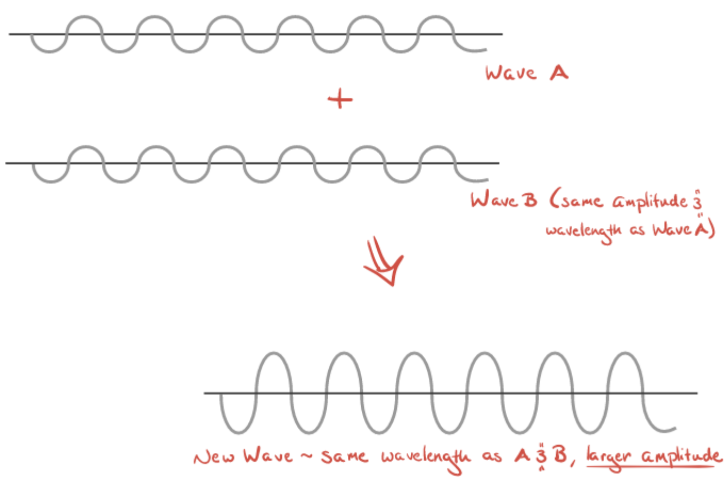
Structure solution & refinement (Small Molecules)

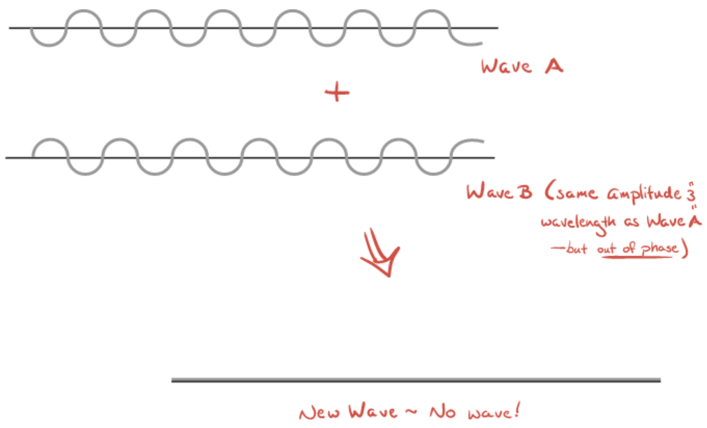
Clara Gomes

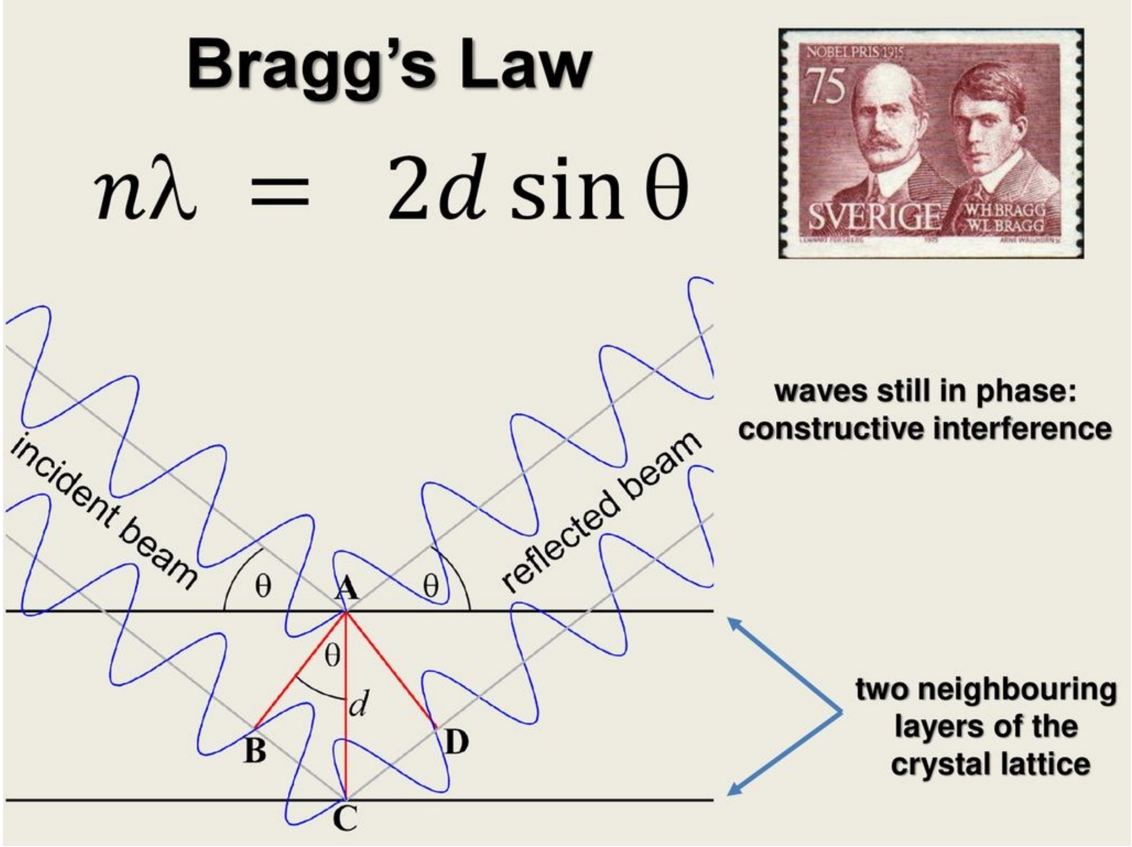
LAQV-REQUIMTE, FCT-NOVA

7th European Crystallographic School (**ECS7**) – June 12, 2022

Waves and scattering


Constructive Interference


Destructive Interference

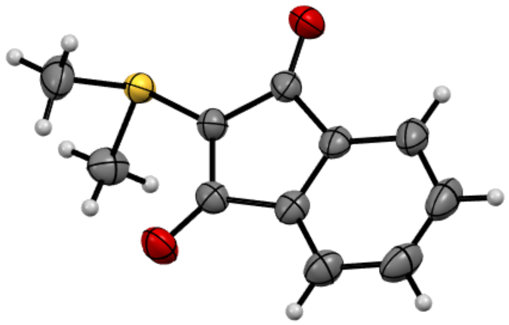

Bragg's Law

$$n\lambda = 2d\sin\theta$$

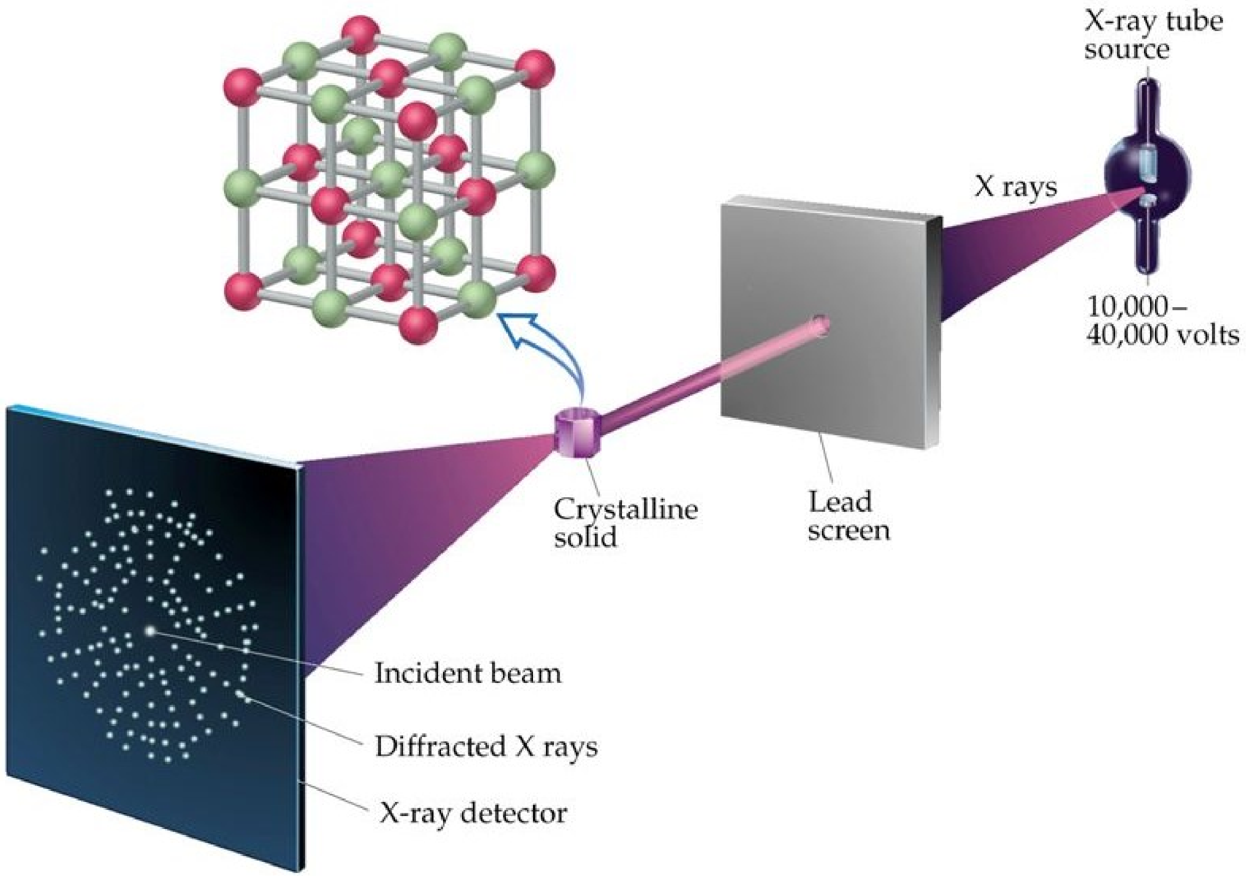
waves still in phase: constructive interference

two neighbouring layers of the crystal lattice

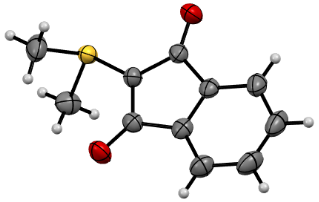
A SCXRD typical experiment



After data
reduction,
structure
solution and
refinement

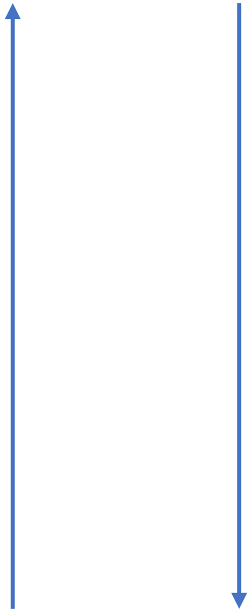


Relationship between a crystal structure and its diffraction pattern



Crystal structure (what we want to know)

Atom positions and displacements; electron density distribution



(Reverse) Fourier transform:

add up the scattered waves, imitating an X-ray microscope lens system

