GGNB Course A57
Macromolecular Structure Determination I
Part III: Phasing Methods, Model Building, and Refinement

Tim Grüne
Dept. of Structural Chemistry, University of Göttingen
September 2011

http://shelx.uni-ac.gwdg.de

tg@shelx.uni-ac.gwdg.de
Isomorphous Replacement
Isomorphous Replacement

Molecular Replacement “borrows” the phases from a (putatively) homologous search model.

The other two methods, isomorphous replacement and anomalous dispersion, indirectly calculate the phases. These two methods belong to the experimental phasing techniques.
Isomorphous Replacement - Some History

Myoglobin, the first macromolecular structure ever to be solved by X-ray crystallography (John Kendrew et al, 1958), used isomorphous replacement to solve the phase problem.

Myoglobin (and also Hemoglobin, Max Perutz, 1959) naturally bind a heavy metal (iron) which make it suitable for isomorphous replacement.

The technique is based on measuring (at least) two datasets:

1. one in the presence of a heavy metal (the derivative dataset)
2. one in the absence of the same heavy metal (the native dataset)

Since molecular replacement was not available because of the lack of search models and since anomalous dispersion was not available because there were not synchrotrons, isomorphous replacement was for a long time the only choice for phasing.
Isomorphous Replacement - Measurement

derivative

Measure

native

---
Isomorphous Replacement - Measurement
Isomorphous Replacement - Measurement

Measure

Subtract

Calculate
Isomorphous Replacement - Summary

• The difference between the intensities from the native and the derivative dataset is like a dataset from the ligand only with the same unit cell dimensions.
• Only for a heavy atoms (Au, Pt, Fe, Ag, Hg) is the difference big enough to be detected over the noise of the data.
• The coordinates of the heavy atom can be calculated with small-molecule-methods. These include so-called direct methods and Patterson methods.
The Patterson Map

The Patterson map of a molecule show peaks for each connection (not bond) of any two atoms.

- Imagine a “small molecule”.

[Diagram showing connections between atoms labeled 1 to 5]
The Patterson Map

The Patterson map of a molecule show peaks for each connection (not bond) of any two atoms.

- Imagine a “small molecule”.
- For each “connection” draw a peak in the coordinate system.
The Patterson Map

The Patterson map of a molecule show peaks for each connection (not bond) of any two atoms.

- Imagine a “small molecule”.
- For each “connection” draw a peak in the coordinate system.
- Do not forget the inverse direction.
The Patterson Map

The Patterson map of a molecule show peaks for each connection (not bond) of any two atoms.

- Imagine a “small molecule”.
- For each “connection” draw a peak in the coordinate system.
- Do not forget the inverse direction.
- For 5 atoms there are 20 connections.
The Patterson Map

The Patterson map of a molecule show peaks for each connection (not bond) of any two atoms.

- Imagine a “small molecule”.
- For each “connection” draw a peak in the coordinate system.
- Do not forget the inverse direction.
- For 5 atoms there are 20 connections.
- $1 \leftrightarrow 3$ and $3 \leftrightarrow 5$ share the same peak.
The Patterson Map

The Patterson map of a molecule show peaks for each connection (not bond) of any two atoms.

- Imagine a “small molecule”.
- For each “connection” draw a peak in the coordinate system.
- Do not forget the inverse direction.
- For 5 atoms there are 20 connections.
- $1 \leftrightarrow 3$ and $3 \leftrightarrow 5$ share the same peak.
- Every atom connects to itself.
Meaning of the Patterson Map

The very same pattern (the Patterson map) we just calculated from the structure can also be obtained from the measured intensity data of a crystal from this structure.

Technically, the second method is a convolution of the data with itself.

The principle path is:

data set $\Rightarrow$ Patterson map $\Rightarrow$ atom coordinates
Interpretation of the Patterson Map

For large molecules with more than a couple of atoms, the Patterson map quickly becomes overcrowded and not interpretable any more.
Interpretation of the Patterson Map

For large molecules with more than a couple of atoms, the Patterson map quickly becomes overcrowded and not interpretable any more.

