Refinement with REFMAC5

DLS-CCP4 Data Collection & Structure Solution Workshop

8th December 2021

Rob Nicholls
nicholls@mrc-lmb.cam.ac.uk

Lucrezia Catapano
lucrezia@mrc-lmb.cam.ac.uk

MRC Laboratory of Molecular Biology

REFMAC5
Macromolecular structure refinement

COOT
Visualization and model building

A few key tools for refinement with CCP4

ProSMART
Restraint generation and comparative structural analysis

AceDRG
Ligand restraint dictionary and conformer generation

LibG
Nucleic acid restraint generation

LORESTR
Automated pipeline for low-resolution refinement

MRC-LMB, Cambridge:

Purpose of Refinement

Crystallographic refinement has two purposes:

- **<u>Fit atomic model into observed X-ray crystallographic data</u>**

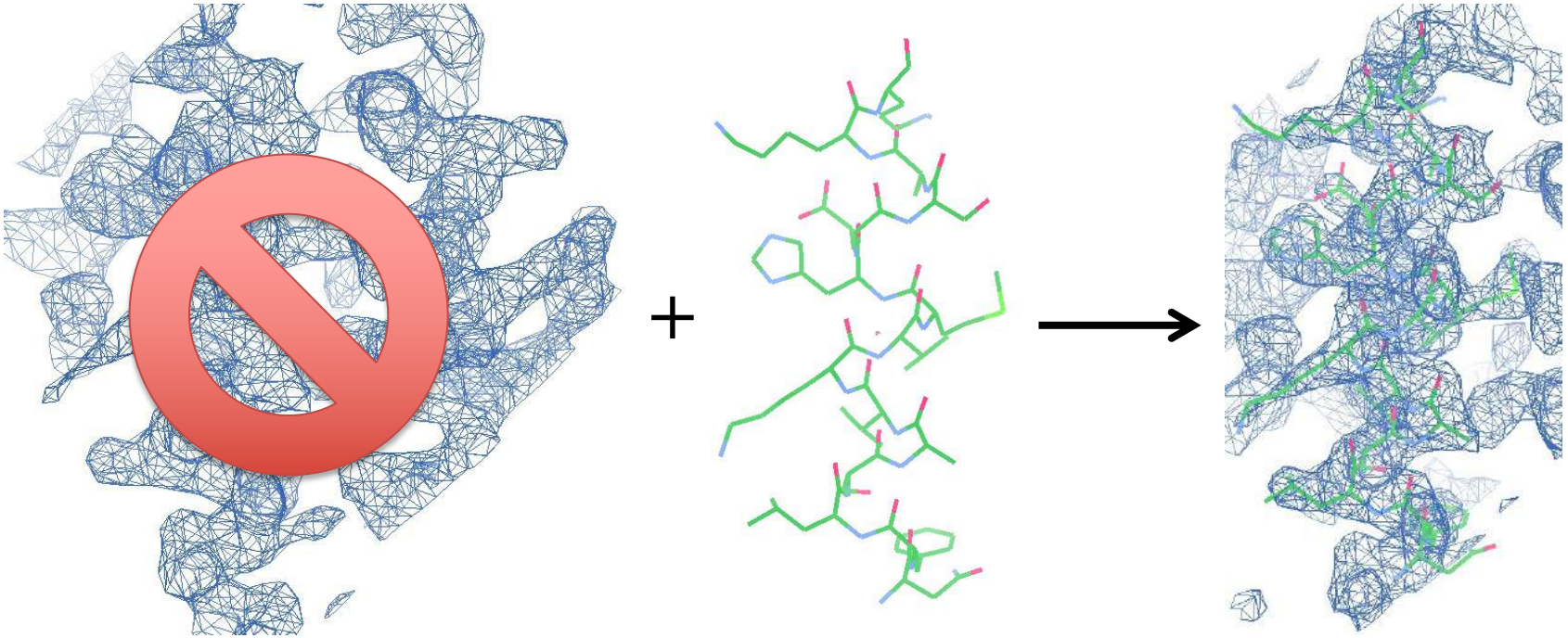
  Model should agree with the observed data

