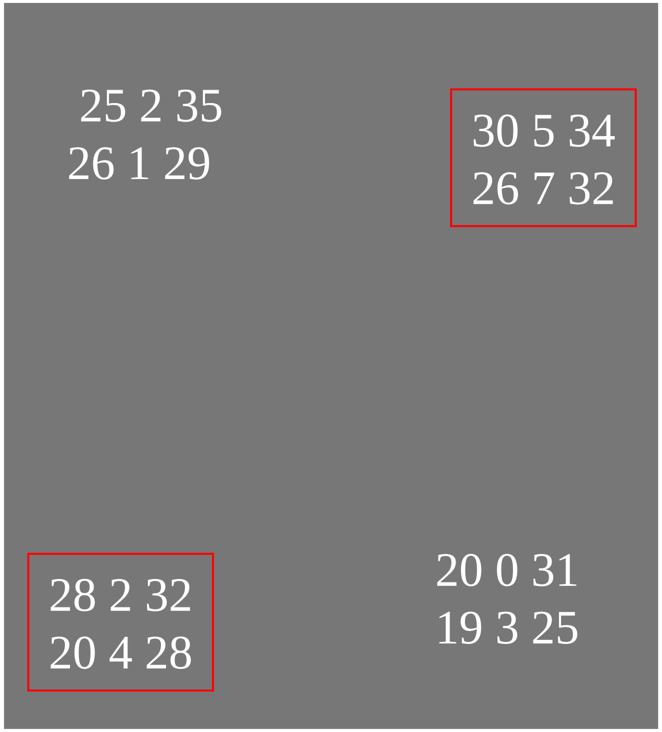
- Particle + unit cell symmetry
- Symmetry preserved in diffraction space

