Refinement with REFMAC

DLS-CCP4 Workshop

2019

Diamond Light Source, Oxfordshire, UK

Oleg Kovalevskiy

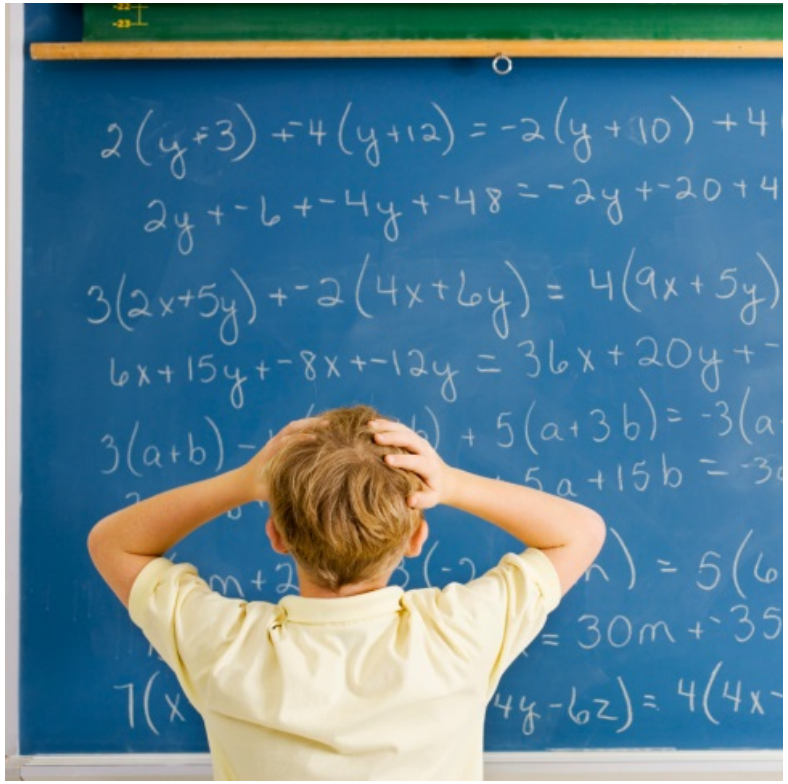
oleg.kovalevskiy@stfc.ac.uk

Science and Technology Facilities Council

CCP4

Contents

(almost) **No math due to time constraints, focus on practical usage**



Contents

- **Refinement**
  - Purpose of Refinement
  - Crystallographic Data
  - Model Parameterisation
  - Restraints
- **Low-resolution Refinement**
  - Jelly-body restraints
  - ProSMART External Restraints
  - LIBG Restraints
  - LORESTR Automated Pipeline

Purpose of Refinement

Crystallographic refinement has one major purpose:

<u>to fit atomic model into observed X-ray crystallographic data</u>

Model should agree with the observed data

Model must be chemically and structurally sensible

and one immediate most valuable consequence:

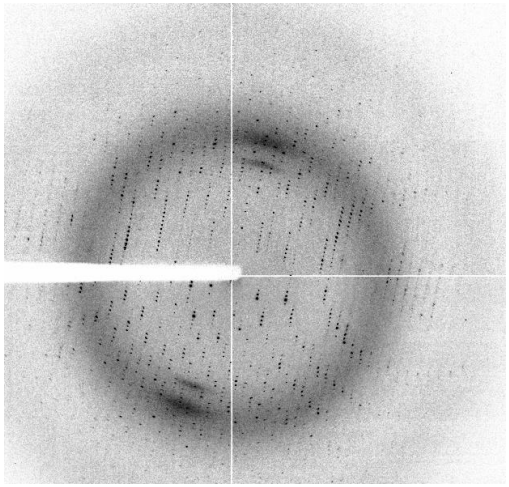
<u>to calculate best possible electron density map</u>

Allowing the atomic model to be visualised, criticised and analysed

Direct relation between model quality and phase quality

(corresponds to the quality of the electron density maps)

Fourier Transform

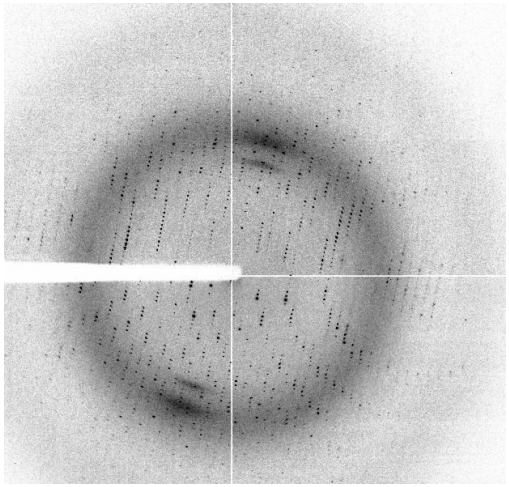


H, K, L	$\lvert F_{obs} \rvert$	ϕ
…		
5, 5, 5	348	-
5, 5, 6	392	-
5, 5, 7	157	-
5, 5, 8	312	-
…		

We have observed amplitudes: $\lvert F_{obs} \rvert$

But we don't have phases: φ

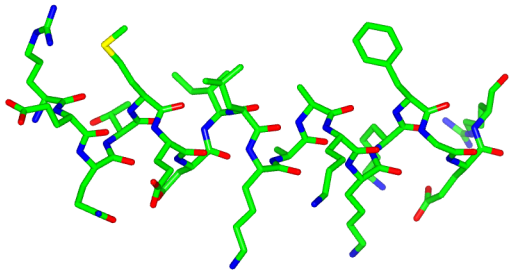
Fourier Transform



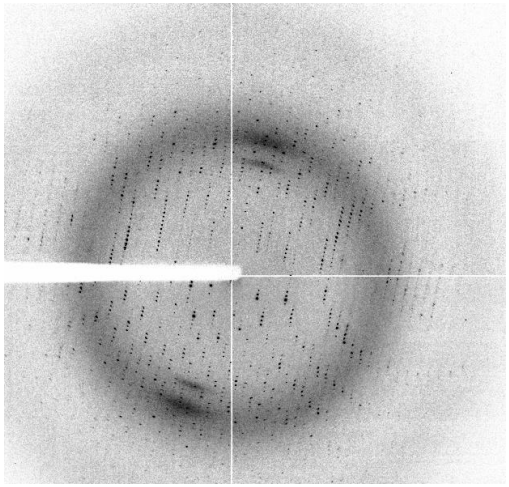
H, K, L	$\lvert F_{obs} \rvert$	ϕ
…		
5, 5, 5	348	-
5, 5, 6	392	-
5, 5, 7	157	-
5, 5, 8	312	-
…		

Suppose we have a starting model:

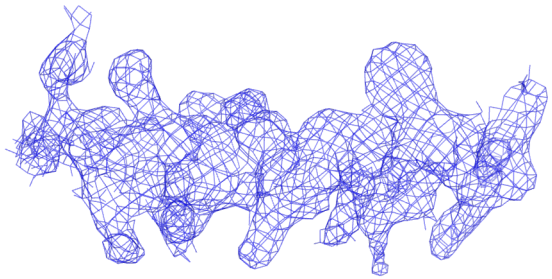
H, K, L	$\lvert F_{calc} \rvert$	ϕ_{calc}
…		
5, 5, 5	355	27°
5, 5, 6	387	8°
5, 5, 7	146	75°
5, 5, 8	340	31°
…		

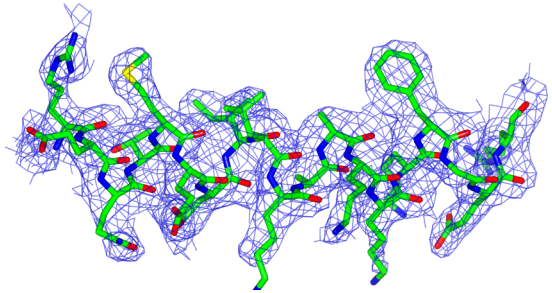


Model Refinement

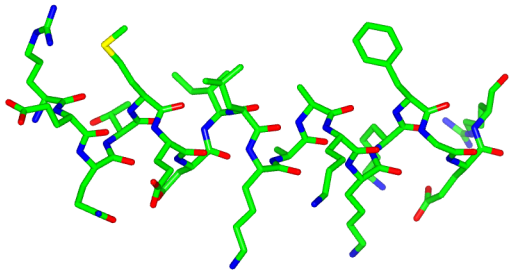


H, K, L	$\lvert F_{obs} \rvert$	ϕ
…		
5, 5, 5	348	-
5, 5, 6	392	-
5, 5, 7	157	-
5, 5, 8	312	-
…		





H, K, L	$\lvert F_{calc} \rvert$	ϕ_{calc}
…		
5, 5, 5	355	27°
5, 5, 6	387	8°
5, 5, 7	146	75°
5, 5, 8	340	31°
…		



Idea:

Iteratively improve the model, optimising the agreement between $\lvert F_{obs} \rvert$ and $\lvert F_{calc} \rvert$

Purpose: improve phase estimates: φ_{calc}

Model Refinement

Idea:

Iteratively improve the model to optimise the agreement between $|F_{obs}|$ and $|F_{calc}|$

*Note – we are **not** refining against a density map*

We are optimising the agreement between $|F_{obs}|$ and $|F_{calc}|$

How to assess correspondence between the model and experimental observations?

R-factor:

$$R = \frac{\sum \left| |F_{obs}| - |F_{calc}| \right|}{\sum |F_{obs}|}$$

Model Refinement

Refinement essentially tries to minimise the R-factor

How do we know that the model is reliable?

What if we improve the amplitudes $|F_{calc}|$ but worsen the phases φ_{calc}?

Such overfitting can happen if there are too many parameters

How to validate?

- **R_{free}** – reserve a portion of data for cross-validation (usually 5%)
- **Chemical & structural validation** – ensure that the model is physically sensible
- **Inspect electron density map** – manual intervention

Map Calculation

Two types of maps:

- $2F_{obs}-F_{calc}$: *"standard" electron density* – represents crystal contents
- $F_{obs}-F_{calc}$: *difference density* – represents differences