  Model must be chemically and structurally sensible

- **<u>Calculate best possible electron density map</u>**

  Allowing the atom model to be visualised, criticised and analysed

Purpose of Refinement



Data

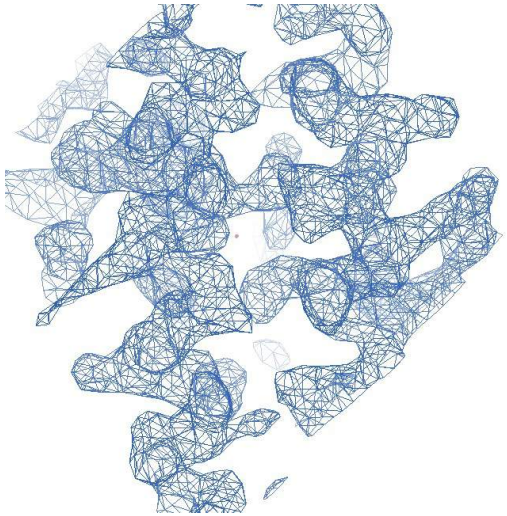
Atomic model

Fit and refine

Crystallographic Data

Different types of data:

- Amplitudes of structure factors from single crystals:
  Observed amplitudes: $|F_{obs}|$
  Estimated uncertainties: σ_{obs}



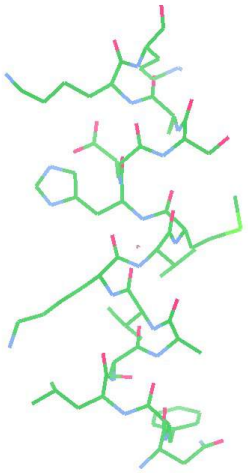
- Intensities/amplitudes from “twinned” crystals
- SAD – amplitudes available for $|F_+|$ and $|F_-|$
- Amplitudes available from multiple crystal forms

Model Refinement

We have observed amplitudes: $|F_{obs}|$

But we don't have phases: φ

Suppose we have a starting model:



FFT

F_{calc} $|F_{calc}|, \varphi_{calc}$

calculated model structure factor

Idea:

Iteratively improve the model, optimising the agreement between $|F_{obs}|$ and $|F_{calc}|$

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 actually refining against a density map

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

How to assess success, model quality?

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

REFMAC outputs coefficients for 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}

Warning:

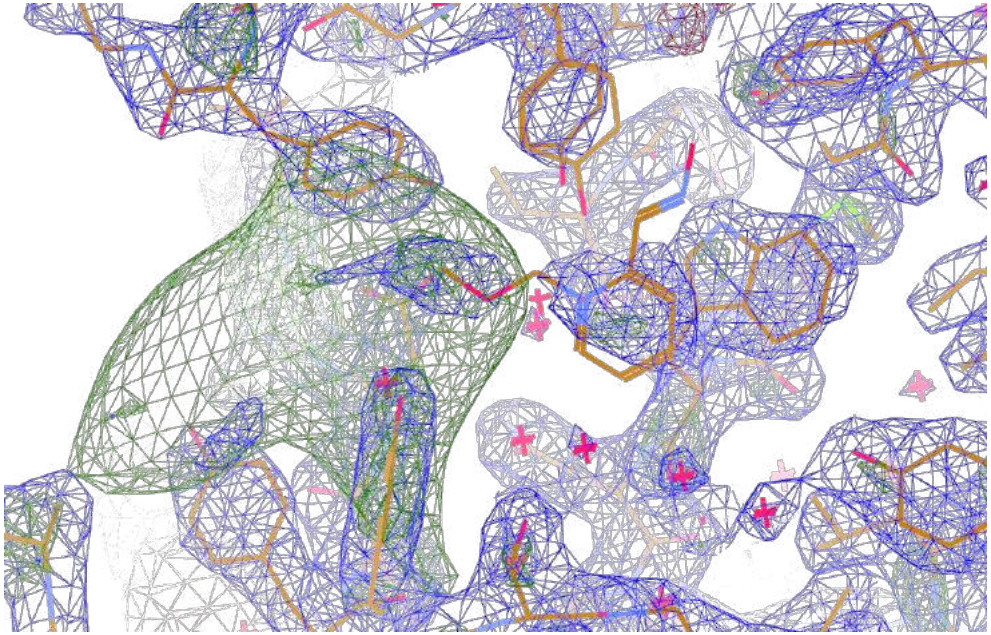
Model updated

↓

φ_{calc} updated

↓

Density changes



Note – contrast with Coot real space refinement, and REFMAC5 cryo-EM refinement

Model Refinement

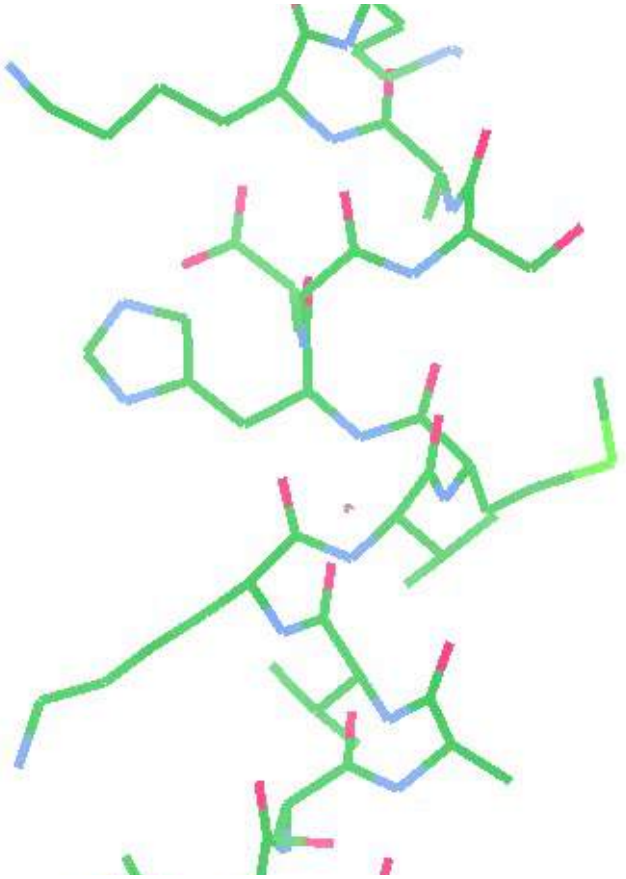
We now know:

- What sort of data we have
- How to assess model quality
- How to get phase estimates from the current model
- How to calculate electron density maps

So what is the model, and how do we refine it?

Model Parameterisation

Standard refinable parameters



Atomic model:

- Position – (x,y,z) coordinates
- ADP (B or U factors)
- (Occupancies)

Overall parameters (scaling)

- Overall B-factor (and anisotropic U)
- Solvent treatment

Note – different to data anisotropy (which is dealt with during data processing)