T=3 Icosahedron



Averaging



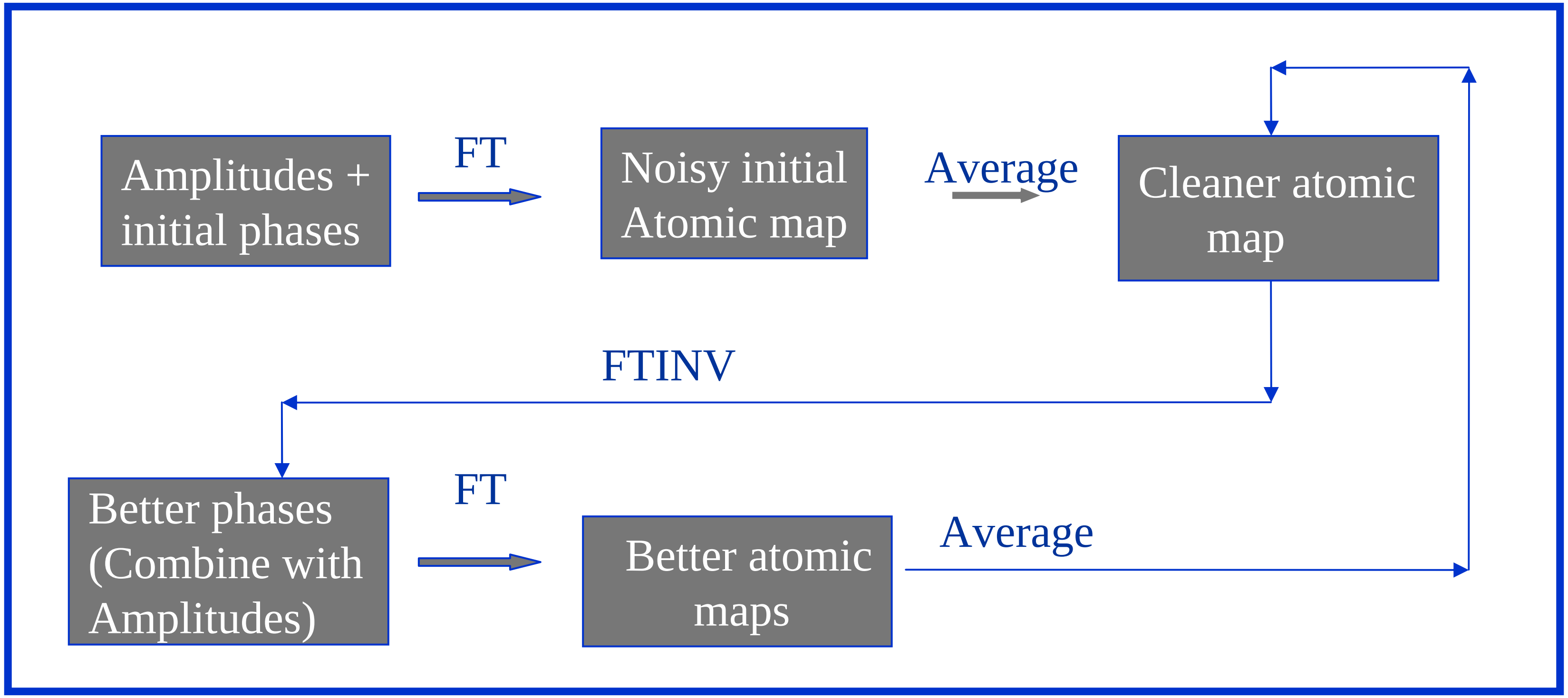

25 2 35
26 1 29

30 5 34
26 7 32

28 2 32
20 4 28

20 0 31
19 3 25

Averaging and Iterative Improvement of Phases