Calculation only; physically unachievable

Fourier transform:

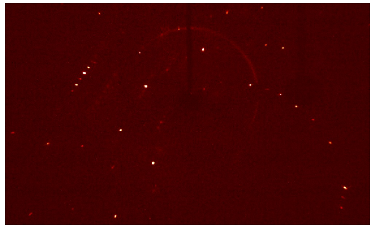
add up the contributions to scattering by all the atoms

Calculation from model structure

Diffraction experiment

Diffraction pattern (what we can measure, in part)

Amplitudes [and phases] of X-ray waves



Data obtained from X-ray Crystallography

Directly:

- Positions of reflections
- Intensities of reflections and their estimated standard deviations.

Indirectly, from reflection positions:

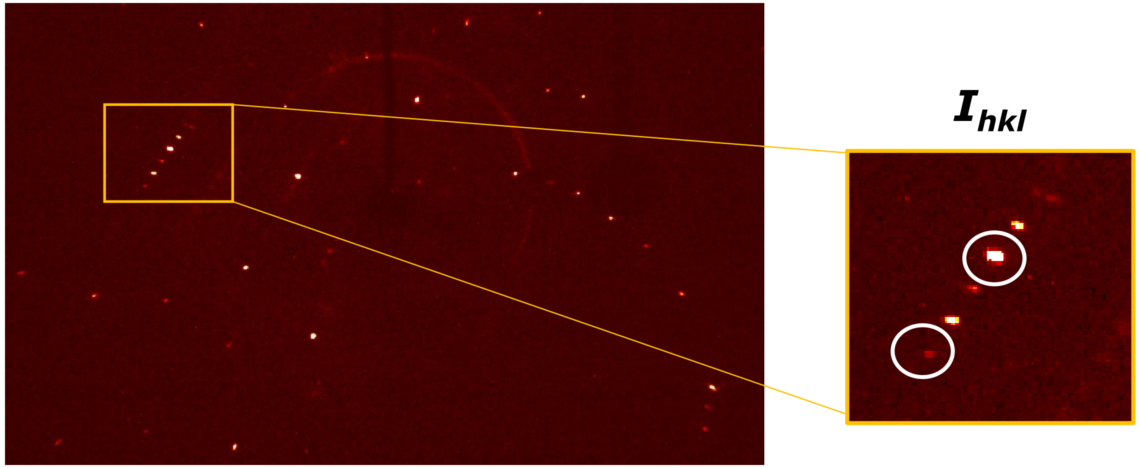
- Smallest crystallographic repeating unit (the “unit cell” of the structure, *a*, *b* and *c* axes, α, β and γ angles)
- Metric symmetry of the structure (Laue symmetry).

Indirectly, from reflection intensities:

- Symmetry of the structure (Laue symmetry and space group)
- Phases of reflections (from model – calculated)
- Positions and thermal parameters of atoms
- Bond distances and angles, molecular symmetry, etc

How do reflections relate to a structure?