Maps are calculated using phase estimates from the current model: φ_{calc}

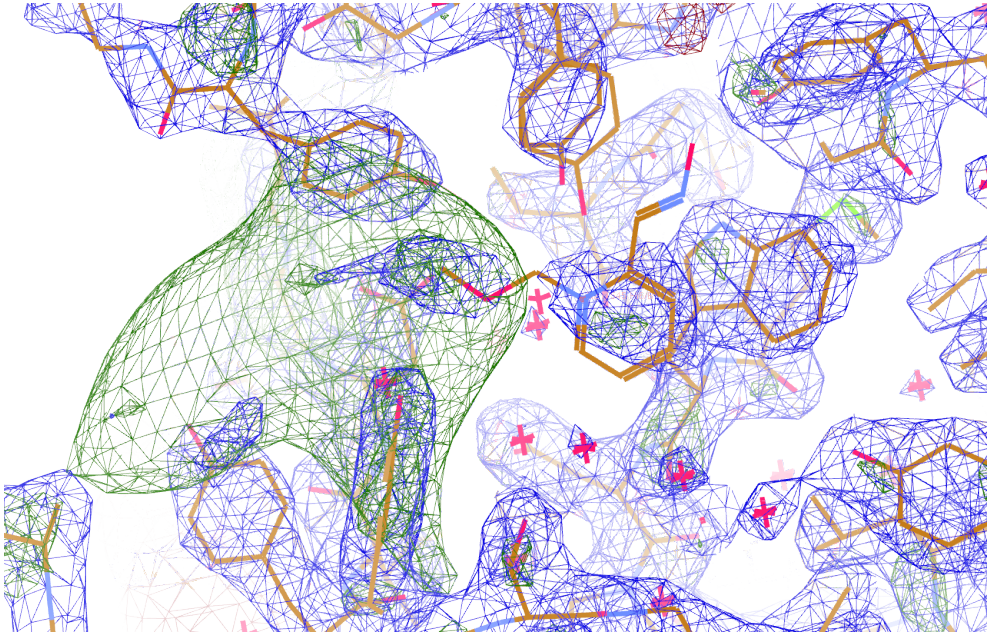
Model updated

↓

φ_{calc} updated

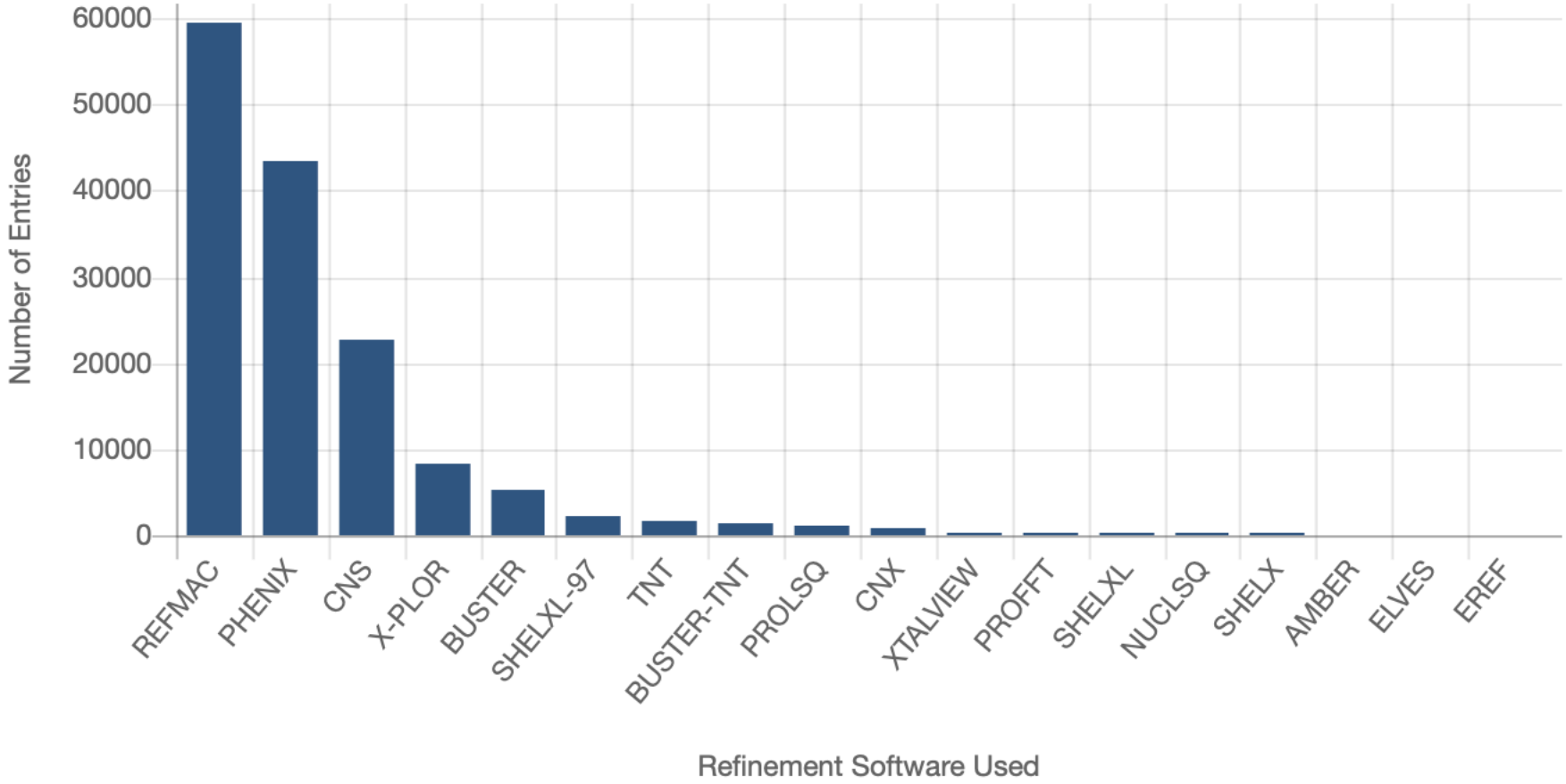
↓

Density changes



Note – contrast with real space refinement

Available Refinement Programs



https://www.rcsb.org/stats/distribution_software

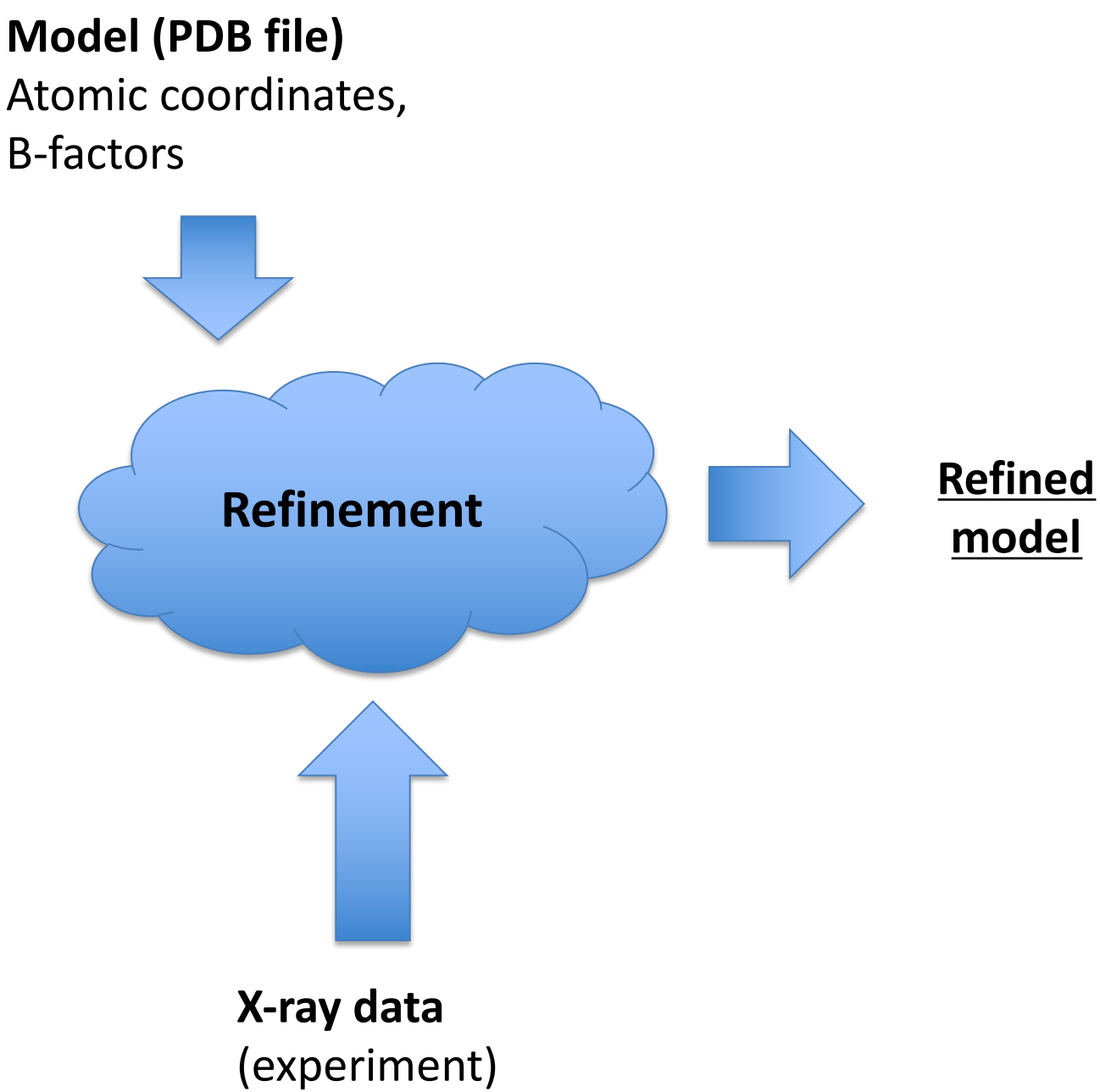
Your Crystal Peculiarities

Why there is no single universal refinement protocol?

Why do you need to adjust refinement parameters?

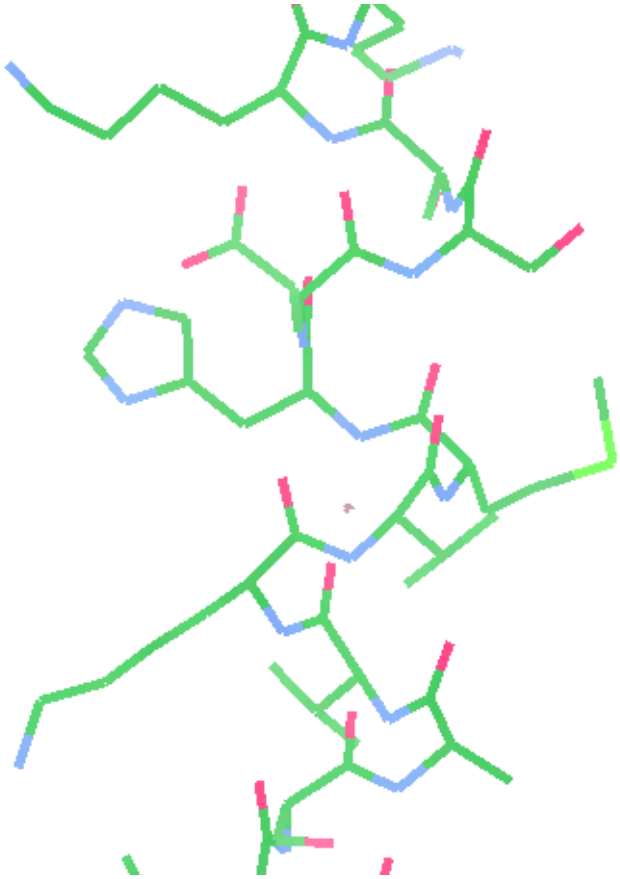
Three important things to consider:

1. You describe your data with the model. Data of different quality shall be described differently (mathematically speaking, different number parameters of the model could be estimated given particular data).
2. The model is defined not only by the PDB file, but also in the parameters of the refinement program.
3. Different quality of the model at different stages of refinement


Model (PDB file)
Atomic coordinates,
B-factors
Refinement
Refined model
X-ray data
(experiment)

Model Parameterisation

Standard refinable parameters



Atomic model:

- Position – (x,y,z) coordinates
- Uncertainty – B-factors
- (Occupancies)

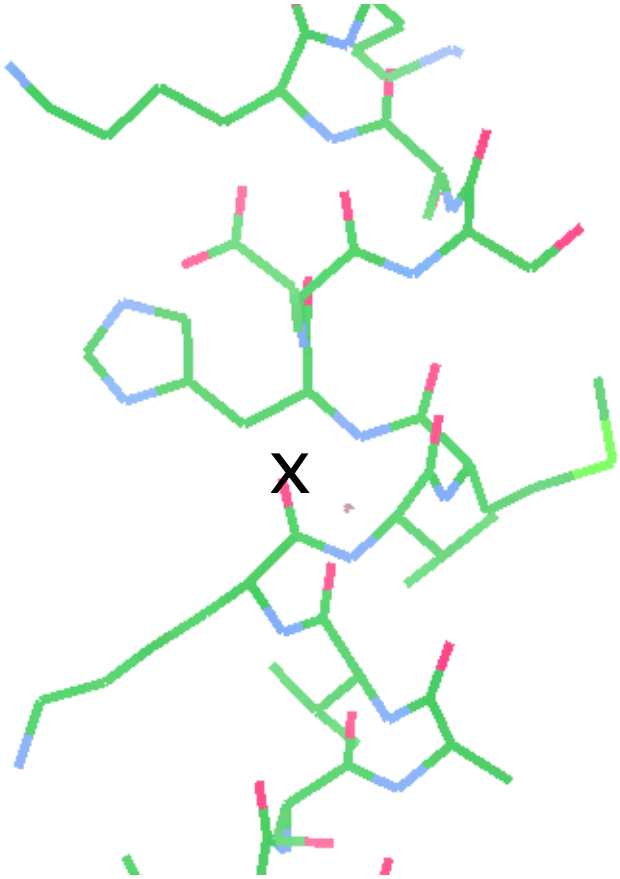
```
ATOM      5  CB  ASP A   8     -30.909   9.723  18.264  1.00 33.70          C
ATOM      6  CG  ASP A   8     -31.252   9.345  16.825  1.00 41.96          C
ATOM      7  OD1 ASP A   8     -31.072  10.248  15.981  1.00 46.18          O
```