Amplitudes + initial phases

FT

Noisy initial Atomic map

Average

Cleaner atomic map

FTINV

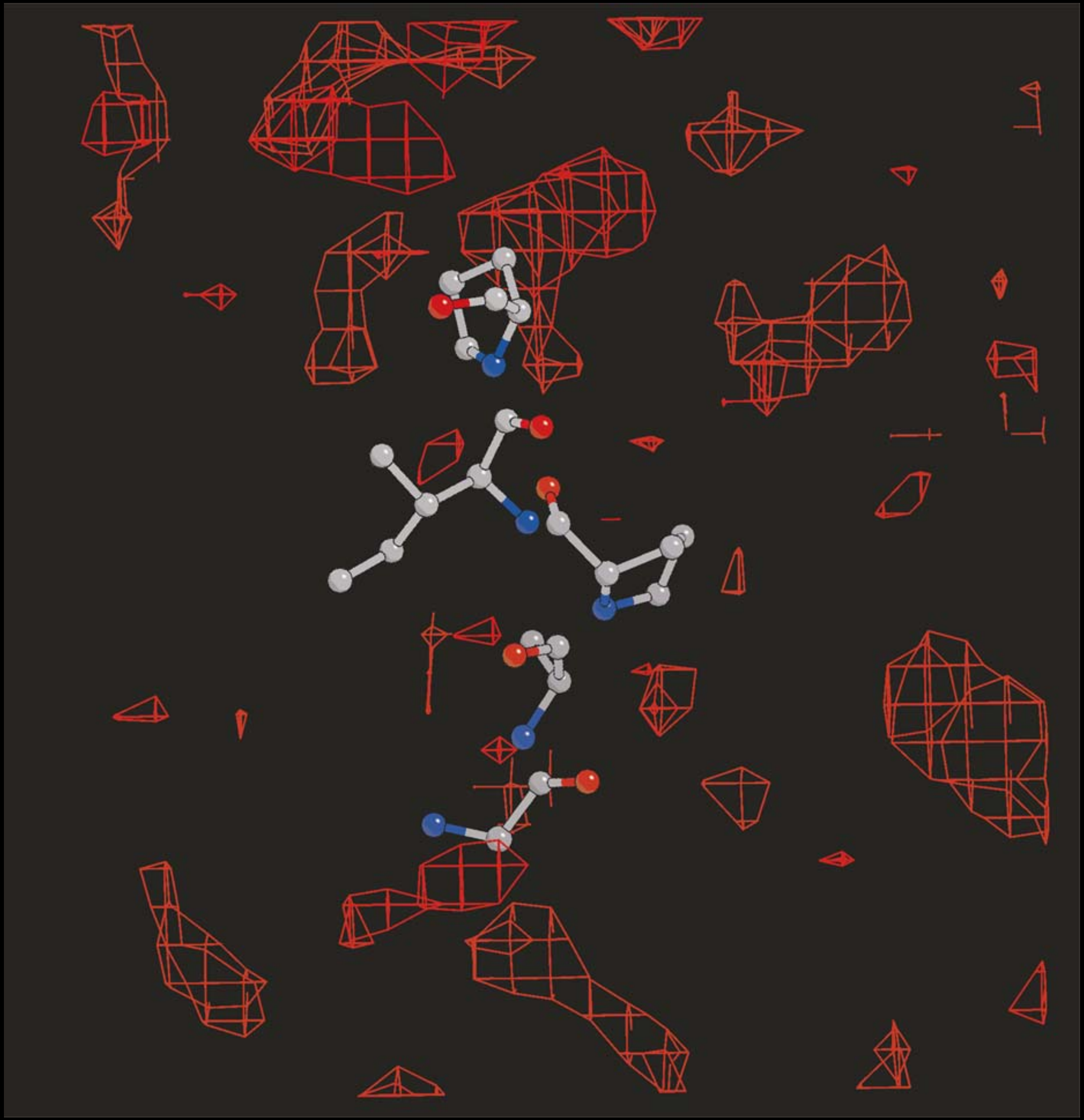
Better phases (Combine with Amplitudes)