Diffraction pattern



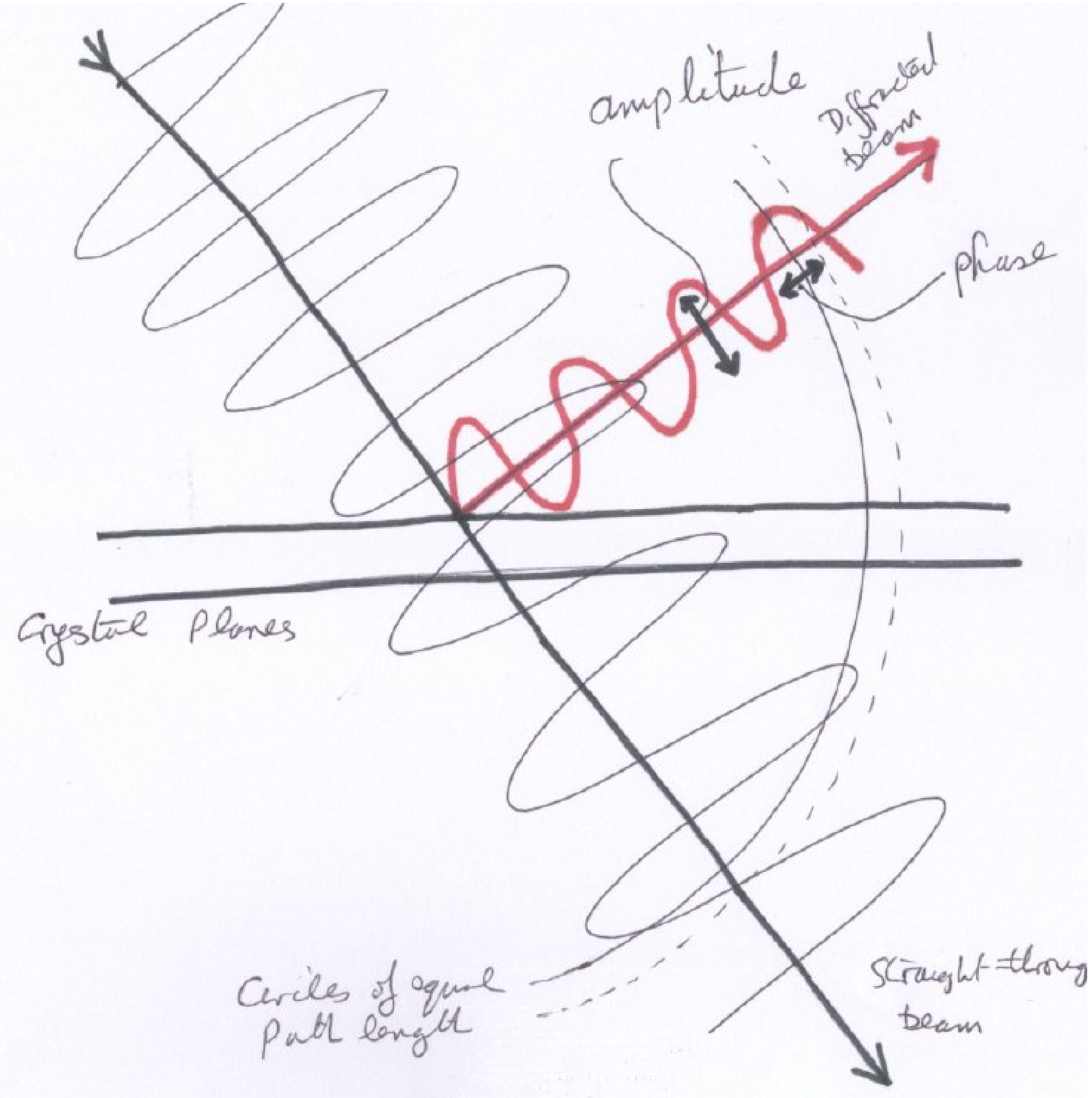
h	k	l	I_{hkl}	σ_{hkl}
-1	0	0	0.10	0.30
-1	0	0	0.00	0.10
1	0	0	0.10	0.30
-2	0	0	4074.69	171.98
-2	0	0	3640.14	171.88
2	0	0	3451.65	171.88
-3	0	0	0.80	0.50
-3	0	0	0.80	0.70
-3	0	0	0.30	0.90
3	0	0	0.30	0.70
-4	0	0	441.16	20.70
-4	0	0	437.76	20.50
-4	0	0	440.06	20.90
-5	0	0	0.60	1.40
-5	0	0	2.80	2.10

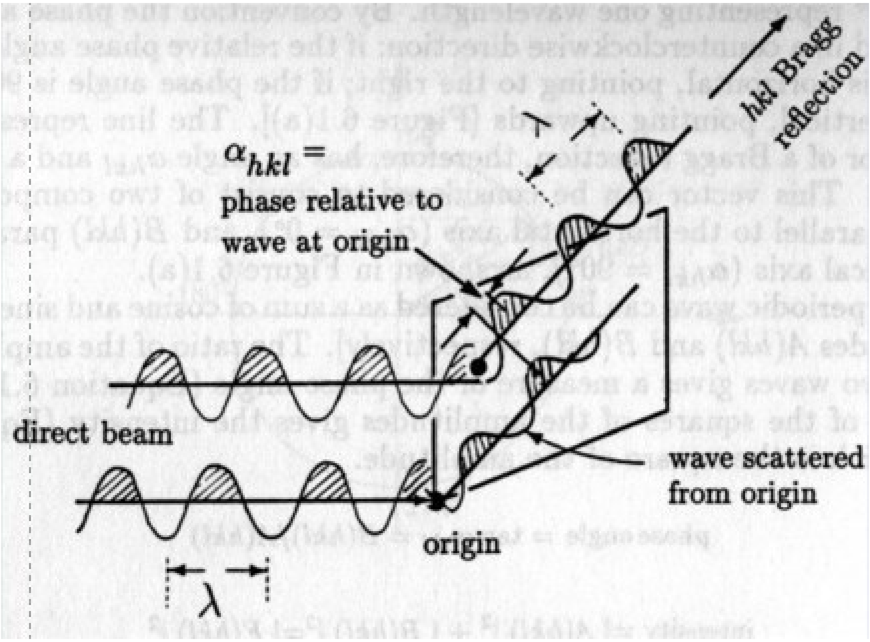
The reflections are indexed (attribution of *hkl* indices – Miller indices) and a list of intensities, I_{hkl}, and the associated standard deviation σ_{hkl} for each measured reflection is obtained.

The intensity I_{hkl} is proportional to the structure factor F^2_{hkl}

$$F(h,k,l) = \underbrace{\left|F(h,k,l)\right|}_{\text{Amplitude}} \cdot e^{i\overset{\text{phase}}{\boxed{\alpha}}}$$

Amplitude and phase





The phase of a diffracted beam is actually the phase lag with respect to the incident beam.

It is not easy to measure.

How do reflections relate to a structure?

For a **one atom** system (from Braggs’ law):

$$F(hkl) = |F(hkl)|e^{i\alpha(hkl)} = |F(hkl)|e^{i(2\pi(hxj + kyj + lzj)}$$

- F(hkl) is the **structure factor** of the reflection hkl (intensity and phase of the diffracted beam)
- |F(hkl)| is the **max. amplitude**, or magnitude of a diffracted beam in phase with the direct beam ($\alpha = 0$)
- x_j, y_j and z_j are the fractional coordinates of the atom j in the crystal lattice
- $2\pi(hx_i + ky_i + lz_i) = \alpha$ is the **phase** of the diffracted beam

How do reflections relate to a structure?

Multi-atom system: the contributions of all atoms in the unit cell have to be added up for every diffracted beam hkl:

$$F(hkl) = \sum_j f'_j \, e^{i\alpha_j(hkl)} = \sum_j f'_j \, e^{i(2\pi(hxj + kyj + lzj)}$$

F(hkl) is the structure factor of the reflection hkl (intensity and phase of the diffracted beam)

f'_j is the dampened structure factor of the atom *j*

x_j, y_j and z_j are the fractional coordinates of the atom j in the crystal lattice

$\alpha_j = 2\pi(hx_j + ky_j + lz_j)$ is the phase of the diffracted beam for the atom j

How do reflections relate to a structure?

The **atomic structure factor *f*** is a mathematical description of how a material scatters incident radiation.

$$f' = f\, e^{\frac{-2\,\pi^2\, u^2}{d^2}} \quad \text{(isotropic case)}$$

f' = dampened atomic scattering factor (reduction of scattering from an atom with θ)
f = undampened atomic scattering factor (number of electrons of atom);
u = measure of atomic vibration (u^2 is the mean vibrational amplitude, the atomic displacement factor or temperature factor U)
d = d-spacing (from Braggs’ law)

How do reflections relate to a structure?

Fourier transform of diffraction pattern (cannot be done experimentally!) yields the **electron density** (unit e/Å^3) in the crystal lattice as a function of F(hkl) (intensity and phase of the reflection):

$$\rho_{XYZ} = {}^{1}/_{V} \sum_{hkl} F(hkl)\, e^{-i\alpha(hkl)} = {}^{1}/_{V} \sum_{hkl} F(hkl)\, e^{-i2\pi}(hX + kY + lZ)$$

V = volume of the unit cell
F(hkl) = structure factor for reflection hkl (intensity and phase, see previous page)
X, Y and Z are arbitrary points on the e- density map, not atomic positions!