Model Parameterisation

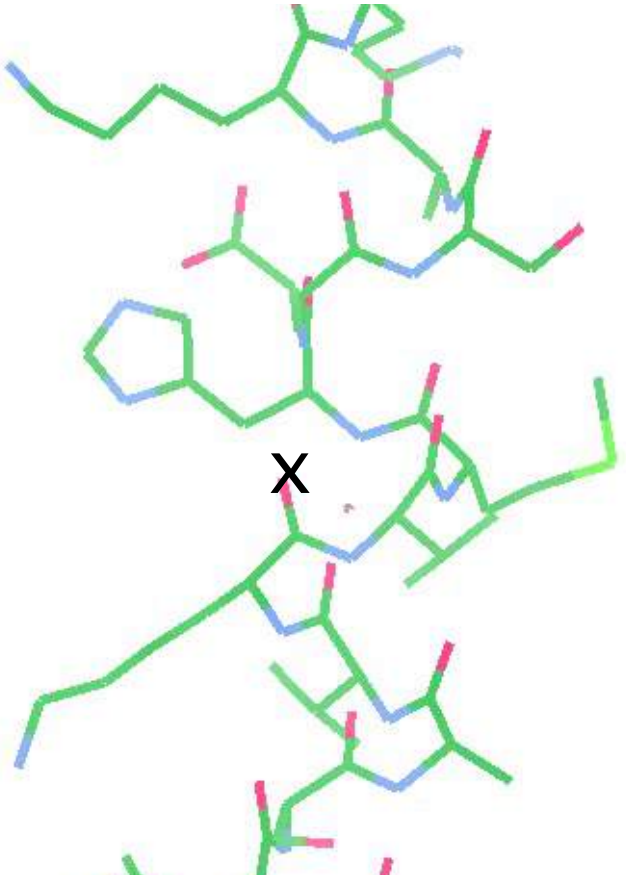
Standard refinable parameters

Atomic model:

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

Overall parameters (scaling)

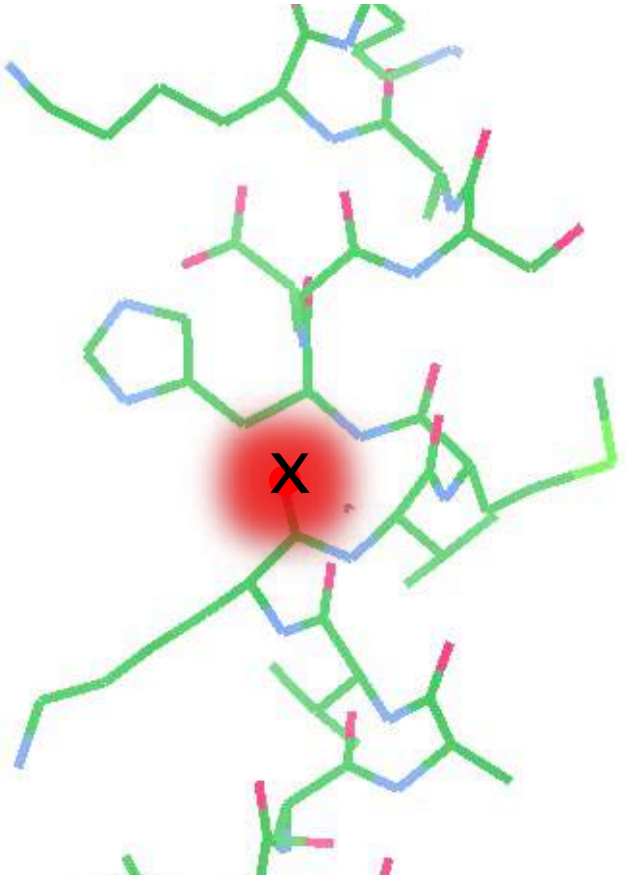
- Overall B-factor (and anisotropic U)
- Solvent treatment



Note – different to data anisotropy (which is dealt with during data processing)

Model Parameterisation

Standard refinable parameters



Atomic model:

- Position – (x,y,z) coordinates
- **ADP (B or U factors)**
- (Occupancies)