FT

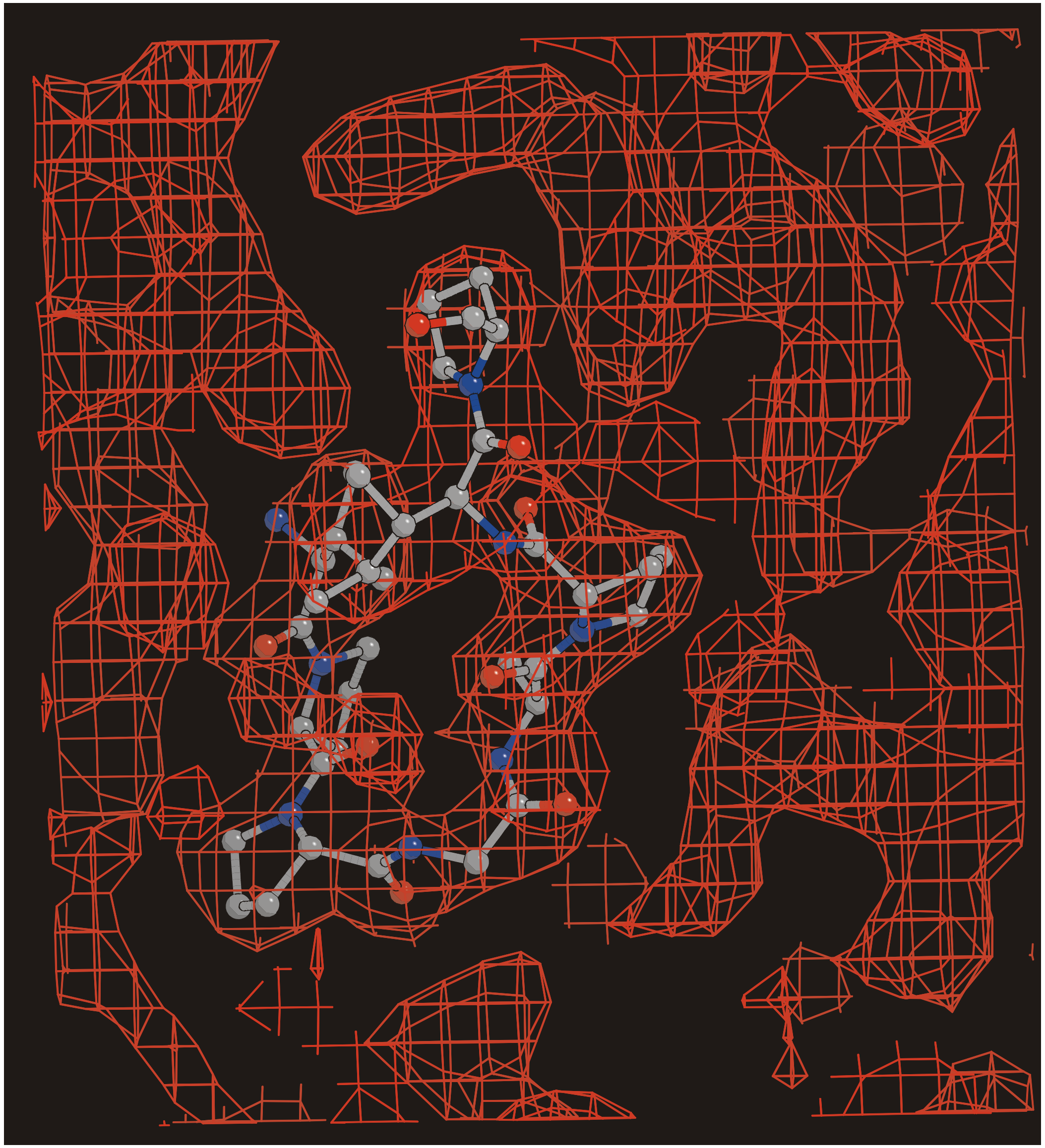
Better atomic maps

Average

Poor Initial Phases

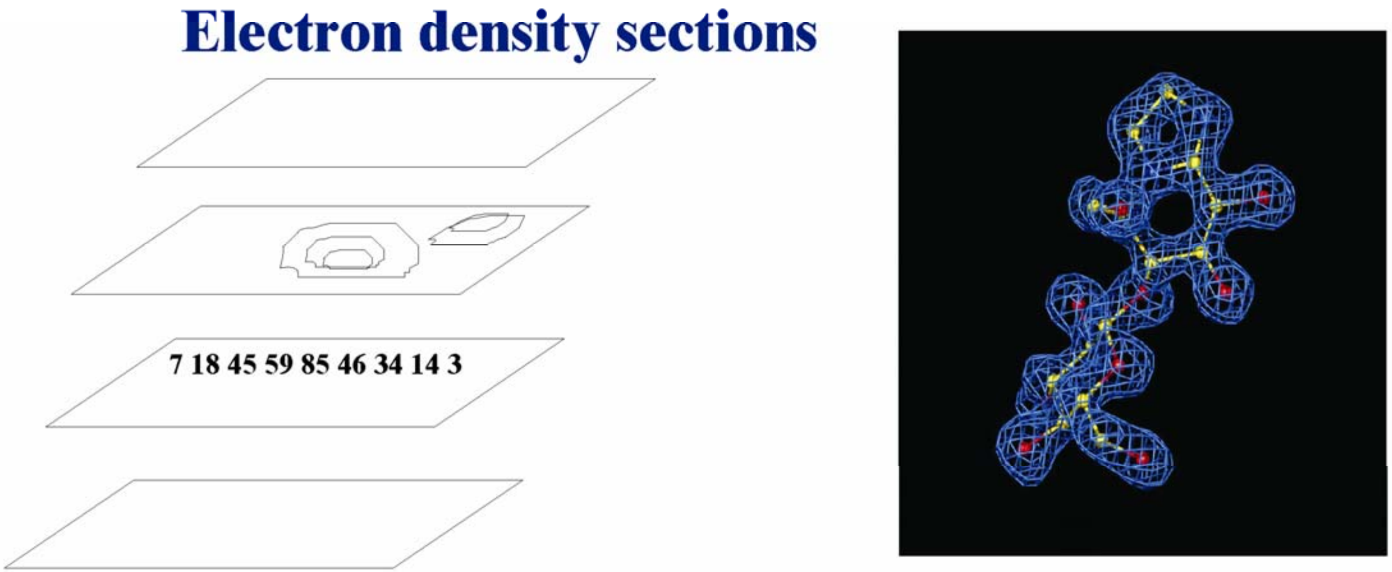


Improved Phases after Cycles of Averaging



Display of Electron Density

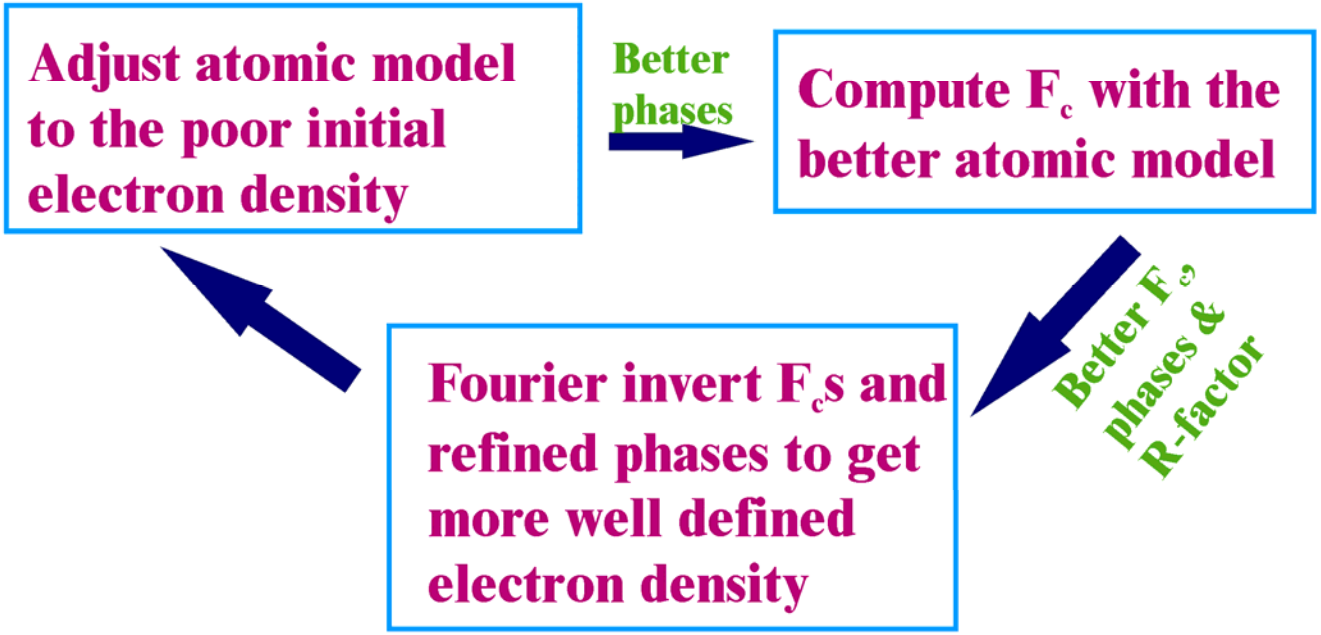
Electron density maps are displayed as iso-contour figures


Electron density sections

Refinement of Atomic Model

Poor initial phases => F_c does not agree with F_{obs}

$$\text{Discrepancy or R-factor} = \sum_{h\,k\,l} \frac{\big|\,|F_{obs}| - |F_c|\,\big|}{|F_c|} \quad \text{is high}$$



Adjust atomic model to the poor initial electron density

Better phases

Compute F_c with the better atomic model

Better F_c, phases & R-factor

Fourier invert F_cs and refined phases to get more well defined electron density

Refinement is Like Curve Fitting

Fit the best 2-parameter curve through 9 points

No. of observations: 9
No. of parameters: 2

Observational equations

$y_1 = a_{11}x_1 + a_{12}x_2$

$y_2 = a_{21}x_1 + a_{22}x_2$

……..

.

.

$y_9 = a_{91}x_1 + a_{92}x_2$

Ax = y, where A is a (9x2) matrix that gives the best estimate of the parameters x to match the observational column vector y

Minimize $(y-Ax)^t(y-Ax)$ to get the best estimates of the parameters x

Normal Equations in Refinement

Observational equations: $A\psi = b$

A: Matrix that calculates the F_cs from the positional and the thermal parameters ψ

b: Experimentally obtained structure factors F_{obs}s

Residual vector: $r = b - A\psi$

Minimize the objective function $M = r^T r$

$$\boxed{\frac{\partial M}{\partial \psi} = 0 \Rightarrow A^t A\psi = A^t b}$$

Normal equations

Observational equations are non-linear in parameters:

$A(\delta\psi) = b$ $\boxed{A^t A\delta\psi = A^t b}$ **Conjugate Gradient**

Full LSQ Refinement not Possible

Refinement parmeters: Positional and thermal

Protein of 30kDa: ~300 residues => ~3000 atoms
(3 positional + 1 thermal parameter) / atom
12000 refinement parameters.