How do reflections relate to a structure?

Rewriting as a continuous function (rather than a sum) yields:

$$\mathrm{F(hkl)} = \int \rho(\mathrm{XYZ})\ e^{i\alpha(hkl)} \mathrm{dV}$$

and

$$\rho(\mathrm{XYZ}) = {}^{1}/_{V} \sum_h \sum_k \sum_l F(hkl)\ e^{-i\alpha(hkl)}$$

In this form, **F(hkl) is the inverse FT of** ρ**(XYZ)**, and ρ**(XYZ) is the FT of F(hkl)**

Solving of an X-ray crystal structure – the phase problem

- Determination of a Crystal Structure requires to know the **unit cell dimensions**, the **space group**, and identify the **positions of the atoms** within the unit cell

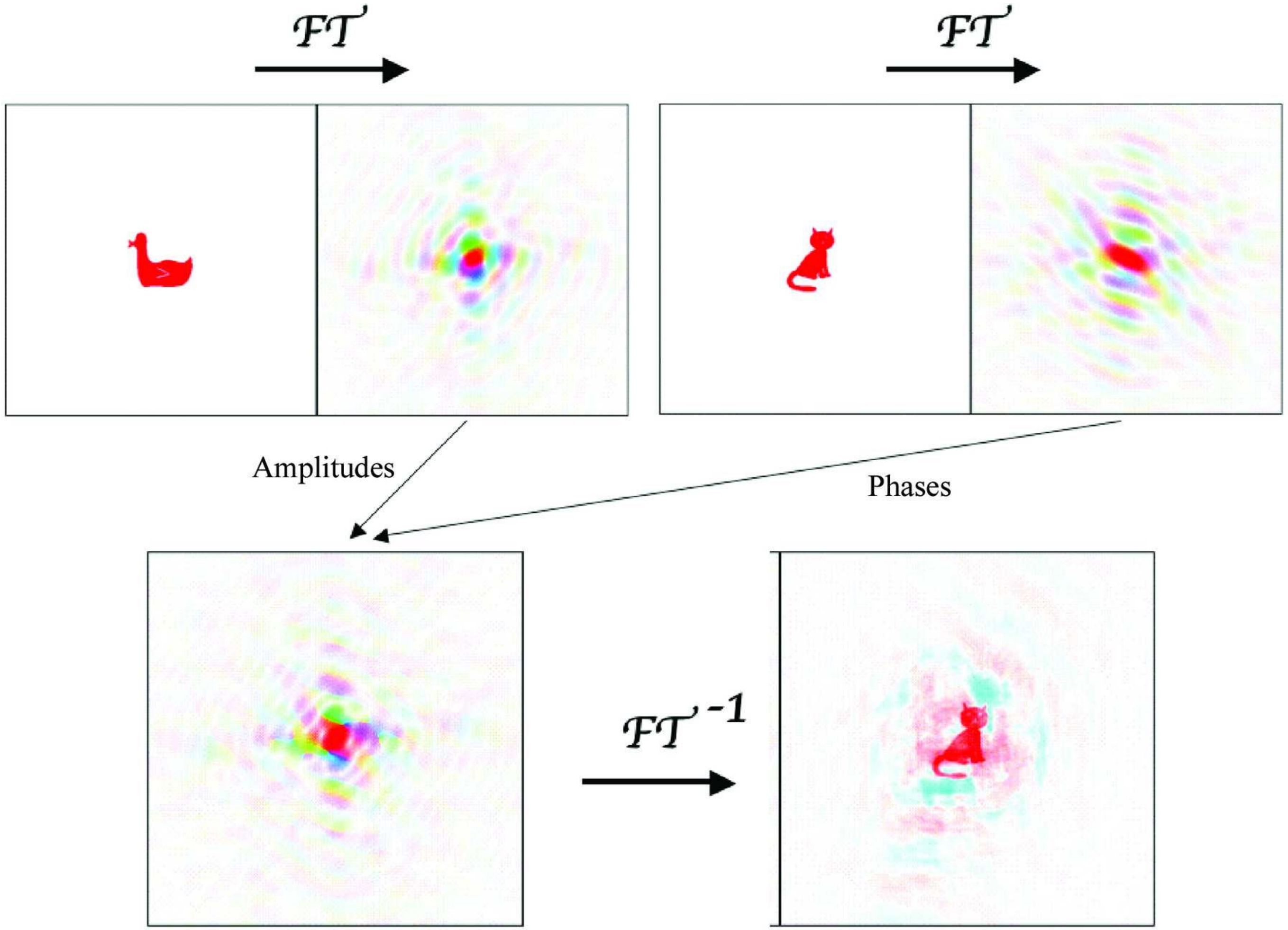
- Mathematically, the electron density function and the X-ray structure factors are related to each other by Fourier transformation.

- To calculate the electron density function – i.e. atom positions - from the reflections, one would have to perform a **Fourier synthesis** using amplitudes and phases of all reflections.

- However, only the intensities (squared amplitudes) can be measured. The phase information is lost (measurement is not time-resolved, the relative phase shifts of the reflections hitting the x-ray detector cannot be determined).

- Unfortunately, the phases are more important for Fourier synthesis than the intensities. They contain all the information about the x, y and z positons of the atoms! This difficulty is known as the **phase problem** of crystallography.

Solving of an X-ray crystal structure – the phase problem



Kevin Cowtan: http://www.ysbl.york.ac.uk/~cowtan/fourier/fourier.html

Approaches to solve the phase problem

1.The Patterson or heavy atom method:

Atoms that are heavier than all others in the crystal will give rise to the strongest reflections. Heavy is a relative term however, all atoms after Rb are normally considered to be heavy. Using the Patterson method it should be possible to **calculate the position of a heavy atom** and this could be **used to phase** most of the reflections and **calculate an electron density map**.

2.Direct Methods: The Direct Methods approach uses probabilistic/statistical techniques that tries to guess the reflection phases and then awards a Figure of Merit to the guess. It uses **only the intensities** and a **random starting set of phases**.The most popular is the Combined Figure of Merit or **CFOM**). It is normal practice to accept solutions with CFOM values in the range 0.03-0.15 and to try to use these phases to generate an electron density map.

The chemical formula should be correct, or structure solution fails. Direct methods need **atomic resolution** data to work**.**

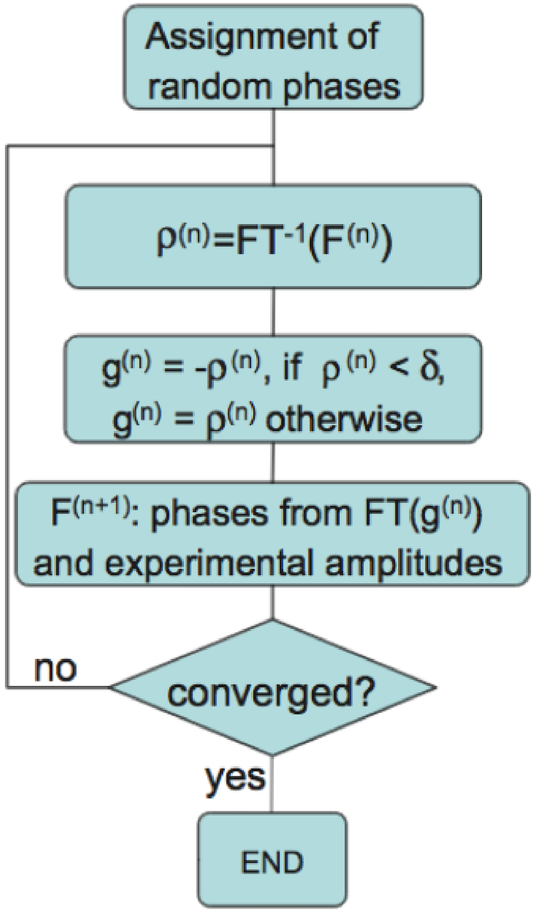
Approaches to solve the phase problem

3. The charge flipping method:

It is a **dual-space** method. The iteration is initialized by **assigning random phases** to the experimental structure-factor amplitudes. From this trial solution an **electron density** is calculated by **inverse Fourier transform**. This electron density is modified in such a way, that all grid points with density **below** a certain positive **threshold** δ are multiplied by -1 (**flipped**). New temporary structure factors are calculated by Fourier transform of this modified density. The phases of these temporary structure factors are combined with the experimental amplitudes and such a set of structure factors enters the next cycle of iteration.