For small molecules, though, where most peaks can be located, one can actually invert the process just presented and deduce the atom positions from the Patterson map.
From Substructure to Phases
Structure Factor per Atom

The intensity $I(hkl)$ is the square of the structure factor amplitude.
Structure Factor per Atom

The intensity $I(hkl)$ is the square of the structure factor amplitude.

- Each atom contributes to the structure factor amplitude.
Structure Factor per Atom

The intensity $I(hkl)$ is the square of the structure factor amplitude.

- Each atom contributes to the structure factor amplitude.
- Each contribution can be calculated from the position of the particular atom.
Structure Factor per Atom

The intensity $I(hkl)$ is the square of the structure factor amplitude.

- Each atom contributes to the structure factor amplitude.
- Each contribution can be calculated from the position of the particular atom.
- With isomorphous replacement, one single contribution is sufficient to calculate the phase $\phi(hkl)$ (up to two-fold ambiguity).
Solving the Phase Problem with the Harker Construction

- The heavy atom positions can be calculated from the difference between derivative and native data set using the Patterson map.
- The coordinates of the heavy atom(s) allow to calculate the structure factors for the heavy atoms.
- The Harker construction requires
  - The structure factor \( F_{h.a.}(hkl) \) of the heavy metal (calculated).
  - The intensity \( I_{\text{der}}(hkl) \) of the derivative structure (measured).
  - The intensity \( I_{\text{nat}}(hkl) \) of the native structure (measured).
- The Harker construction delivers
  - the phase \( \phi(hkl) \) of the native and the derivative structure.
Harker Construction

The Harker construction has to be carried out for every measured reflection separately.

- Draw the calculated structure factor from the heavy metal in the complex plane.

$\text{Im}(F(hkl))$

$F_{\text{h.a.}}(hkl)$

$\text{Re}(F(hkl))$
Harker Construction

The Harker construction has to be carried out for every measured reflection separately.

- Draw the calculated structure factor from the heavy metal in the complex plane.
- Draw a circle around its end-point with radius \( \sqrt{I_{\text{der}}(hkl)} \).
Harker Construction

The Harker construction has to be carried out for every measured reflection separately.

- Draw the calculated structure factor from the heavy metal in the complex plane.
- Draw a circle around its end-point with radius $\sqrt{I_{\text{der}}(hkl)}$.
- Draw a circle around the origin with radius $\sqrt{I_{\text{nat}}(hkl)}$. 
Harker Construction

The Harker construction has to be carried out for every measured reflection separately.

- Draw the **calculated** structure factor from the heavy metal in the complex plane.
- Draw a circle around its end-point with radius $\sqrt{I_{\text{der}}(hkl)}$.
- Draw a circle around the origin with radius $\sqrt{I_{\text{nat}}(hkl)}$.
- The structure factor of the native data **must** be one of the two intersection points.
Harker Construction - Resolving the Ambiguity

The two-fold ambiguity can be resolved by collecting a second derivative dataset with a second type of heavy metal.
MIR and SIR

Using two or more different derivative datasets (in addition to the native dataset) is called *Multiple Isomorphous Replacement* (MIR).

When using only one derivative dataset, it is called *Single Isomorphous Replacement* (SIR). The twofold ambiguity is then resolved simply by taking the mid-point of both choices.

While this seems like a very crude assumption, it is usually good enough to have starting phases.
Isomorphous Replacement - Lack of Isomorphism

Isomorphous replacement provides us with an estimate of the real phases - unlike Molecular replacement, where they come from a different structure. So Isomorphous replacement does not suffer from model bias.

The main disadvantage of isomorphous replacement is that one needs at least two data sets from two different crystals, the native and the derivative one. When preparing the derivative crystal, either by soaking the crystal in a solution with the heavy metal compound, or by co-crystallisation, the crystal can change a little: the native and the derivative can have different unit cell dimensions.