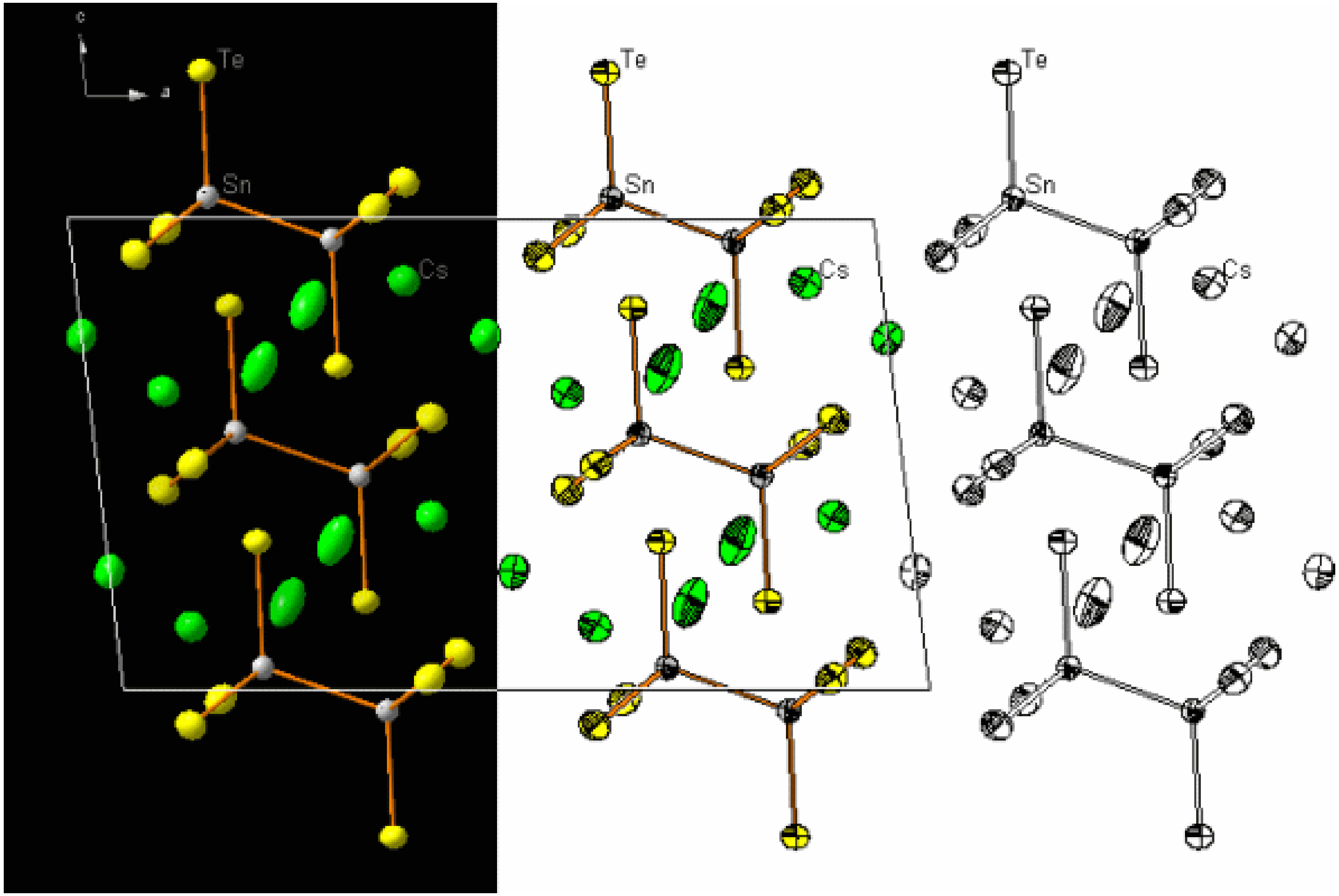
Unit cell: 80x70x40 A => ~16000 reflections within the 2A Ewald sphere

Overdeterminacy ratio = # observations / # parameters = 16K/12K ~ 1.4 (Inadequate for LSQ refinement)

Introduce geometric and energy restraints/constraints (Restrained/constrained least squares refinement)