Model Parameterisation

Standard refinable parameters

Atomic model:

- **<u>Position – (x,y,z) coordinates</u>**
- Uncertainty – B-factors
- (Occupancies)



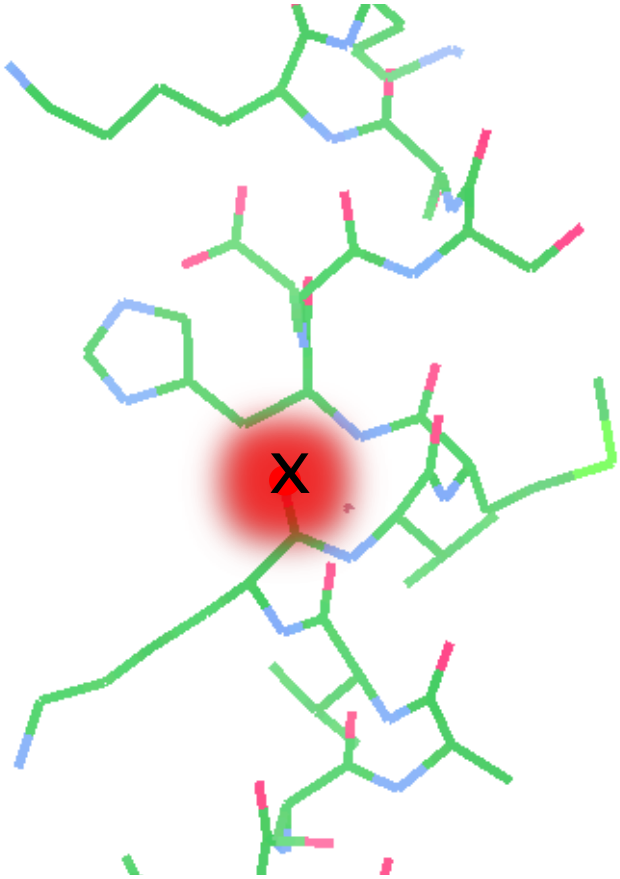

```
ATOM      5  CB  ASP A   8      -30.909   9.723  18.264  1.00 33.70          C
ATOM      6  CG  ASP A   8      -31.252   9.345  16.825  1.00 41.96          C
ATOM      7  OD1 ASP A   8      -31.072  10.248  15.981  1.00 46.18          O
```

Model Parameterisation

Standard refinable parameters

Atomic model:

- Position – (x,y,z) coordinates
- **<u>Uncertainty – B-factors</u>**
- (Occupancies)



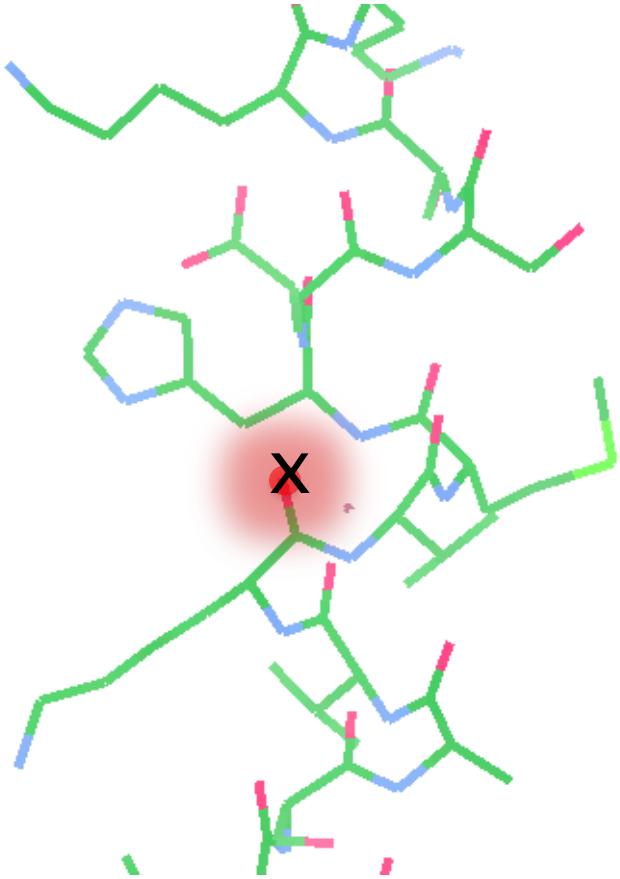

```
ATOM      5  CB  ASP A   8     -30.909   9.723  18.264  1.00 33.70          C
ATOM      6  CG  ASP A   8     -31.252   9.345  16.825  1.00 41.96          C
ATOM      7  OD1 ASP A   8     -31.072  10.248  15.981  1.00 46.18          O
```

Model Parameterisation

Standard refinable parameters

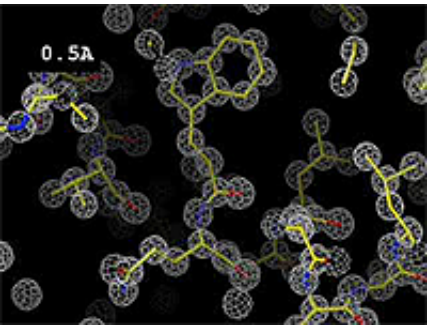
Atomic model:

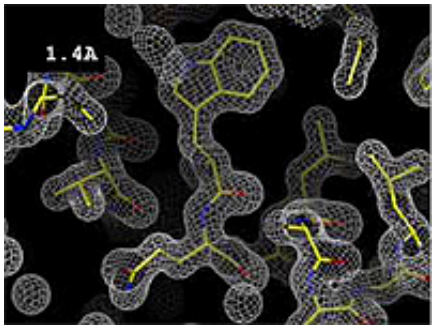
- Position – (x,y,z) coordinates
- Uncertainty – B-factors
- **<u>(Occupancies)</u>**

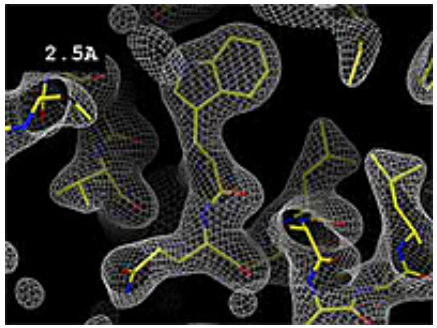


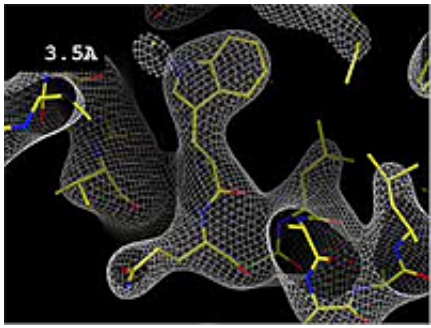

```
ATOM      5  CB  ASP A   8     -30.909   9.723  18.264  1.00 33.70          C
ATOM      6  CG  ASP A   8     -31.252   9.345  16.825  1.00 41.96          C
ATOM      7  OD1 ASP A   8     -31.072  10.248  15.981  1.00 46.18          O
```

Model Parameterisation



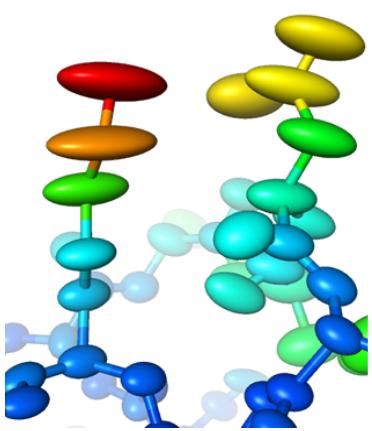
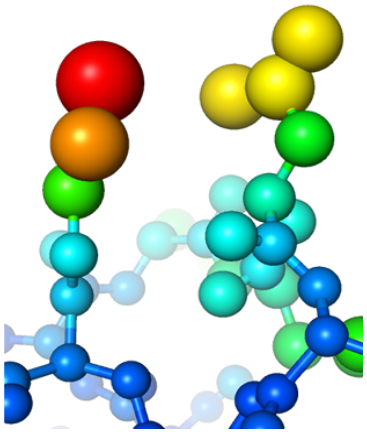




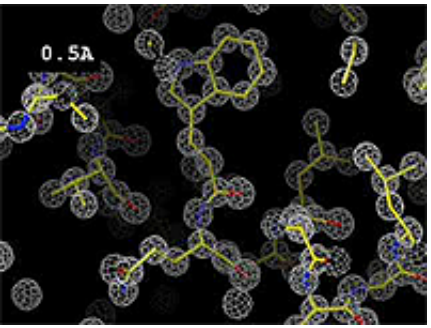


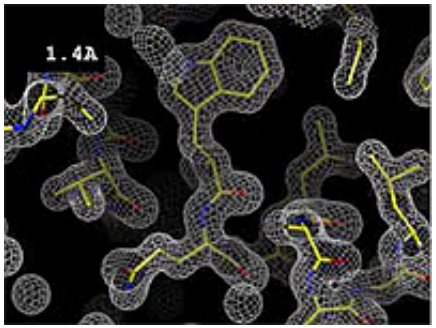
Notes on B-factors

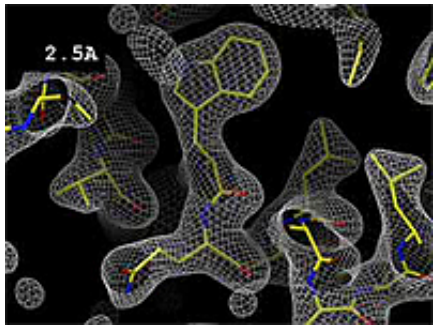
- Atomic B-factors can be modelled in a different way according to the quality of the data: **anisotropic** (6 parameters per atom), **isotropic** (1 parameter per atom), **TLS** (20 parameters per group of atoms)
- B-factors describe relative positional uncertainty

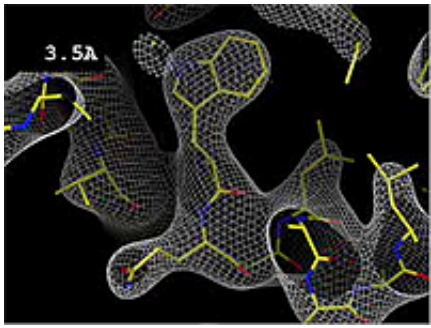



Model Parameterisation









Notes on B-factors

- Atomic B-factors can be modelled in a different way according to the quality of the data: **anisotropic** (6 parameters per atom), **isotropic** (1 parameter per atom), **TLS** (20 parameters per group of atoms)
- B-factors describe relative positional uncertainty
- B-factors are sometimes also referred to as atomic displacement parameters (ADPs) or thermal/temperature factors
- Should not compare atomic B-factors between different models