Changes in the unit cell as small as 0.5% can already render SIR and MIR as unsuccessful!
Anomalous Dispersion
Anomalous Dispersion

Isomorphous replacement is based on finding the coordinates of some heavy metal compound based on comparing two different data sets: the native and the derivative.

Anomalous Dispersion techniques follow the same goal but with a different approach: the breakdown of Friedel's law
Friedel’s law

Under normal conditions, the reflections $I(hkl)$ and $I(-h - k - l)$ are related with each other:

$$I(hkl) = I(-h - k - l) \quad \phi(hkl) = -\phi(-h - k - l)$$

This relationship is called Friedel’s Law, and the two reflections $(hkl)$ and $(-h - k - l)$ are called a Friedel Pair.
Breakdown of Friedel’s law

When the energy of the X-rays is at or beyond the transition energy for a shell transition of an atom, Friedel’s Law is no longer true. This is called breakdown of Friedel’s Law. Since every atom contributes to every reflection, we find that $I(hkl) \neq I(−h − k − l)$ (and the phases do not match anymore, either) even when only some of the atom types in the crystal match the wavelength.

In this case, the two reflections $(hkl)$ and $(−h − k − l)$ are called a Bijvoet Pair.

For the major part of atom types in organic compounds ($N, C, O, H$), X-rays are not in the range where anomalous dispersion occurs. Like with isomorphous replacement one has to rely on heavy atoms as in the case of MIR/SIR.
Substructure Solution

The positions of the atoms that are affected by anomalous dispersion make the so-called substructure. Similar to isomorphous replacement, the substructure can be solved by small molecule techniques, but in this case from the differences of the Bijvoet pairs (instead of the difference between a native and a derivative dataset).

The contribution of the substructure to the whole diffraction experiment can be extracted from the anomalous data with the help of an equation derived by J. Karle (1980) and by W. A. Hendrickson, J. L. Smith, and S. Sheriff (1985).
MAD and SAD Data Collection

Once the intensities from the substructure atoms are known, the positions of the substructure atoms can be found as in the case of isomorphous replacement, using Patterson or direct methods, and also the subsequent procedure of deducing the phases for the total structure factors is the same.

In contrast to MIR, the 2-fold ambiguity is not resolved by using different heavy atoms, but by collecting data at different wavelengths: the strength of anomalous signal varies with wavelength.

This is called MAD, multi-wavelength anomalous dispersion, as opposed to SAD, single wavelength anomalous dispersion, which uses the mean of the two possibilities.
(Dis-)Advantages of MAD/ SAD

The main advantage of MAD and SAD phasing: only one single crystal is required, i.e. the problem of non-isomorphism between different crystals is overcome.

Disadvantages

**MAD:** 1. Adjustable Wavelength - one needs a synchrotron
   2. Long exposure: 2-3 datasets from one crystal can lead to radiation damage, i.e. destruction of the crystal by formation of free radicals during exposure.

**SAD:** The data must be very accurate, since the differences between the Bijvoet pairs is very small (about 1% of the total intensity).
More Advantages of MAD/ SAD

- **SeMet**: (biological) replacement of $S$ with $Se$ in Met-residues
  - “natural” state of protein $\rightarrow$ crystallisation conditions similar to native protein
  - feasible in most recombinant expression systems (bacterial, insect cells, yeast, mammalian cells; see H. Walden, Acta Cryst 2010, D66)

- **S-SAD**: SAD phasing based on anomalous signal of $S$ in Cys-residues (and Met, but Met is often too flexible for this weak signal)
  - Convenient, because one single crystal without further manipulation suffices
  - requires very accurate data, preferably better than 2Å on inhouse source.
Resolution Ranges of Phasing Methods

MR  In principle no resolution limit, but worse than, say, 2.5-2.8Å model bias becomes a serious problem

MAD resolution better than 3.0Å has a good chance of success, but also 4-5Å can work.