Anisotropic B-valuesThermal Ellipsoids

$$B_{isotropic} = const * \langle u^2 \rangle$$



http://www.crystalimpact.com/diamond/v2feature-ellipsoids.htm

R_{free} in Refinement of Atomic Models

$$\textbf{Discrepancy or R-factor} = \sum \frac{\left| |F_{obs}| - |F_c| \right|}{|F_c|}$$

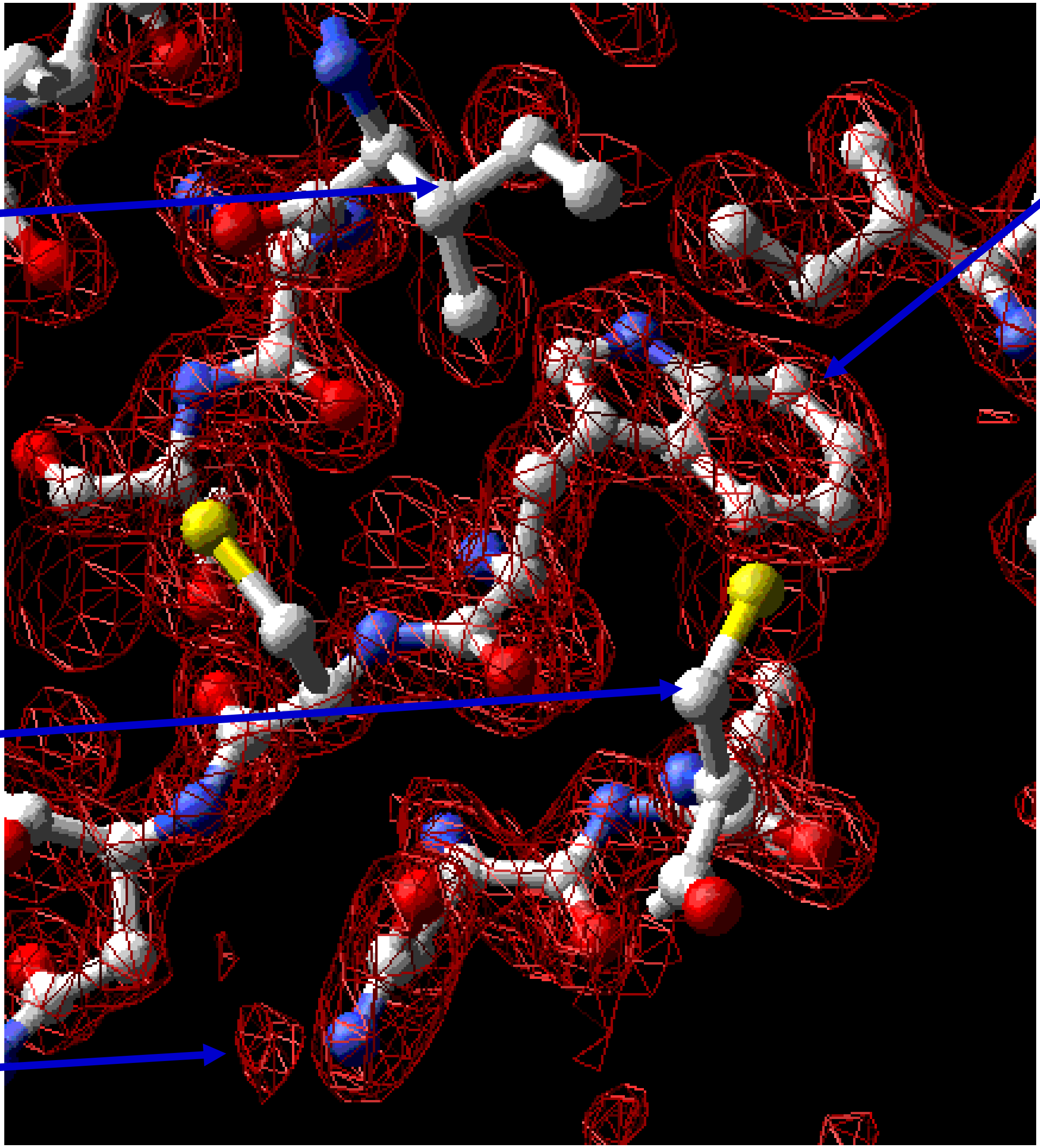
Q: How good does this model predict the data that it has not seen?