δ is the only parameter of the whole algorithm.

The charge flipping method works well if the **data has good resolution**.



Approaches to solve the phase problem

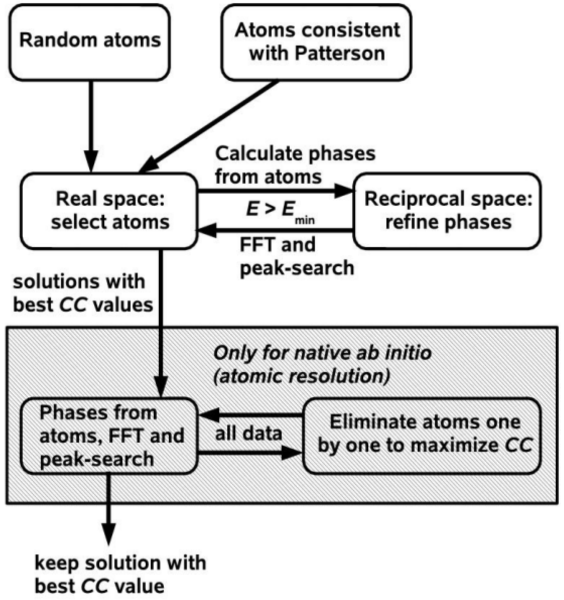
4. Intrinsic Phasing (SHELXT) - approach

- Assuming the **Laue** group to be **known**, equivalent intensities are averaged and the data then **expanded to P1**.
- The phase problema is solved by **dual space recycling** in P1. Result = **electron density map + corresponding phases**.
- The phases are used to determine the **correct space group and translation** necessary for the electron density map to fit it.
- The phases are averaged in this space group and used to calculate an improved density that is improved by further Fourier recycling.
- The maxima of the density are assigned to atoms.

Does not require the formula to be correct.

Limitations:

- When disorder, pseudosymmetry or twinning are present.
- Non-centrosymmetric structures with centrosymmetric arrangement of very heavy atoms.



Random atoms

Atoms consistent with Patterson

Calculate phases from atoms

Real space: select atoms

$E > E_{min}$

Reciprocal space: refine phases

FFT and peak-search

solutions with best *CC* values

Only for native ab initio (atomic resolution)

Phases from atoms, FFT and peak-search

all data

Eliminate atoms one by one to maximize *CC*

keep solution with best *CC* value

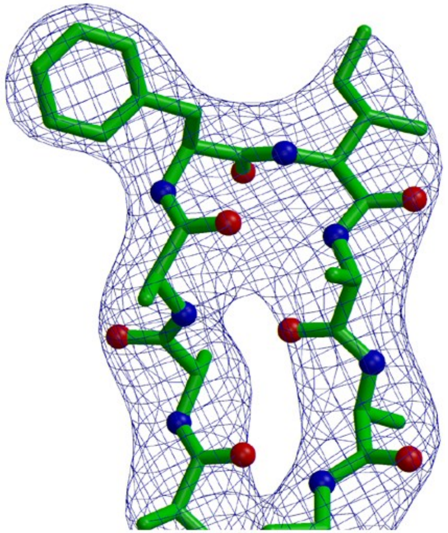
Resolution – why is it important?

“Resolution” is related with the level of detailed that can be obtained: a high resolution map gives more detailed image of the atomic structure

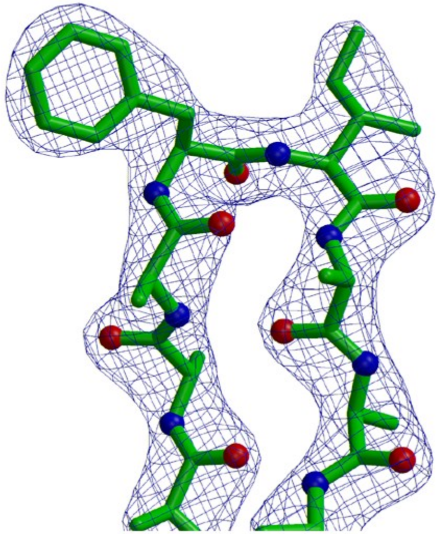
Resolution: sin θ / λ

Low resolution
low q, high d

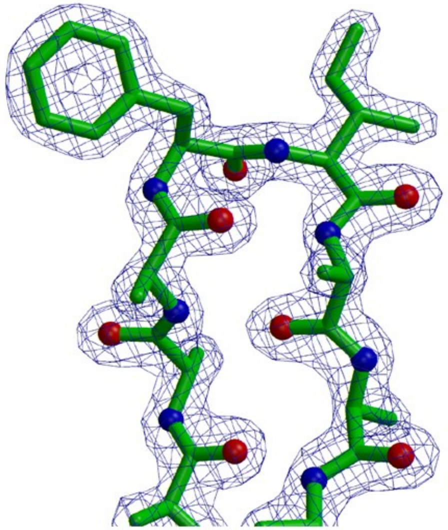
High resolution
high q, low d



Map at 4 Å
Diffuse



Map at 2.9 Å



Map at 2 Å
Almost contains the atomic detail

Refinement - SHELXL

What does “refinement” mean?

- Refinement is the “improvement or clarification of something by the making of small changes”

- It is the way from the first solution to the final accurate and publishable model (removes the initial errors in the model by an iterative optimization process)

  - ✓ Initial errors include: incorrect atom coordinates, no atom types assigned at all, groups of lighter atoms, disorders, hydrogen positions, etc.

This can be a short highway to happiness or a long and rugged road through pain and sorrow!

Least Squares Refinement

Concept

- By means of Fourier transformation, a complete set of structure factors is calculated from the atomic model. The **calculated intensities** are then **compared** with the **measured intensities**, and the best model is that which minimizes M:

$$M = \sum w \left(F_o^2 - F_c^2 \right)^2$$

or

$$M = \sum w \left(|F_o| - |F_c| \right)^2$$

F^2_o or F^2_{obs}(hkl) = observed structure factor

F^2_c or F^2_{calc}(hkl) = calculated structure factor

w = weighting factor, reflects our confidence in this particular datum (derived from the standard uncertainty σ of that measurement) $w = 1/\sigma^2 (F_{obs}{}^2)$

- The least squares process is repeated until the suggested shifts are insignificantly small (until the parameter shift is much smaller than the esd on the parameter).

Least Squares Refinement

Refinement against *F* or F^2?

- Refinement **against *F***:

- Involves the extraction of a root (remember: $I \propto F^2$) → leads to mathematical **problems** with **very weak reflections** or reflections with negative measured intensities.
- To overcome this: negative measurements must be set to zero.