SAD approx. 2.5Å; for S-SAD: 2.0Å
With todays lecture we have gathered

- a long list of reflections
- an estimate for the phase of each reflection

We can now calculate our first electron density map from which we start model building and refinement.
Part IV: Model Building & Refinement
Resolution: Example Images

Low resolution:
Helix region of a molecular replacement solution at 3.4 Å.
“Humps” for the side chains can be seen, but not identified.
No “staircase” helix, rather a rod/cylinder.
Resolution: Example Images

**Low resolution:**
Loop or coil region of the same molecule at 3.4 Å.
Breaks in density, no density for side chains.
Resolution: Example Images

Medium to high resolution:
Thermolysin at 1.9 Å.
Side chains can be distinguished (one Phe-ring even shows a hole).
Single atoms are not visible, but e.g. S in Met shows stronger density than C, N.
Resolution: Example Images

Ultra high resolution:
Guanine in a 0.95 Å DNA structure.
Separate atoms visible - so “visible” that there is not even connected density for the main chain.
Model building would be: “place atom” - “name atom”.
The actual result from our experiments is an electron density map $\rho(x, y, z)$.
The Purpose of the Model

Scientists run experiments to get results. The electron density map is the result of our X-ray experiment. So why don’t we stop here?

The map of a is difficult to interpret and unhandy. It does not tell much about the (bio-)chemistry of the underlying molecule.

The model brings much more use to the map. It distinguishes atom types, shows what residues are involved, and e.g. what residues/atoms are involved in the active site, . . .
The PDB-file
The PDB-file is the most common format for macromolecular structural information. Its content can be displayed in many ways.

- ball–and–stick
- CPK (space filling)
- $C_\alpha$ trace (smooth)
- $C_\alpha$ trace (B-factor)
- ball-and-stick (B-factor)
- ribbons
The PDB-format

The PDB-file is a plain text file which stores the information of the model. Every line is at most 80 characters wide (this dates back to when computers were fed with punch-cards).

```
HEADER LIGASE 28-APR-99 1CLI
TITLE X-RAY CRYSTAL STRUCTURE OF AMINOIMIDAZOLE RIBONUCLEOTIDE
AUTHOR C.LI,T.J.KAPPOCK,J.STUBBE,T.M.WEAVER,S.E.EALICK
REMARK 2 RESOLUTION. 2.50 ANGSTROMS.

CRYST1 71.170 211.680 94.450 90.00 90.00 90.00 P 21 21 21 16

ATOM 1 N THR A 5 15.163 80.897 61.279 1.00 20.99 N
ATOM 2 CA THR A 5 15.093 82.326 61.723 1.00 22.09 C
ATOM 3 C THR A 5 16.450 83.017 61.598 1.00 21.68 C

The ATOM lines contain the coordinates and atom types. All other lines contain additional information (publication, resolution, refinement statistics, ...) which should be read when working with a PDB-file.
```
Protein Data Base (PDB)

Macromolecular Structures are stored at the Protein Database (www.pdb.org. Access to the PDB is free. For nucleic acids, there is also the Nucleic Acid Database (NDB, ndb-server.rutgers.edu). Since recently, everybody who submits a new structure to the PDB also has to submit the structure factors, i.e. the experimental data used to create the model. This helps preventing abuse and detecting erroneous structures.

Organic small molecules are stored in the Cambridge Crystal Structure Database (CCSD), inorganic small molecules are stored in the Inorganic Crystal Structure Database (ICSD), both of which are licensed products.
Model Building & Refinement
Refinement Cycle

- Data
- Crystallographer
- Model
- Calculate Map
- New Model
- Better w.r.t. map
- Better w.r.t. data and stereochemistry
- Refinement by program
- Build model/match model to map
The Refinement Cycle: Words of Caution

- Only the intensities $I(hkl)$ are experimental data.
- The initial map comes from the estimates from phasing.
- The improvement of the model means an improvement of the phases.
- One requires a model to produce a map and a map to improve the model.