TLS Groups

Describe rigid body motion – e.g. for chains/domains/subunits

Suitable for medium resolution, when full anisotropy is impossible

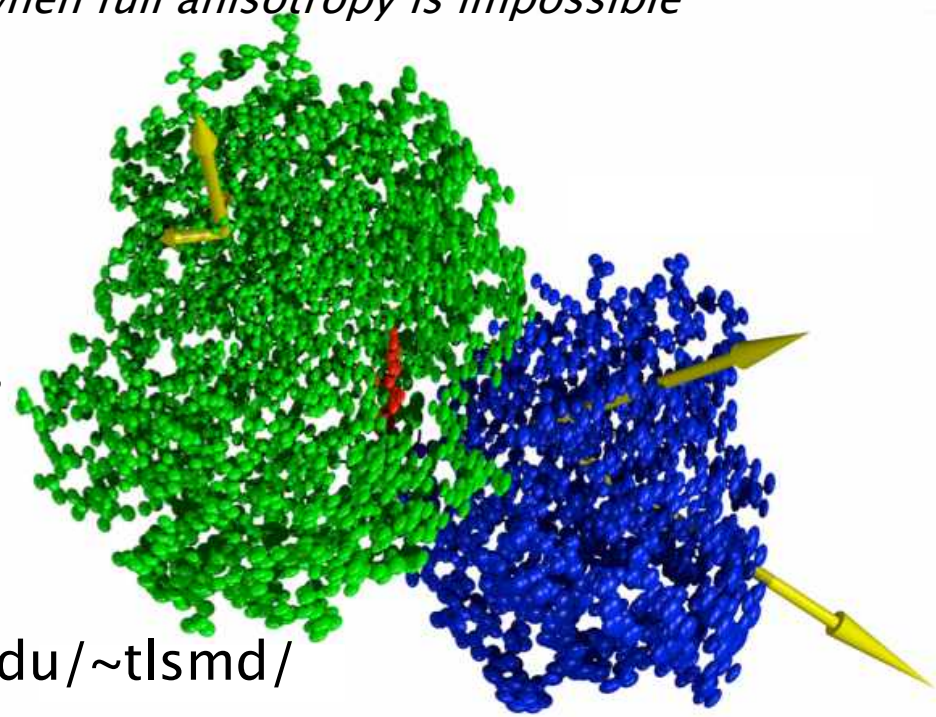
Per group (20 parameters):

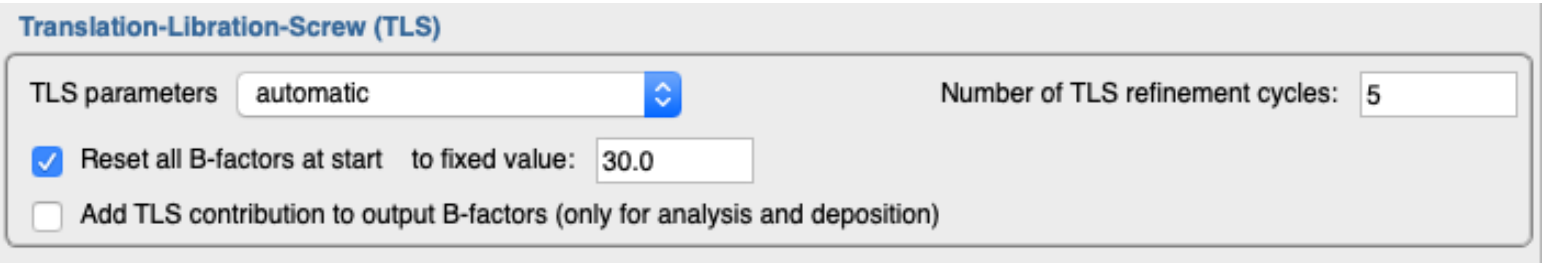
- **T**ranslation – 6 parameters
- **L**ibration – 6 parameters
- **S**crew rotation – 8 parameters

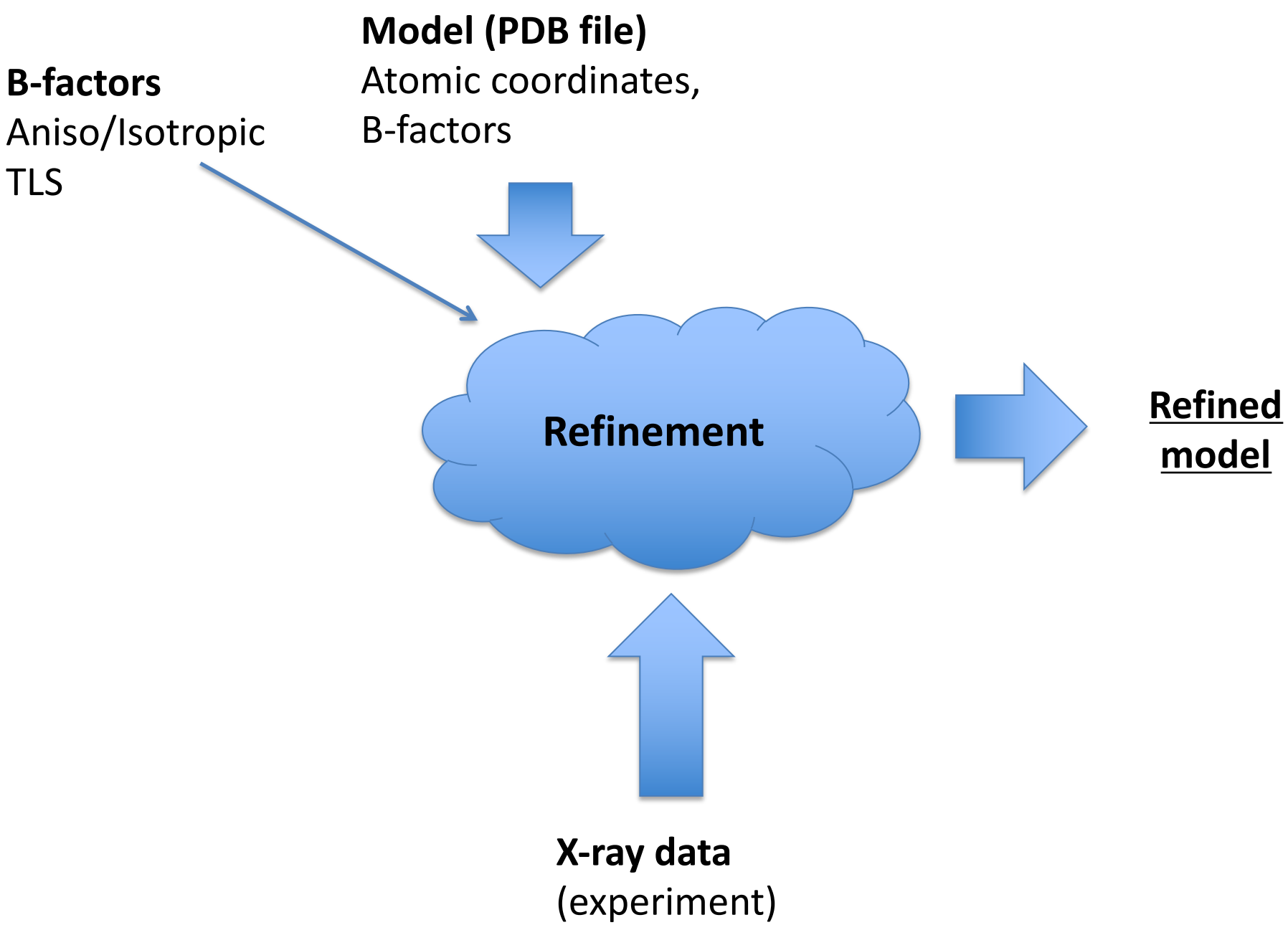
Define groups using CCP4i

or TLSMD webserver:

http://skuld.bmsc.washington.edu/~tlsmd/



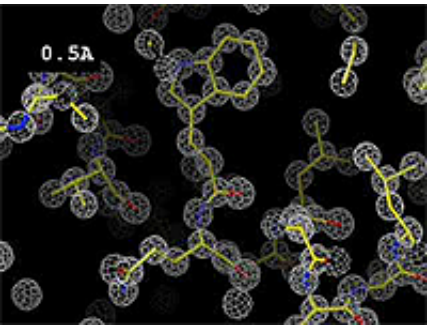



Model (PDB file)
Atomic coordinates,
B-factors
B-factors
Aniso/Isotropic
TLS
Refinement
Refined model
X-ray data
(experiment)

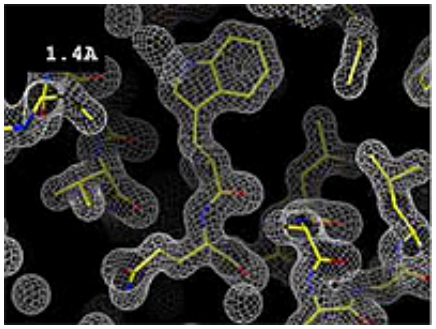
Model Refinement

Refinement strategy will differ for different quality of original data and it can also exploit particular features of your crystal:

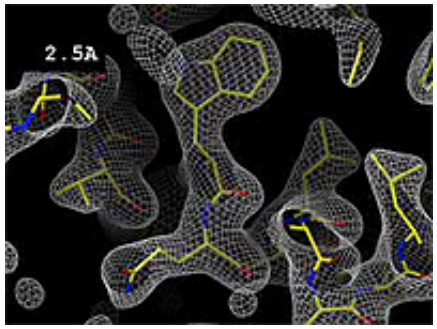
1.0Å **2.0Å** **3.0Å** **4.0Å**



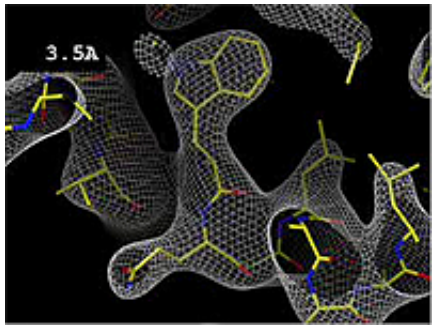
Unrestrained?



Chemical restraints



Chemical restraints

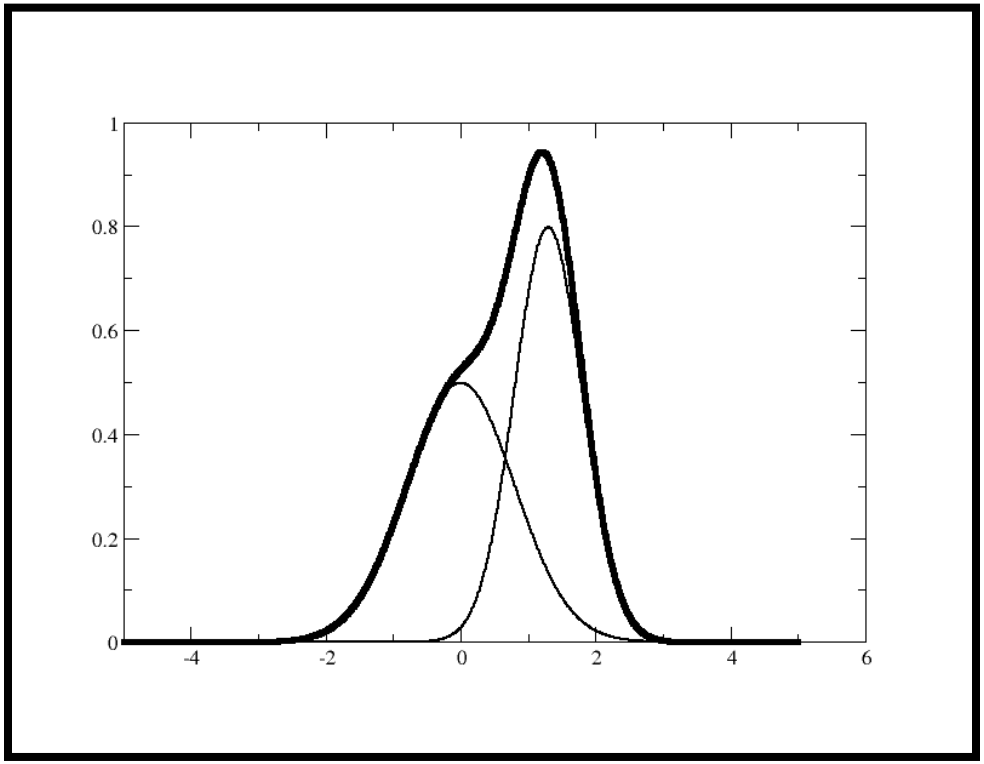


Chemical restraints
Jelly-body restraints
External restraints

Why Restraints?