Curve fitting analogy: Does a 4th pt lie on a 3-pt quadratic curve?

R_{free}: Choose a random 5-10% test data sub-set and calculate R-factor for this test set.

If the model is good, R_{free} should closely follow R of the remaining data set.

Refined Electron Density



Poorly defined ed

Trp

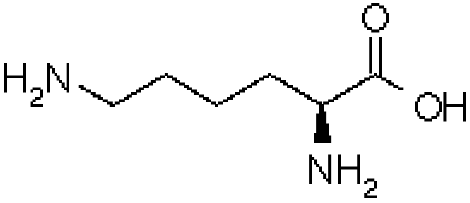
Lack of clear density

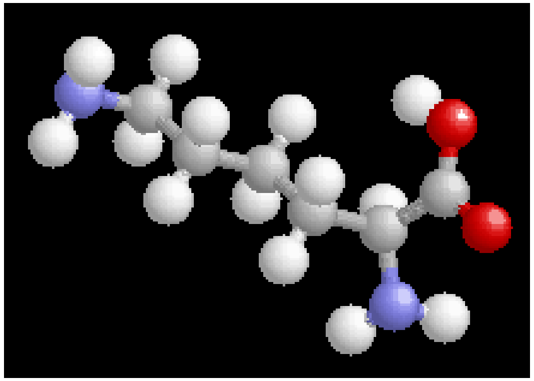
Uninterpreted density

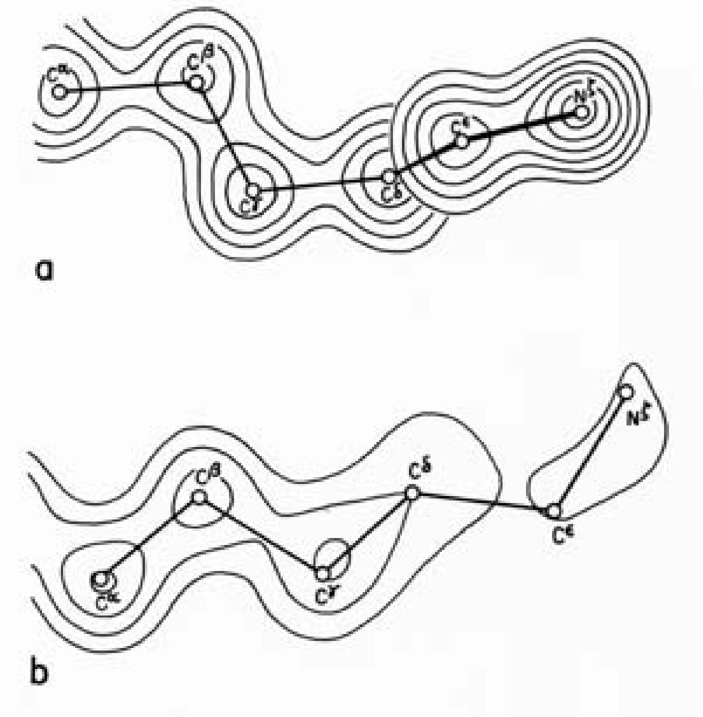
Conformers in Lys, Asn…

http://www.usm.maine.edu/~rhodes/

Manual Editing Using Graphics







Disordered Lysine

Manual model editing is necessary

Concluding Remarks

- X-ray diffraction may be analyzed by FT/series. Alternative methods also reported

- Phases may be determined using several methods

- Phases may be improved by averaging and/or refining the atomic model with model fitting

- Well-refined models are essential to correctly interpret biological functions

Resources

- Cantor, R.C. & Schimmel, P.R. (1980). *Biophysical Chemistry Part II: Techniques for the study of biological structure and function*. New York; W.H. Freeman & Co, Chapter 13.
- Blundell, T.L. & Johnson, L.N. (1976). *Protein Crystallography*. London; Academic Press.
- Stout, G.H. & Jensen, L.M. (1968). *X-ray Structure Determination: A* Practical Guide. New York: Macmillan