The dependency model $\rightarrow$ map $\rightarrow$ model $\rightarrow$ map $\ldots$ is error prone. It is important to understand the risks in order to produce a good model.
The Electron Density Map

After refinement one usually looks at two types of electron density maps:

- **the “normal” map**: should embed the model
- **the difference map** shows errors: parts with too many atoms are negative (Coot: red), parts where atoms are missing are positive (Coot: green).

**Noise** shows up in both types of maps and can easily be mistaken as structural features → risk of overfitting. The correct interpretation of the maps requires chemical and biological knowledge.
Model Building

Model Building marks the improvement of the model to the electron density. Atoms or residues are added/removed, solvent atoms and ligands are placed, corrections are carried out.

Model Building aims at reducing the peaks of the difference map.

Model Building is done manually by the crystallographer.
Refinement

Refinement adjusts the model to the data $I(hkl)$ instead of the map. It changes to the model are usually small, but they involve the whole molecule at once. It also takes the B-factor into account, and at high resolution even the occupancy.

Refinement is carried out by computer programs.
Refinement Programs

The most common programs for macromolecular structure refinement include:

- Refmac5 (G. Murshudov et al.)
- phenix.refine (P. Adams et al.)
- shelxl (G. Sheldrick)
- CNS (A. Brunger et al.)
- TNT (D. Tronrud)
Model Building Programs

Programs for model building are graphical user interfaces. The most common ones include:

- Coot (P. Emsley et al.)
- O (A. Jones et al.)
- MIFit (D. McRee)
- Turbofrodo (A. Roussel et al.)
What is being refined

The molecules inside the unit cell are composed of atoms. In (macromolecular) X-ray crystallography, atoms are characterised by

- their type (N, O, C, Ca, ...).
- their position (x-, y-, z- coordinates).
- their B-factor.
- their occupancy.
Occupancy and B-factor

The electron density map obtained from X-ray data is the average of all unit cells in the crystals. Most atoms do not deviate too much from the average (otherwise there would be no crystal).

Two types of deviations can be described in crystallography by

B-factor and occupancy

When the deviations are too big and too arbitrary, there are no data and the atoms cannot be modelled.
B-Factor

The B-Factor of an atom describes its thermal motion. Even though data are usually collected at 100K, the atoms are not frozen, but move slightly. Also small domain movements can be captured by the B-factor.

At about 1.5Å resolution and better, every atom has six parameters which describe the anisotropic thermal displacement (ADP) of the atom in three directions independently.

From 1.5Å - 3.5Å resolution there are not enough data for such detailed description and the thermal motion is described by only one isotropic B-factor.

At worse than 3.5Å resolution this is even further reduced to one B-factor per residue and eventually one parameter for the whole molecule.
Isotropic vs. Anisotropic B-Factor

Isotropic B-Factors

Anisotropic B-Factors
B-Factor Indicates Domain Mobility

Colouring of a protein by B-factor per residue

- **blue**: low B-factor
- **green**: medium B-factor
- **red**: high B-factor

The *core* is rather stable (low B-factors), the borders/loop-regions are flexible (high B-factors)
An Example for Occupancy

High-resolution map (1.3 Å)
In most parts the positions of the backbone and side-chains are visible.
At the centre the density looks a little “bloppy”.
An Example for Occupancy

High-resolution map (1.3 Å)
In most parts the positions of the backbone and side-chains are visible.
At the centre the density looks a little "blobby".

The main-chain splits into two parts: 40% of all unit cells contain one conformation, 60% the other one.
Occupyancy vs. B-factor

The occupancy describes **discrete conformations** of side chains or even whole parts of a molecule.

The B-factor describes **small** movements of atoms.

Any other larger flexibility is not displayed by crystallography.
Model Building
Being Lazy: Automated Model Building

There are programs for automated model building, e.g.:

Arp/Warp (A. Perrakis, V. Lamzin), Buccaneer (K. Cowtan), or Resolve (T. Terwilliger), Shelxe (G. Sheldrick, Poly-Ala only)

For proteins these work well at 2.5Å and better resolution and with better than 2.0 Å resolution, an automatically built model can be a nearly fully refined structure.