Example: two-atom ideal case

Distance between atoms 1.3Å. B-factors 20 and 50



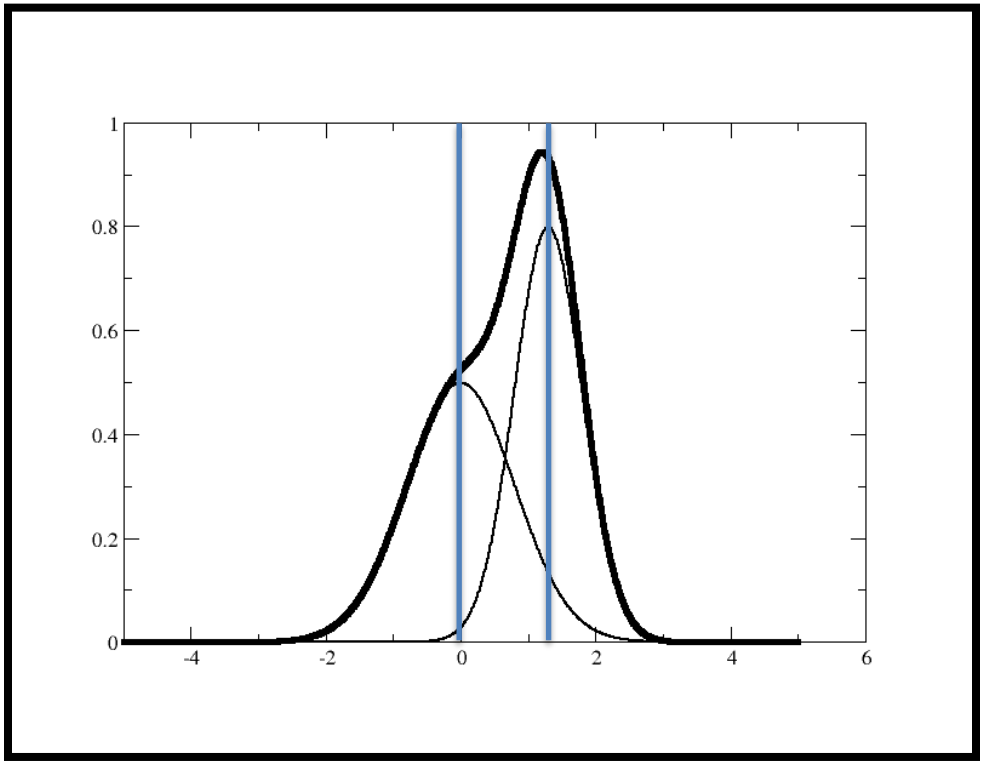
Thin lines – single atoms

Bold line – sum of the two atoms

Why Restraints?

Example: two-atom ideal case

Distance between atoms 1.3Å. B-factors 20 and 50



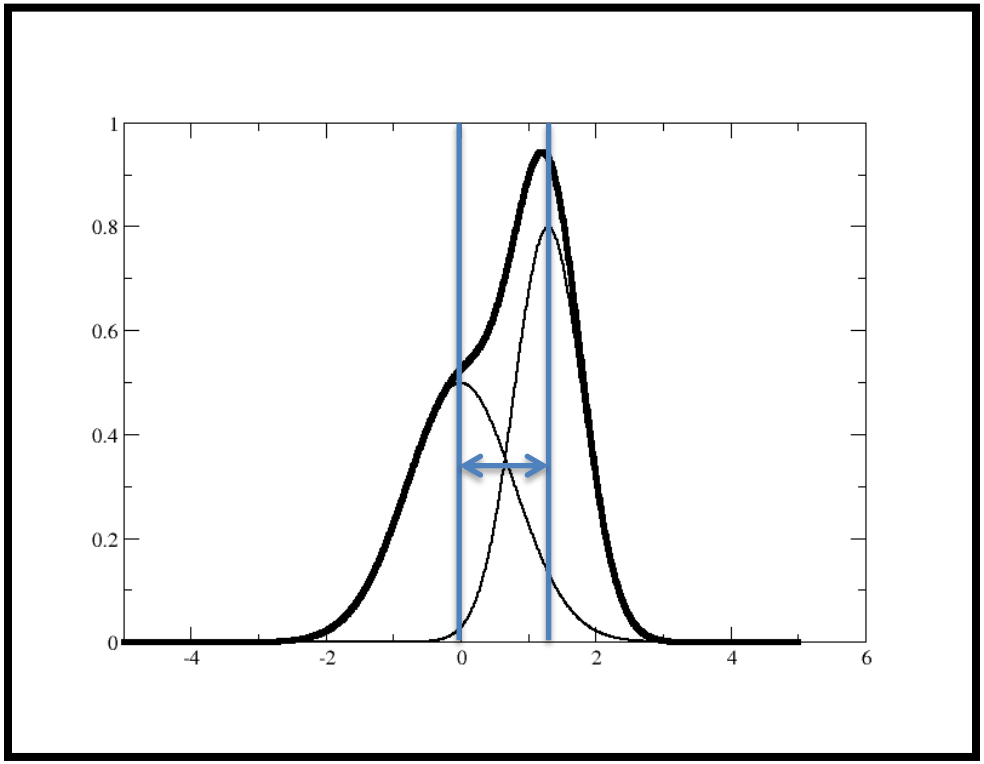
Thin lines – single atoms

Bold line – sum of the two atoms

Why Restraints?

Example: two-atom ideal case

Distance between atoms 1.3Å. B-factors 20 and 50



Thin lines – single atoms

Bold line – sum of the two atoms

Restraints

Standard restraints (used by default) include:

- Bond lengths
- Angles
- Chirals
- Planes
- Some torsion angles
- B-values
- VDW repulsions

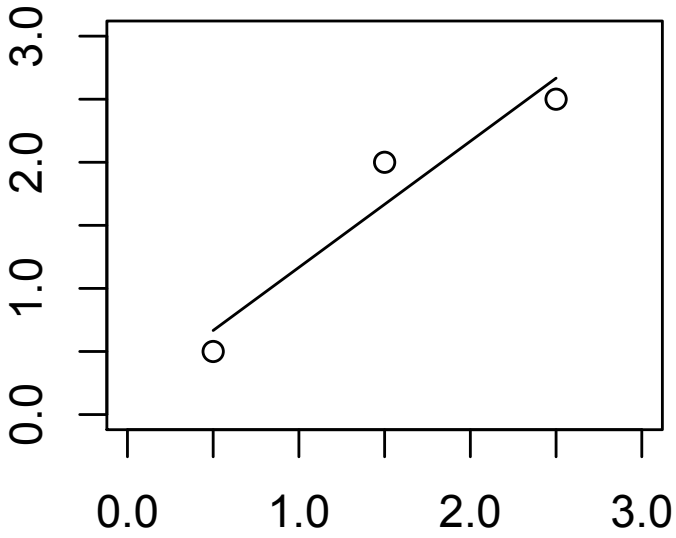
These help to ensure that the model is chemically sensible

Note – we generally deal with restraints, not constraints

Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.



Example: Fitting a line $y = a + bx$

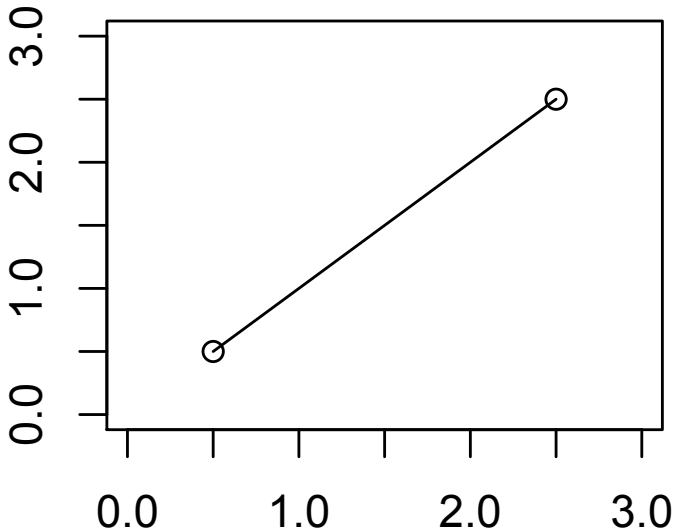
Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.

Can fit a line

Line is unreliable



Overfitting
Model Bias

Example: Fitting a line $y = a + bx$

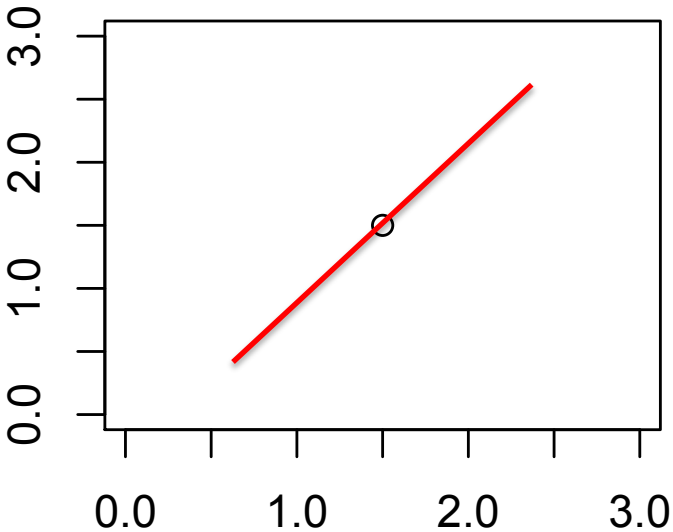
Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.

Insufficient observations!

Unstable refinement



Ill-posed problem

Example: Fitting a line $y = a + bx$

Restraints

How to improve the observation:parameter ratio.

1. Reduce number of parameters

N_{obs} is resolution dependent…

High resolution	:	Anisotropic B-factors	– 9 params per atom
Low resolution	:	Isotropic B-factors	– 4 params per atom

- Rigid body refinement – 9 parameters per body

Restraints

How to improve the observation:parameter ratio.

1. Reduce number of parameters

2. Increase number of restraints

   - B-value restraints
   - NCS restraints
   - H-bond and secondary-structure restraints
   - Restraints to homologous known structures
   - Nucleic acid base-pair and base-stacking restraints
   - Jelly-body restraints

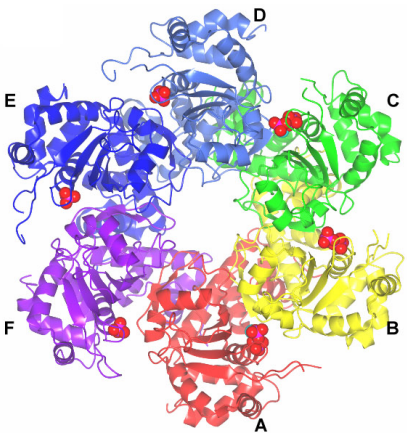
What do we know about Macromolecules?