⬇

- This introduces bias → ignoring information from very weak reflection **affects the structure determination** and problems with the $\sigma(F)$ estimation from $\sigma(F^2)$ lead to **inaccuracies in the refinement**.
- Advantages: R1 << wR2 and refinement is faster (less data)

- Refinement **against F^2**:

- Does not cause any of the above-mentioned problems.
- **Advantages**: makes the refinement of twinned structures mathematically simpler, and refinement against F^2 is less likely to settle into a local minimum.

Parameters, constraints and restraints

Parameters

- For each atom in the model, general position in the unit cell: Three **atomic coordinates** (x,y,z) and one or six **atomic displacement parameters** (one for isotropic, six for anisotropic models)
- **Flack parameter** for non-centrosymmetric structures.
- Etc.

Constraints

Are equations rigidly relating two or more parameters or assigning fixed numerical values to certain parameters.

- Site occupancy factors (eg: disoreder)
- Special position constraints (eg: on coordinates, in Uij values and on occupancies)
- Rigid group constraints

Restraints

- treated as additional experimental observations (increase the data points to refine against)
- Geometrical: DFIX, DANG, SADI, SAME, etc
- On displacement parameters: ISOR, DELU, SIMU, etc

Residual factors

R-factors – assess the quality of the model

- ***wR*** or ***wR2*** in SHELXL (weighted *R*-factor based on F^2)

$$wR = \left[\frac{\sum w \left(F_o^2 - F_c^2\right)^2}{\sum w F_o^2} \right]^{1/2}$$

- The weighting factor ***w*** is individually derived from the standard uncertainties of the measured reflections and expresses the **confidence** we have **in every single reflection**.

- ***R*** or ***R1*** in SHELXL (unweighted *R*-factor based on *F*) – most popular one!!

$$R = \frac{\sum \left| |F_o| - |F_c| \right|}{\sum |F_o|}$$

✓ ***R*** and ***wR2*** should be **as small as possible** (ideally zero, if model and data agree perfectly). If the model is complete and good, the R value should be close to the average error for the experimental diffraction intensities.

Residual factors

R-factors – assess the quality of the model

- ***GooF*** or ***S*** (goodness of fit)

$$S = \left[\frac{\sum w \left(F_o^2 - F_c^2\right)^2}{(N_R - N_P)} \right]^{1/2}$$

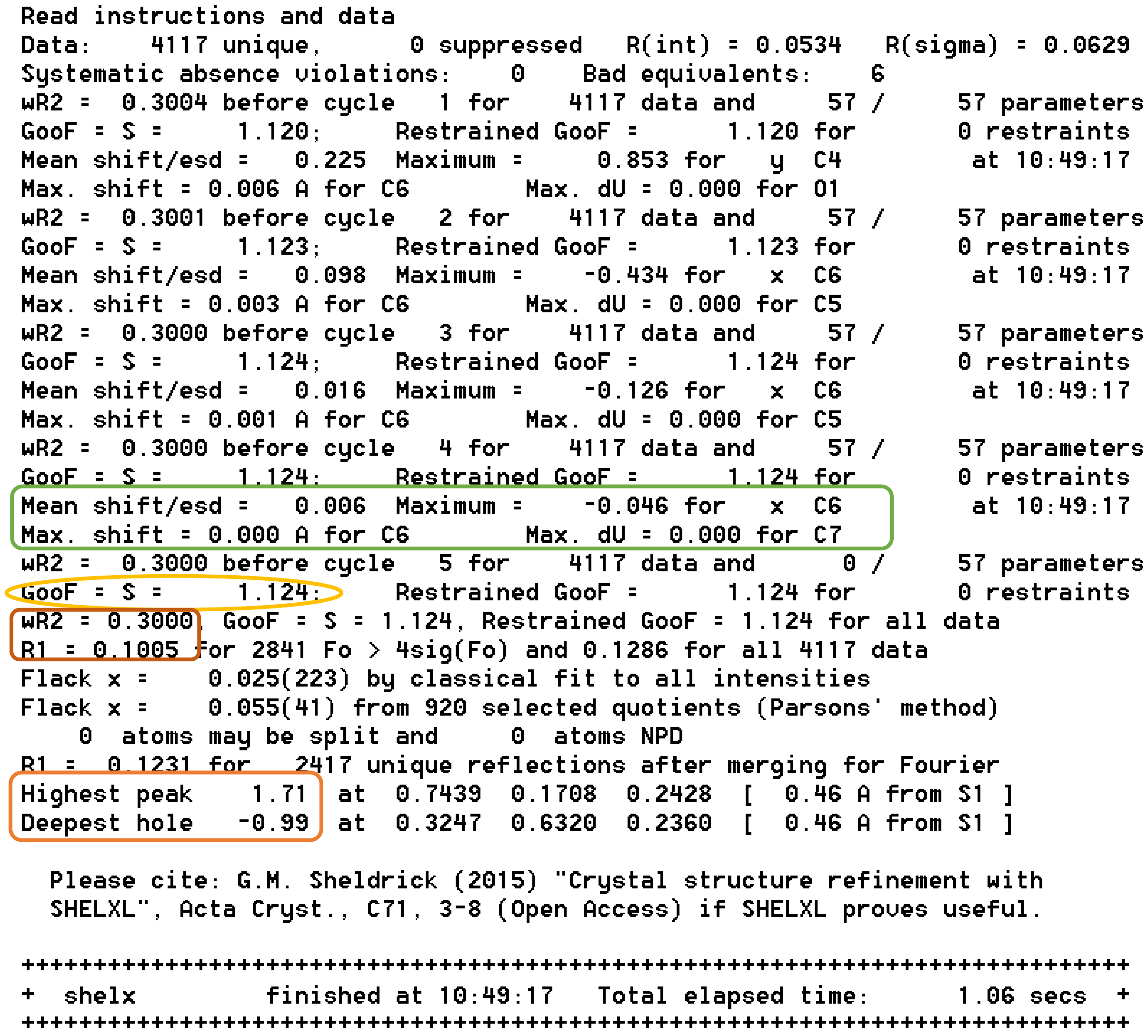
- N_R = number of independent reflections
- N_P = number of refined parameters

(N_R-N_P) is the excess of data *vs* parameters

✓ ***S* < 1** suggests the model is better than the data. Be careful!

✓ *S* should ideally be close to 1. If *S* is far from 1, the estimated standard deviations for bond positions, bond lengths, etc, will be wrong.