But only nearly ...
Manual Model Building

Computer programs do not know about biology, certainly not of a specific molecule/structure. Human interaction is therefore required to pay attention to:

- correct placement of solvent (water) molecules
- multiple conformations
- presence and identification of ligands and/or metal ions
- special interaction for complexes
- exceptions from standard values used in refinement

And finally you also want to interpret your structure.
Getting Started

At medium or better resolution (< 2.5 Å) the first model is most likely created by a program for automated model building, which can be 80% complete or better. After successful molecular replacement, one also at least has most of the backbone and only needs to make “minor” corrections.

At low resolution, such comfort may not be available and one has to create a model “from scratch”.

The best thing to start with is to find secondary structure elements, i.e. α-helices and β-sheets. Especially α-helices are visible even at low resolution.

For nucleic acids, the bases, base-stacking, as well as the phosphate backbone are the features to look out for in the electron density map.
α-Helices: the Christmas Tree

2.4Å map after SeMet-MAD

The side chains, in particular the $C_\beta$-atoms, of an α-helix tend to point backwards to the N-terminus of the sequence. This is a good way to get the direction right of the helix.

Tim Grüne
$\beta$-sheets

2.4Å-map after SeMet-MAD

$\beta$-sheets are also striking, but their direction is not as obvious and they can easily be placed the wrong way round.

Tim Grüne
2.4Å-map after SeMet-MAD

β-sheets are also striking, but their direction is not as obvious and they can easily be placed the wrong way round.
Sequence Docking

Model building begins with the placement of the $C_{\alpha}$ atoms (they sit where the side chains protrude off the main chain), which are then turned into a poly-alanine model (this fixes the direction of the chain).

In order to place the side chains correctly, it is good to start with bulky, large side chains like Trp, Phe, Tyr.

Marker atoms from the phasing experiment are also good anchors, especially the Se-atoms after SeMet-MAD.
Phase Improvement

The electron density map is calculated from phases that are calculated from the model — the model is the storage container for the phases. The more complete the model, the better the map, which in turn facilitates model building.
Water, Ions, and Ligands

When adding waters, or ions, or ligands to a structure, they must **chemically make sense**!

Active site of thermolysin:
One could fill this patch of density (green) with water molecules, until no (green) density is left. However, this would be overfitting – our chemical understanding tells us that these are **not** water molecules, but something else.

Water molecules are fairly unrestrained (no bonds) - they can mimic anything!

Better leave this density in the structure undefined (but point out in the publication that there is unexplained, probably disordered density).
Reinement
What is being refined

Refinement tries to minimise the difference between the measured data $I_{meas}(hkl)$ and the data $I_{calc}(hkl)$ calculated from the PDB-file.

The $I_{calc}(hkl)$ are calculated based on the atoms in the unit cell, i.e., they depend on

- atom coordinates $(x, y, z)$.
- atom B-factors and occupancy.

These are the parameters that have to be properly determined (by the refinement program) to create a good model.
Stereochemistry

Refinement take stereochemistry into account:

- atoms must not clash (not come closer than their van-der-Waals-radii)
- bond distances and bond angles should be close to expected values.

The average bond distances and angles were determined and published by R.A. Engh and R. Huber in 1991 from (accurate) small molecule data.

Examples: $(C_\alpha - C) = 1.525\text{Å} \pm 0.020\text{Å}$, $(C_\alpha - N) = 1.329\text{Å} \pm 0.014\text{Å}$. 
Data to Parameter Ratio

During model building and refinement, the parameters are modified in order to better fit the data \( I_{\text{calc}}(hkl) \).

The more data points, the more reliably the parameters can be fitted to the data.

High-resolution crystal structures have more reflections (data) than low-resolution crystal structures. High-resolution structures have a better data to parameter ratio.