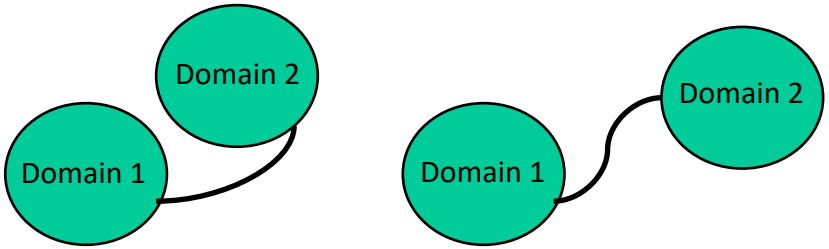
Introduce additional restraints based on our knowledge:

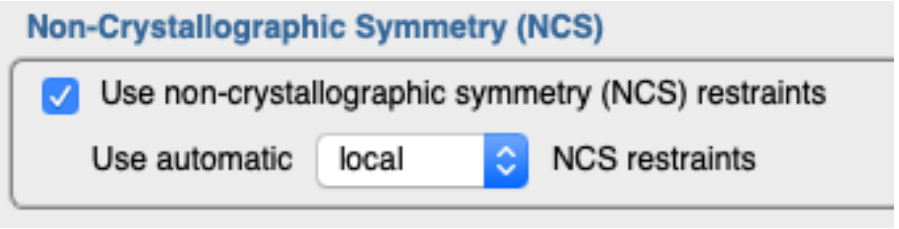
1. Macromolecules consist of atoms bonded to each other in a specific way
   ↳ **Specified by current state of the model**
2. Oscillation of atoms close to each other in 3D cannot be dramatically different
   ↳ **B-factor restraints, TLS restraints**
3. If there are two copies of the same molecule present then they will likely be similar to each other
   ↳ **NCS restraints – local/global depending on resolution**
4. If there are two molecules with sufficiently high sequence identity then it is likely that they will be structurally similar
   ↳ **External restraints to homologous structures – ProSMART**
5. Proteins tend to form secondary structures
   ↳ **Generic H-bonding restraints – ProSMART**
6. DNA/RNA tend to form base-pairs, stacked bases tend to be parallel
   ↳ **Generic base-pair and stacking restraints – LibG**

NCS
(Non-Crystallographic Symmetry Restraints)

1. NCS constraints
   - NCS-related copies are considered to be exactly the same
   - Only one set of atomic parameters per molecule is refined
2. Global NCS restraints
   - Molecules are superimposed
   - Difference between atoms are minimised
3. Local NCS restraints
   - Molecules are assumed to be locally similar
   - However, they may adopt (slightly) different global conformations
   - Restrain differences between local interatomic distances







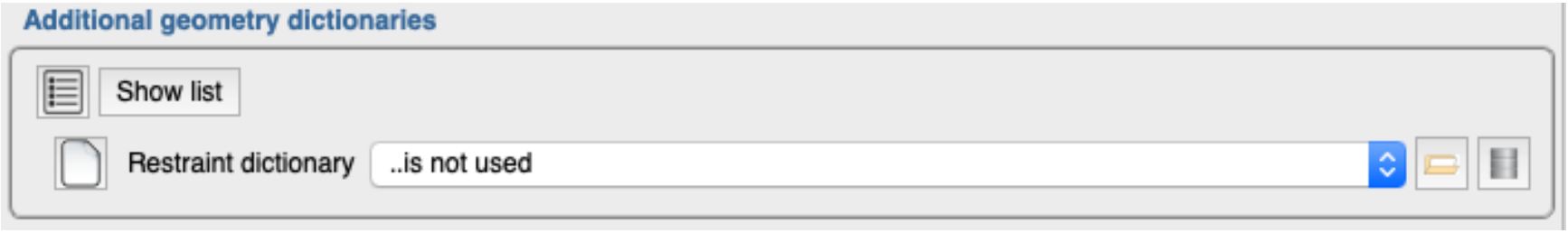
Ligand Refinement

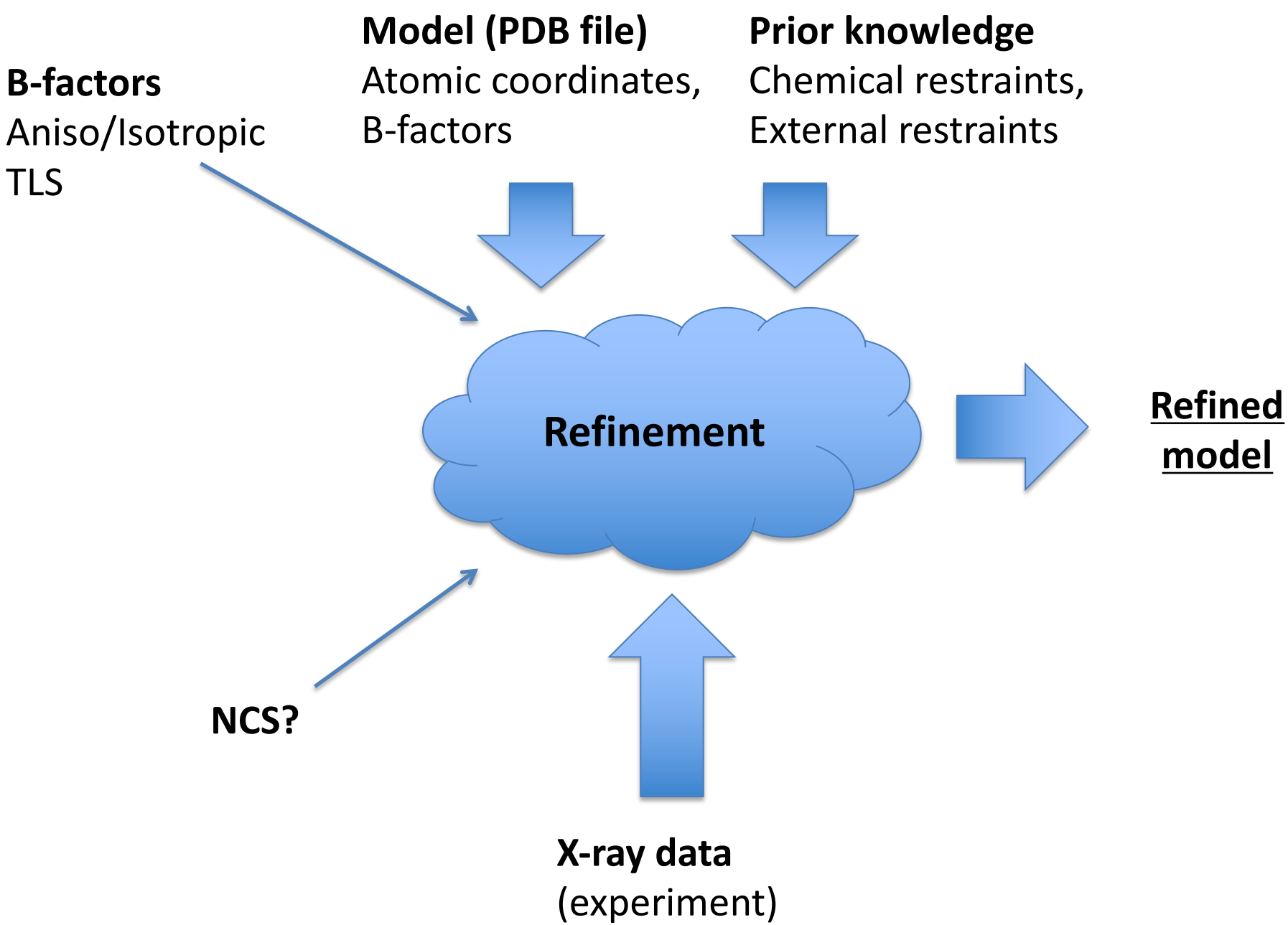
Geometric restraints for protein / nucleic acids are pre-tabulated

Ligands are more complicated

Need a source of prior information

- Common/known structures are dealt with automatically
  - CCP4/REFMAC monomer library has pre-computed descriptions
    - 13000 monomers
    - 60 modifications
    - 70 link entries
- New ligands require description (CIF file)
  - Right tool – ACEDRG




Model (PDB file)
Atomic coordinates,
B-factors
Prior knowledge
Chemical restraints,
External restraints
B-factors
Aniso/Isotropic
TLS
Refinement
Refined model
NCS?
X-ray data
(experiment)

$f_{xray} = -\log[P(obs;model)]$

likelihood of the data

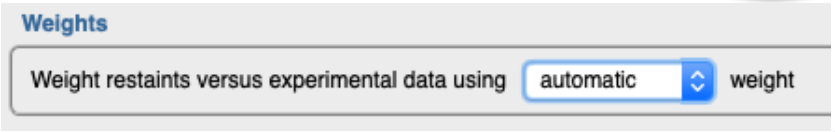
$f_{geom} = -\log[P(model)]$

probability of the model

w : relative weighting

Prior knowledge
Chemical restraints,
External restraints

$$f_{tot} = wf_{xray} + f_{geom}$$



X-ray data
(experiment)

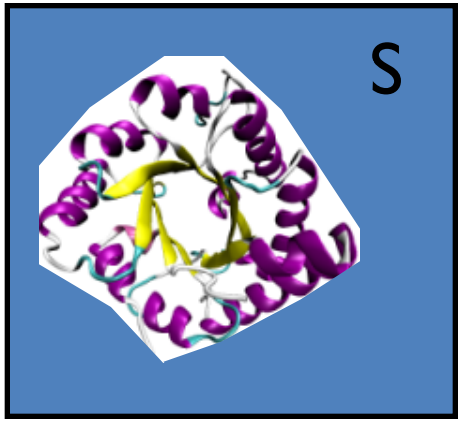
Solvent model

Mask-based bulk solvent correction

Idea:

- Protein region is masked out
- Solvent region is flattened (set to constant)
- Structure factors for solvent are calculated: $F_{solvent}$

$$F_{total} = F_{protein} + \alpha F_{solvent}$$



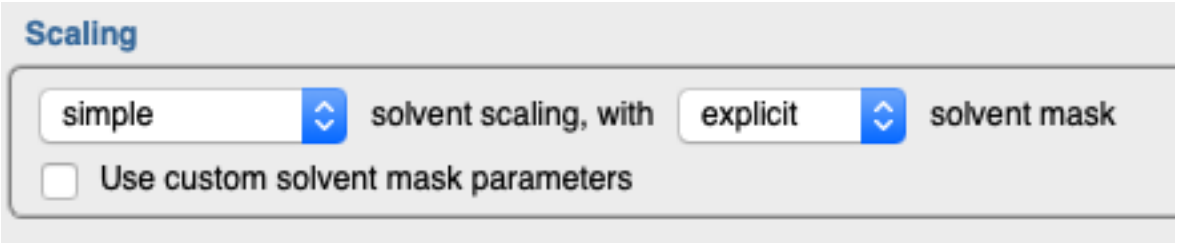
α : solvent scale factor

$$k\, e^{-Bs^2}\, e^{-s^T U s}\, (F_{protein} + \alpha F_{solvent})$$

k : overall scale factor

B : overall B-factor

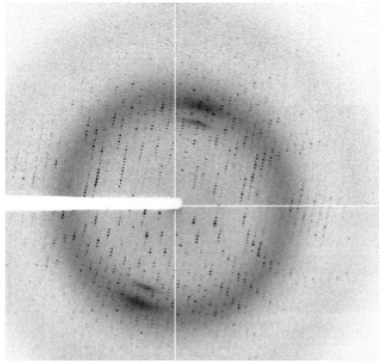
U : overall anisotropic B-factor



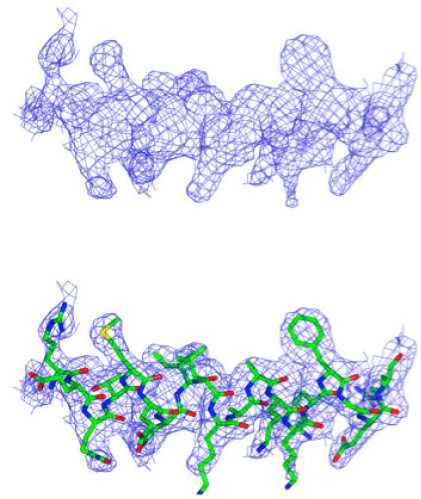
Overall Parameters: Scaling

Problem:

- Observed and calculated amplitudes need to be brought to the same scale so that they can be compared



H, K, L	$\lvert F_{obs} \rvert$	ϕ
...		
5, 5, 5	348	-
5, 5, 6	392	-
5, 5, 7	157	-
5, 5, 8	312	-
...		