Example – SHELXL output

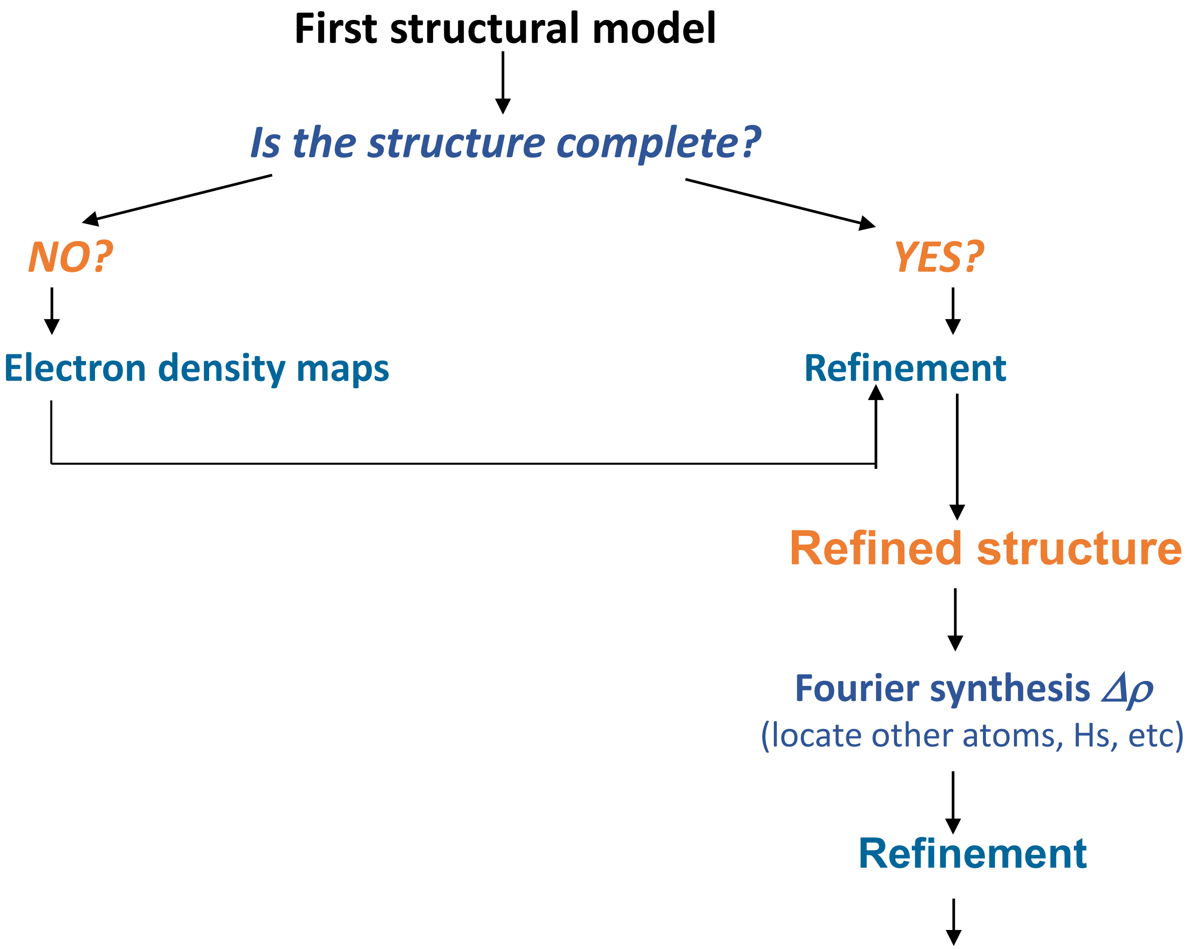


```
Read instructions and data
Data:    4117 unique,     0 suppressed   R(int) = 0.0534   R(sigma) = 0.0629
Systematic absence violations:    0    Bad equivalents:    6
wR2 =  0.3004 before cycle   1 for    4117 data and     57 /     57 parameters
GooF = S =     1.120;     Restrained GooF =      1.120 for       0 restraints
Mean shift/esd =   0.225  Maximum =     0.853 for    y  C4        at 10:49:17
Max. shift = 0.006 A for C6        Max. dU = 0.000 for O1
wR2 =  0.3001 before cycle   2 for    4117 data and     57 /     57 parameters
GooF = S =     1.123;     Restrained GooF =      1.123 for       0 restraints
Mean shift/esd =   0.098  Maximum =    -0.434 for    x  C6        at 10:49:17
Max. shift = 0.003 A for C6        Max. dU = 0.000 for C5
wR2 =  0.3000 before cycle   3 for    4117 data and     57 /     57 parameters
GooF = S =     1.124;     Restrained GooF =      1.124 for       0 restraints
Mean shift/esd =   0.016  Maximum =    -0.126 for    x  C6        at 10:49:17
Max. shift = 0.001 A for C6        Max. dU = 0.000 for C5
wR2 =  0.3000 before cycle   4 for    4117 data and     57 /     57 parameters
GooF = S =     1.124;     Restrained GooF =      1.124 for       0 restraints
Mean shift/esd =   0.006  Maximum =    -0.046 for    x  C6        at 10:49:17
Max. shift = 0.000 A for C6        Max. dU = 0.000 for C7
wR2 =  0.3000 before cycle   5 for    4117 data and      0 /     57 parameters
GooF = S =     1.124;     Restrained GooF =      1.124 for       0 restraints
wR2 = 0.3000, GooF = S = 1.124, Restrained GooF = 1.124 for all data
R1 = 0.1005 for 2841 Fo > 4sig(Fo) and 0.1286 for all 4117 data
Flack x =    0.025(223) by classical fit to all intensities
Flack x =    0.055(41) from 920 selected quotients (Parsons' method)
    0  atoms may be split and     0  atoms NPD
R1 =  0.1231 for   2417 unique reflections after merging for Fourier
Highest peak    1.71  at  0.7439  0.1708  0.2428  [  0.46 A from S1 ]
Deepest hole   -0.99  at  0.3247  0.6320  0.2360  [  0.46 A from S1 ]

  Please cite: G.M. Sheldrick (2015) "Crystal structure refinement with
  SHELXL", Acta Cryst., C71, 3-8 (Open Access) if SHELXL proves useful.

+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
+  shelx        finished at 10:49:17   Total elapsed time:      1.06 secs  +
+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
```

Summing up



First structural model

Is the structure complete?

NO?

Electron density maps

YES?

Refinement

Refined structure

Fourier synthesis $\Delta\rho$
(locate other atoms, Hs, etc)

Refinement

- *Structural analysis*
- *Molecular geometry calculations*

The structure is complete when…

1. The bonds in the model must make sense chemically. Similar bonds should have similar geometries, and all distances and bond angles, etc., should have similar values to the literature.

2. There can be no atoms with temperature factors (*B*-factors) with values too high or too low. Values should range between 0.03 and 0.15 depending on the type of atom.

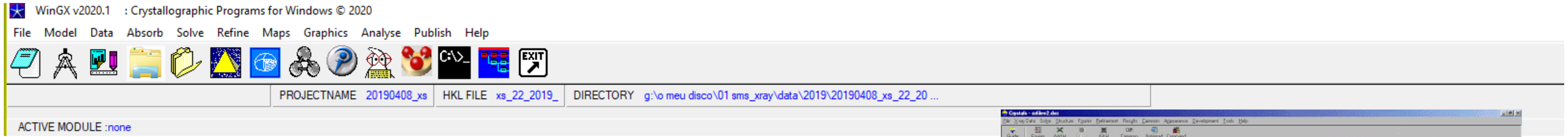
3. The structure must be refined to convergence, i.e., the ratio shift / error should be <0.05.