High-resolution structures are therefore more reliable than low-resolution structures.
### Data to Parameter Ratio vs. Resolution

<table>
<thead>
<tr>
<th>Resolution[Å]</th>
<th>refined parameters&lt;sup&gt;a&lt;/sup&gt;</th>
<th>data/parameters ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.0</td>
<td>x, y, z</td>
<td>0.9:1</td>
</tr>
<tr>
<td>2.3</td>
<td>x, y, z; B</td>
<td>1.5:1</td>
</tr>
<tr>
<td>1.8</td>
<td>x, y, z; B</td>
<td>3.1:1</td>
</tr>
<tr>
<td>1.5</td>
<td>x, y, z; B</td>
<td>5.4:1</td>
</tr>
<tr>
<td>1.5</td>
<td>x, y, z; U&lt;sub&gt;11&lt;/sub&gt; U&lt;sub&gt;12&lt;/sub&gt; U&lt;sub&gt;13&lt;/sub&gt; U&lt;sub&gt;23&lt;/sub&gt; U&lt;sub&gt;22&lt;/sub&gt; U&lt;sub&gt;33&lt;/sub&gt;</td>
<td>2.4:1</td>
</tr>
<tr>
<td>1.1</td>
<td>x, y, z; U&lt;sub&gt;11&lt;/sub&gt; U&lt;sub&gt;12&lt;/sub&gt; U&lt;sub&gt;13&lt;/sub&gt; U&lt;sub&gt;23&lt;/sub&gt; U&lt;sub&gt;22&lt;/sub&gt; U&lt;sub&gt;33&lt;/sub&gt;</td>
<td>6.1:1</td>
</tr>
<tr>
<td>0.8</td>
<td>x, y, z; U&lt;sub&gt;11&lt;/sub&gt; U&lt;sub&gt;12&lt;/sub&gt; U&lt;sub&gt;13&lt;/sub&gt; U&lt;sub&gt;23&lt;/sub&gt; U&lt;sub&gt;22&lt;/sub&gt; U&lt;sub&gt;33&lt;/sub&gt;</td>
<td>16:1</td>
</tr>
</tbody>
</table>

<sup>a</sup>x, y, z: coordinates; B: isotropic B-value; U<sub>ij</sub>: anisotropic B-values

Effectively at worse than 1.8Å, there would not be enough data points to create a reliable model. The data to parameter ratio can be improved by additional — (bio—) chemical etc. — information.
Restraints and Constraints

Constraints and restraints are introduced into refinement in order to improve the data to parameter ratio.

**Constraints** “reduce the number of parameters”. They are “must have” or “must be” expressions (mathematically: equalities)

- e.g.: “temperature factor must be isotropic instead of anisotropic”:
  (3+1=4) parameters per atom instead of (3+6=9) parameters per atom

**Restraints** “increase the number of data”. “Should be” or “should be approximately” expressions (mathematically: inequalities)

- e.g. angle \((N, C, O) \approx 122^\circ\).

The Engh-Huber parameters are restraints.

Only small molecules at very high resolution (< 0.8 Å) can be freely refined, i.e. without using constraints and restraints.
Model Bias and Overfitting

The refinement programs minimise the difference between $I_{\text{meas}}(hkl)$ and $I_{\text{calc}}(hkl)$.

... including the difference density for this beautifully displayed Phe which is missing in the model.
Model Bias and Overfitting

At high resolution such strong difference density would not disappear, it is too strongly anchored in the data $I(hkl)$, no matter what the phases say.

- ...at low or medium resolution ...

But ...

- ...at the beginning of model building ...
- ...after molecular replacement ...

the phases are still poor, and when the data are weak (low or medium resolution), it may happen that the refinement program levels out such features from the difference map, or enhance errors in the model.

Therefore: Always do **as much model building** as possible before running the refinement program.
Summary and Outlook

Model Building and Refinement are a bit of a *vicious circle*: Because of the *lack of reliable experimental phases*, the model is required for creating the model.

Especially at weak resolution it is easy to introduce or overlook errors in the model that one cannot remove anymore.

Therefore it is important to understand what refinement does and to *validate* the structure one has build.