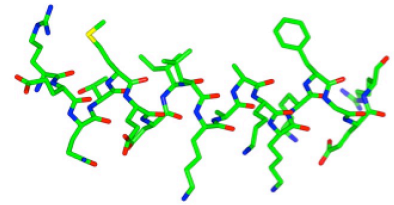


Idea:

Iteratively improve the model, optimising the agreement between $\lvert F_{obs} \rvert$ and $\lvert F_{calc} \rvert$

Purpose: improve phase estimates: φ_{calc}

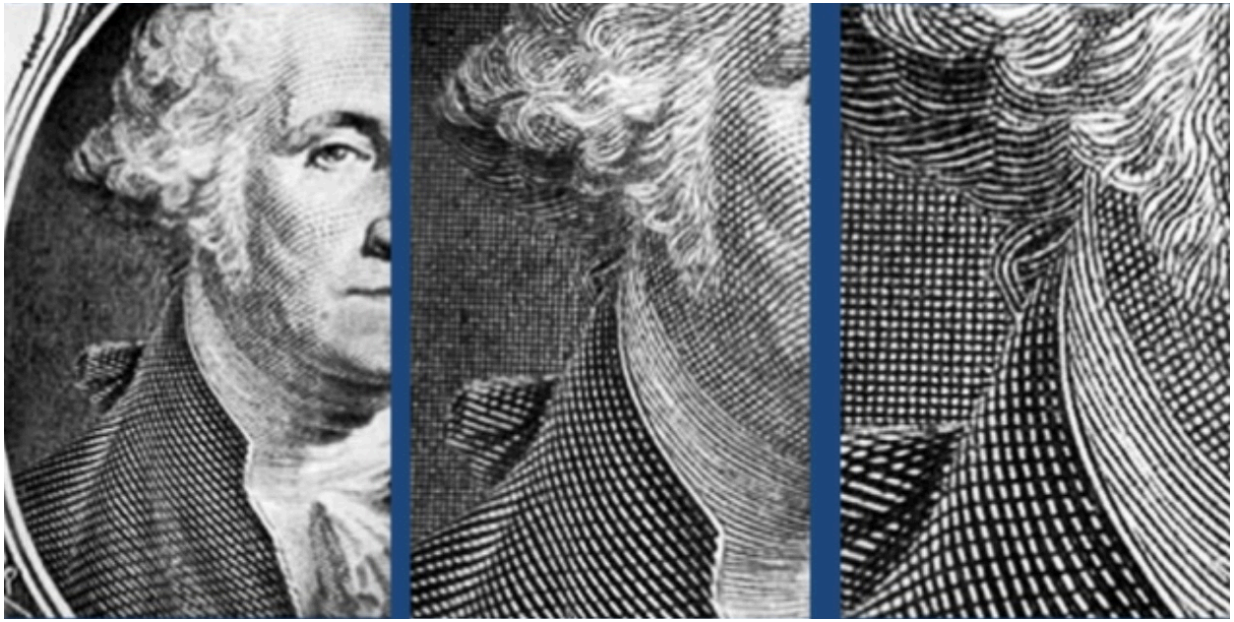
H, K, L	$\lvert F_{calc} \rvert$	ϕ_{calc}
...		
5, 5, 5	355	27°
5, 5, 6	387	8°
5, 5, 7	146	75°
5, 5, 8	340	31°
...		



Overall Parameters: Scaling

Problem:

- Observed and calculated amplitudes need to be brought to the same scale so that they can be compared



Overall Parameters: Scaling

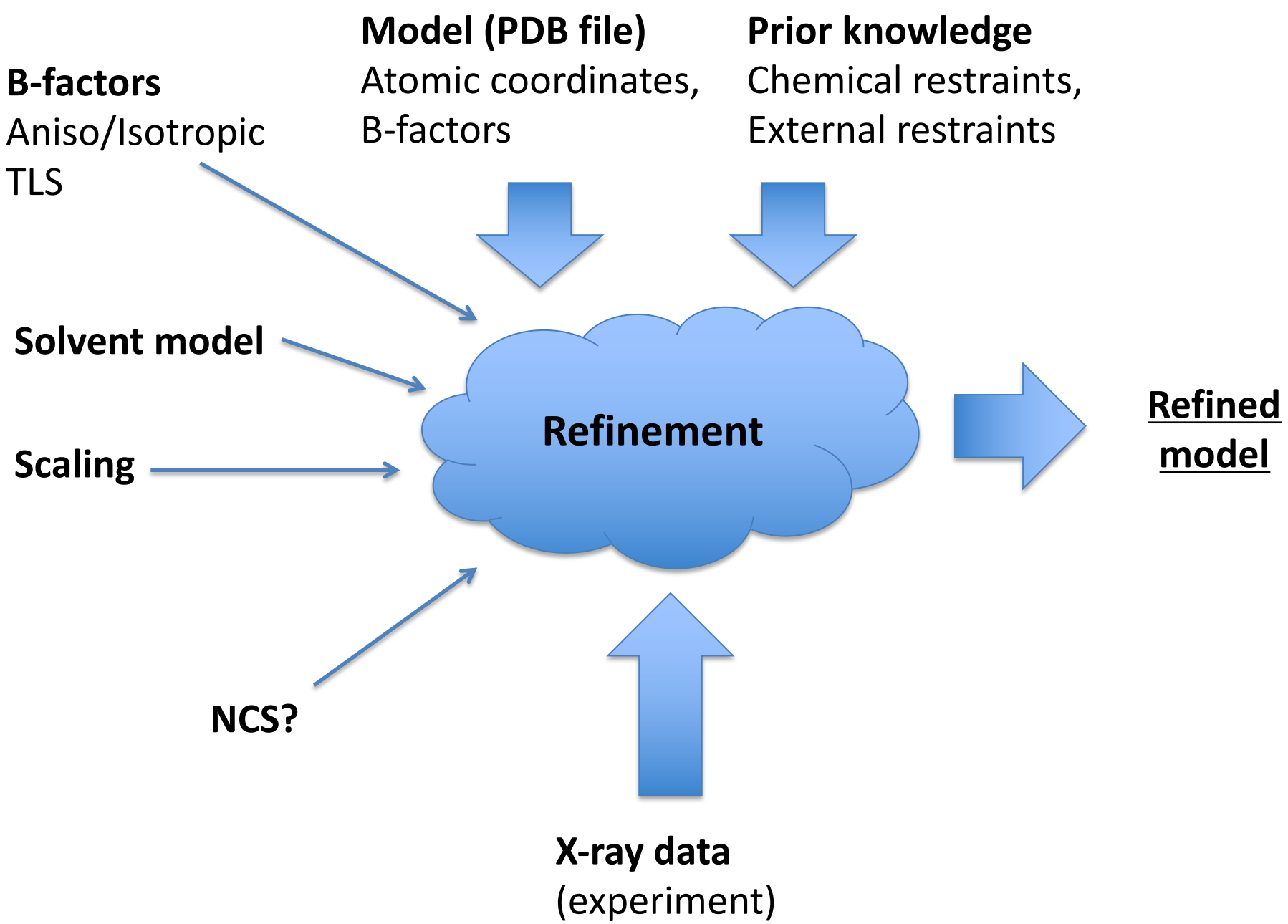
Problem:

- Observed and calculated amplitudes need to be brought to the same scale so that they can be compared

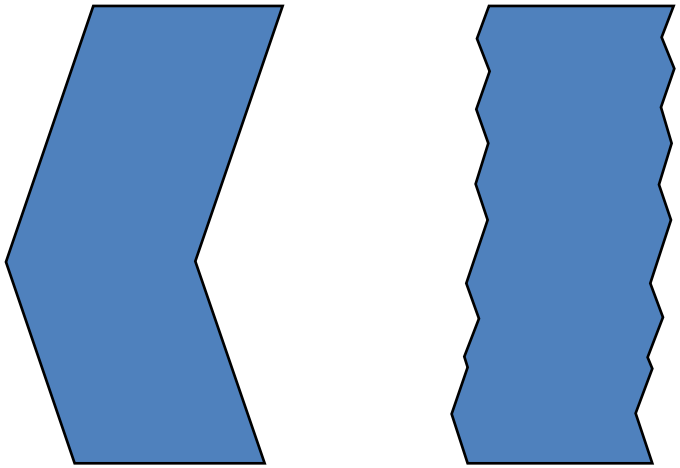
Need to:

- Modify/scale F_{calc}
- Find a scaling function, with some parameters – "overall parameters"

Scale parameters are optimised in ML refinement, along with all other parameters


Model (PDB file)
Atomic coordinates,
B-factors
Prior knowledge
Chemical restraints,
External restraints
B-factors
Aniso/Isotropic
TLS
Solvent model
Scaling
NCS?
Refinement
Refined model
X-ray data
(experiment)

Twin Refinement



A single crystal can be cut out of the twin: yes no

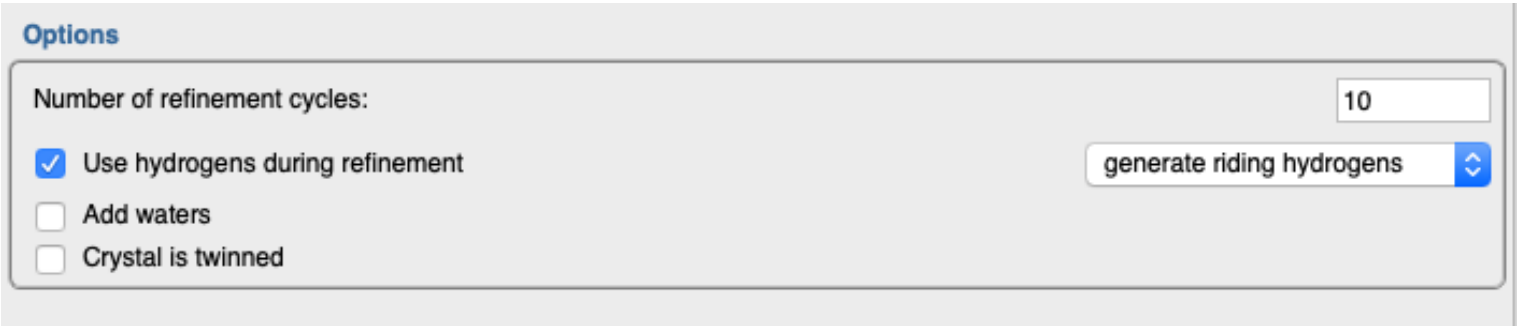
Need to deal with polysynthetic twin during refinement

Twin Refinement

Twin refinement in REFMAC5 is automatic

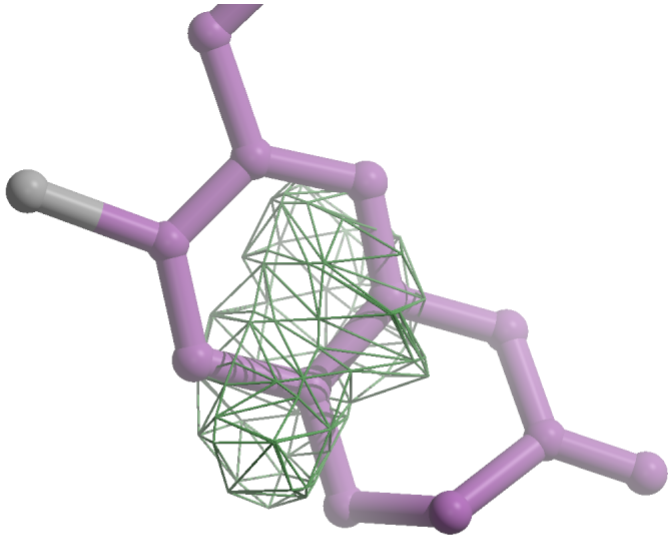
1. Identify potential twin operators
2. For each operator, calculate R_{merge} (R-factor comparing twin-related intensities)
3. If R_{merge}>0.44 remove this operator
4. Refine twin fractions
5. Keep only sufficiently large domains (default 7%)