4. The goodness of fit, S, must have a value close to 1.0.

5. The final difference Fourier map should not be very large or very small peaks.

6. Final *R1* and *wR2* values should be reasonably low compared to data quality.

Software for structure solution and refinement

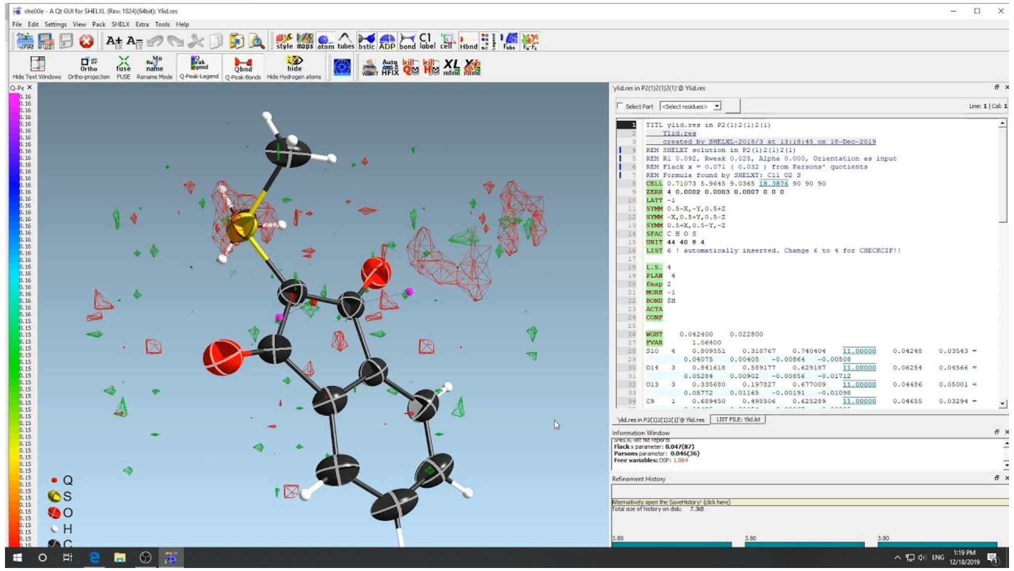


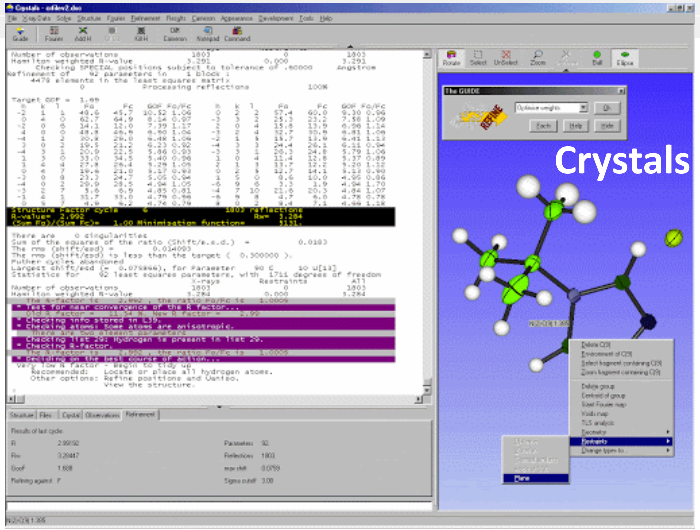
PROJECTNAME 20190408_xs

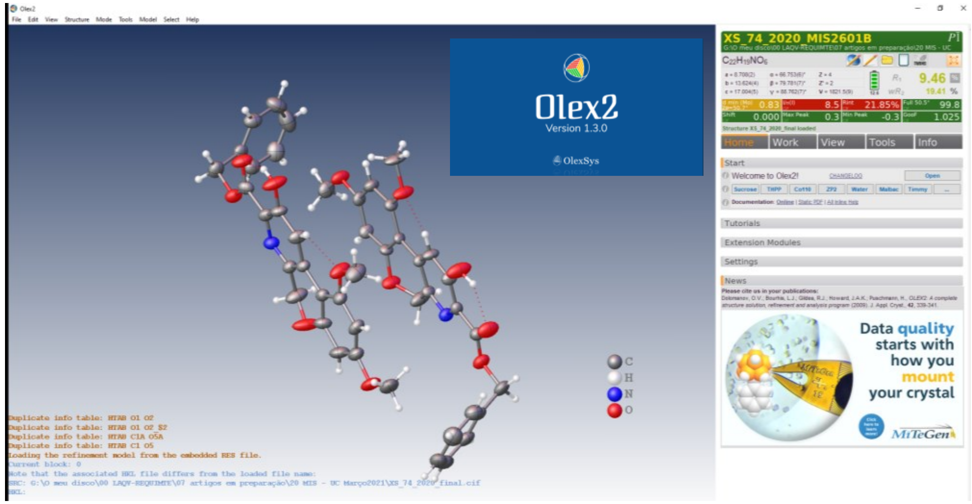
HKL FILE xs_22_2019_

DIRECTORY g:\o meu disco\01 sms_xray\data\2019\20190408_xs_22_20 ...

ACTIVE MODULE :none







SHELX

Input files

hkl file

h	k	l	I_{hkl}	σ_{hkl}
-1	0	0	0.10	0.30
-1	0	0	0.00	0.10
1	0	0	0.10	0.30
-2	0	0	4074.69	171.98
-2	0	0	3640.14	171.88
2	0	0	3451.65	171.88
-3	0	0	0.80	0.50
-3	0	0	0.80	0.70
-3	0	0	0.30	0.90
3	0	0	0.30	0.70
-4	0	0	441.16	20.70
-4	0	0	437.76	20.50
-4	0	0	440.06	20.90
-5	0	0	0.60	1.40
-5	0	0	2.80	2.10

INS file

```
TITL Polymorph1_OrtoR_Pbca in Pbca
CELL 0.71073   9.4493  18.1864  21.2044  90.000  90.000  90.000
ZERR   16.00   0.0014   0.0027   0.0031   0.000   0.000   0.000
LATT  1
SYMM 0.5-X, -Y, 0.5+Z
SYMM -X, 0.5+Y, 0.5-Z
SYMM 0.5+X, 0.5-Y, -Z
SFAC C H N
UNIT 176 160 32
TEMP 0
SIZE 0.20 0.30 0.60
TREF
HKLF 4
END
```

Unit cell parameters

Z, standard deviations

Lattice type (P = 1, I = 2, etc.). Negative for non-centrosymmetric space groups

symmetry operations for that space group

Structure Factors (element symbols) – the first two are C and H, if presente

number of atoms of each type in the unit cell

Temperature

Command for *Direct methods* using XS – omitted once the structure is solved

NOTE: The order of the above seven lines must not be changed, and you can't insert any other command lines in here.

SHELXL manual: shelxl_user_guide.pdf (uni-goettingen.de)

Output – RES file

```
TITL  SIR2019 run in space group P 21 21 21
CELL   0.71073    5.9661   9.0443  18.3969  90.000  90.000  90.000
ZERR         4    0.0001   0.0002   0.0003   0.000   0.000   0.000
LATT -1
SYMM 1/2-X,-Y,1/2+Z
SYMM 1/2+X,1/2-Y,-Z
SYMM -X,1/2+Y,1/2-Z
SFAC C H O S
UNIT 44 40 8 4
L.S.   10
ACTA
BOND   $H
CONF
TEMP     23.0
SIZE   0.20  0.20  0.20
WGHT    0.1
FMAP   2
PLAN     25
FVAR    1.0
S1      4   0.30911   0.18147   0.74051  11.00000   0.03598
O1      3  -0.16486   0.30225   0.67647  11.00000   0.05038
O2      3   0.34014  -0.08947   0.62924  11.00000   0.05054
C1      1  -0.06705   0.19484   0.64965  11.00000   0.03398
C2      1   0.18814   0.00031   0.62536  11.00000   0.03523
C3      1  -0.14245   0.11592   0.58259  11.00000   0.03502
C4      1   0.13646   0.12322   0.67275  11.00000   0.03482
C5      1   0.01049   0.00085   0.56751  11.00000   0.03547
C6      1  -0.32380   0.14069   0.53817  11.00000   0.04653
C7      1  -0.34885   0.04940   0.47656  11.00000   0.05297
C8      1  -0.20028  -0.06187   0.46199  11.00000   0.05293
C9      1  -0.01406  -0.08791   0.50787  11.00000   0.04622
C10     1   0.33791   0.37755   0.72867  11.00000   0.04718
C11     1   0.15067   0.17908   0.82280  11.00000   0.05429

HKLF      4
END
```

Number of least square refinement cycles to run
Writes a cif file.
Writes bond distances and angles (including for H atoms) to the cif file.

Weighting scheme (default 0.2)
How to write Q peaks. Default is 2.
Number of Fourier peaks (Q peaks). Default is 20.
Free Variables.

Columns:

1) Atom name,
2) refers to the SFAC line,
3-5) Fractional coordinates (x, y, z),
6) occupancy of the atom.
7) Displacement Parameter U