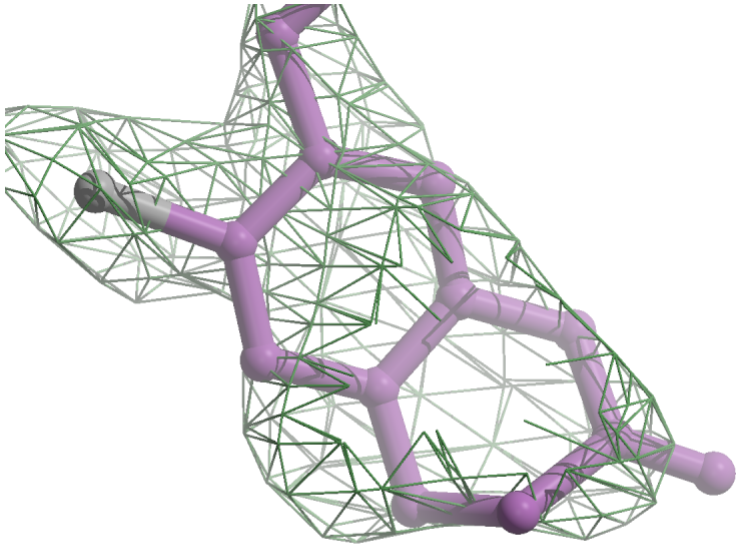
DON'T USE TWIN REFINEMENT IF YOUR Rfree IS HIGHER THAN 40%!



Twin Refinement

Example: Where's the density for my ligand (2.15A)?





F_o-F_c (3σ)

R/R_{free} = 25.5/26.9%

After initial rigid body and restrained refinement

R/R_{free} = 15.9/16.3%

Re-refine with twin on (twin fractions: 0.6/0.4)

Borrowed from Ben Bax, ex-GSK

Stages of model building

Early stages (e.g. straight after MR)

- Run many cycles (up to 200 with jelly body restraints) of refinement

Medium stages – during model building

- 10-20 cycles
- Optimise additional parameters like NCS, TLS, etc
- Once your Rfree is lower than 40%, turn on twinning if twinned
- Play with geometry weight parameter

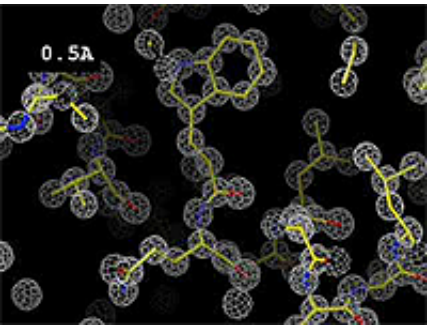
Final stages of refinement

- Even fewer cycles?
- Use optimal additional parameters
- If geometrical quality of the model is not optimal, further play with parameters like geometry weight.

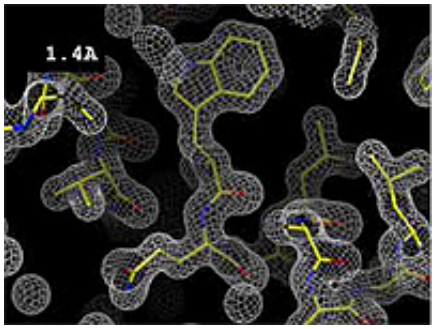
Model Refinement

Refinement strategy will differ for different quality of original data and it can also exploit particular features of your crystal:

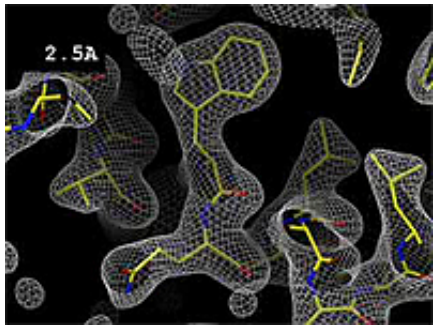
1.0Å **2.0Å** **3.0Å** **4.0Å**



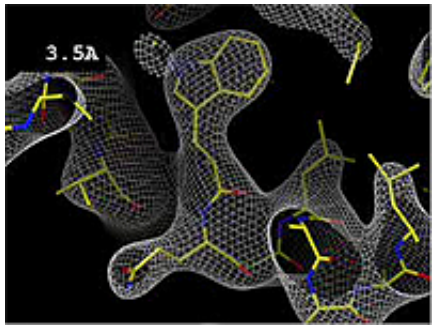
Unrestrained
Anisotropic B
Hydrogens



Chemical restraints
Isotropic B
TLS

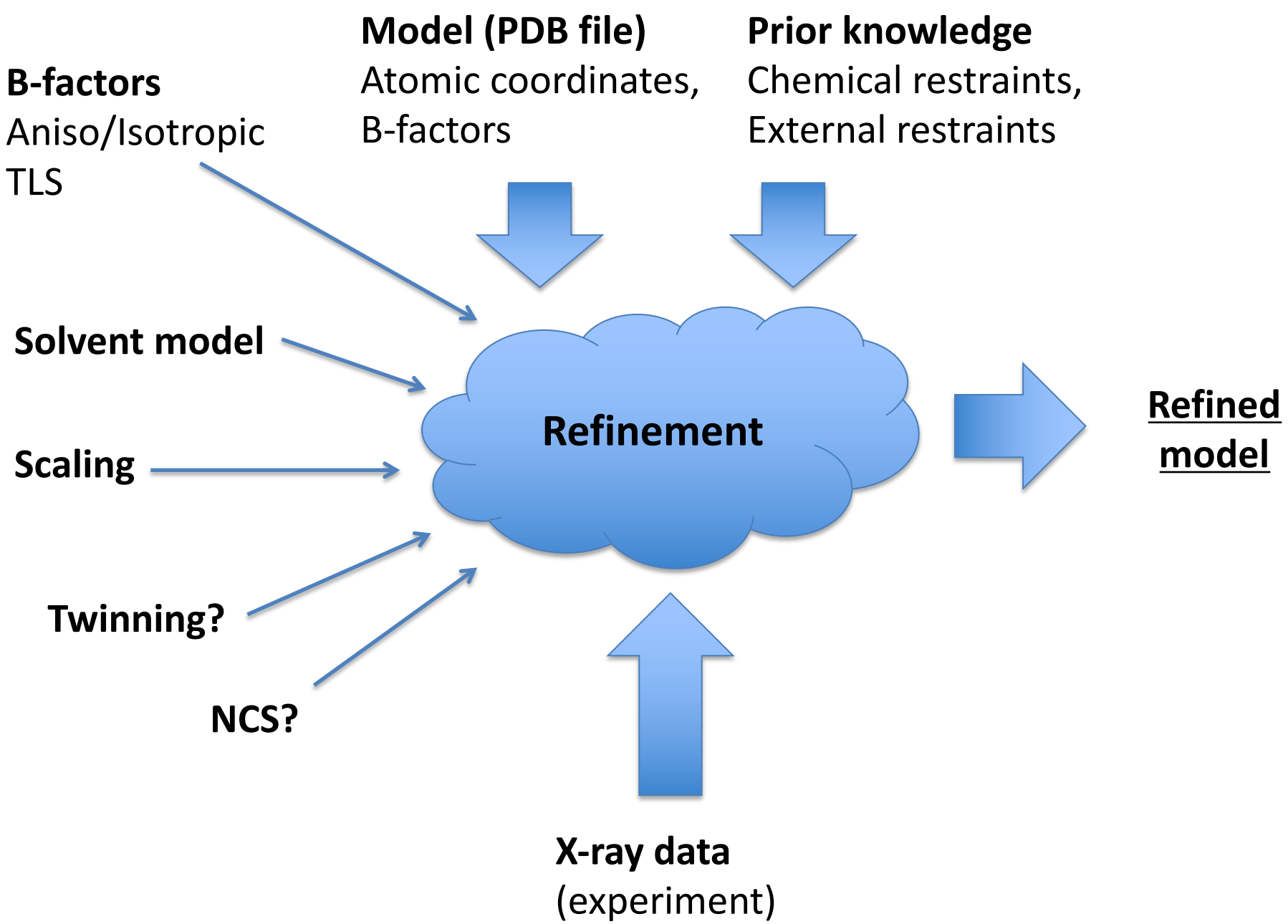


Chemical restraints
Isotropic B
TLS



Chemical restraints
Jelly-body restraints
External restraints
Isotropic B

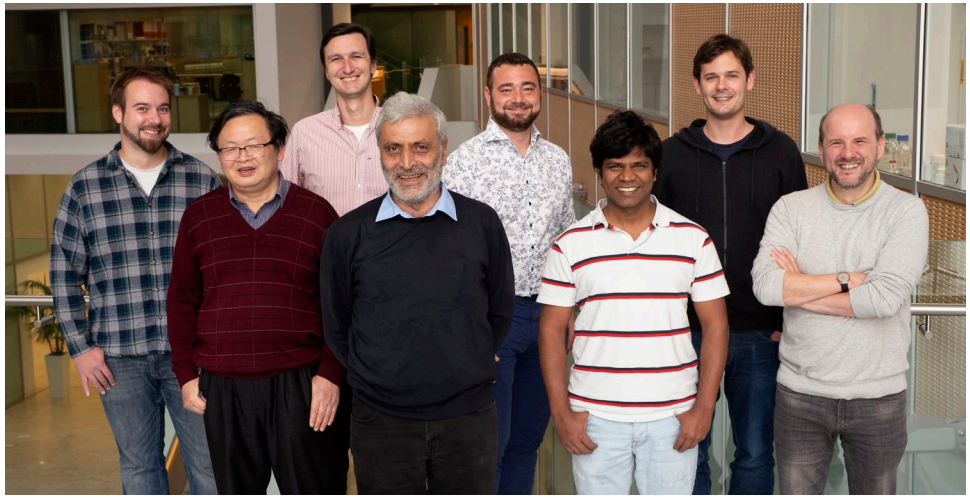
- Twinning
- Several similar subunits – NCS (at moderate and low resolution)
- Available homologues for external restraint generation (low resolution)


Model (PDB file)
Atomic coordinates,
B-factors
Prior knowledge
Chemical restraints,
External restraints
B-factors
Aniso/Isotropic
TLS
Solvent model
Scaling
Twinning?
NCS?
Refinement
Refined model
X-ray data
(experiment)

Acknowledgements

Contact: **oleg.kovalevskiy@stfc.ac.uk**
www2.mrc-lmb.cam.ac.uk/groups/murshudov/

MRC-LMB Computational Structural Biology Group



Left to right:

Rob Nicholls
Fei Long
Oleg Kovalevskiy
Garib Murshudov
Michal Tykac
Rangana Warshamanage
James Parkhurst
Paul Emsley

CCP4 Core

Eugene Krissinel
Andrey Lebedev
Charles Ballard
Ronan Keegan
Ville Uski

Global Phasing

Marcin Wojdyr

CCP4i2

Martin Noble
Stuart McNicholas
Jon Agirre
Liz Potterton

CCP-EM

Martyn Wynn
Tom Burnley
Colin Palmer

Collaborators

Marcus Fischer
Robbie Joosten
Andrea Thorn
Roberto Steiner
Alan Brown
Jude Short
Ana Casañal
Rafiga Masmaliyeva
Azzurra Carlon

Computing

Jake Grimmett
Toby Darling

All colleagues from
MRC-LMB, CCP4, CCP-EM
Users for feedback!

MRC Laboratory of Molecular Biology

CCP4

BBSRC bioscience for the future

UKRI Science and Technology Facilities Council