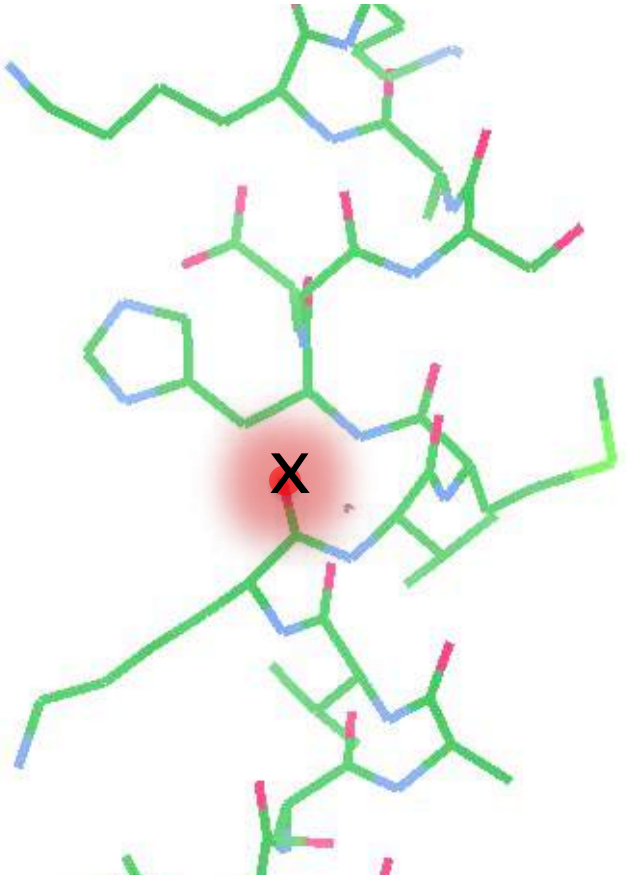
Overall parameters (scaling)

- Overall B-factor (and anisotropic U)
- Solvent treatment

Note – different to data anisotropy (which is dealt with during data processing)

Model Parameterisation

Standard refinable parameters



Atomic model:

- Position – (x,y,z) coordinates
- ADP (B or U factors)
- **Occupancies**

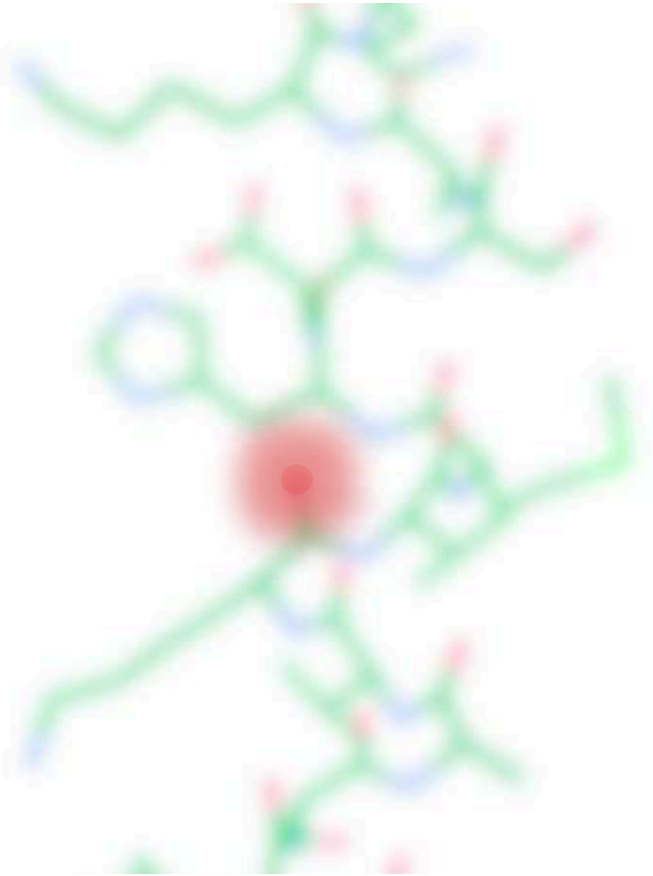
Overall parameters (scaling)

- Overall B-factor (and anisotropic U)
- Solvent treatment

Note – different to data anisotropy (which is dealt with during data processing)

Model Parameterisation

Standard refinable parameters



Atomic model:

- Position – (x,y,z) coordinates
- ADP (B or U factors)
- (Occupancies)

<u>Overall parameters (scaling)</u>

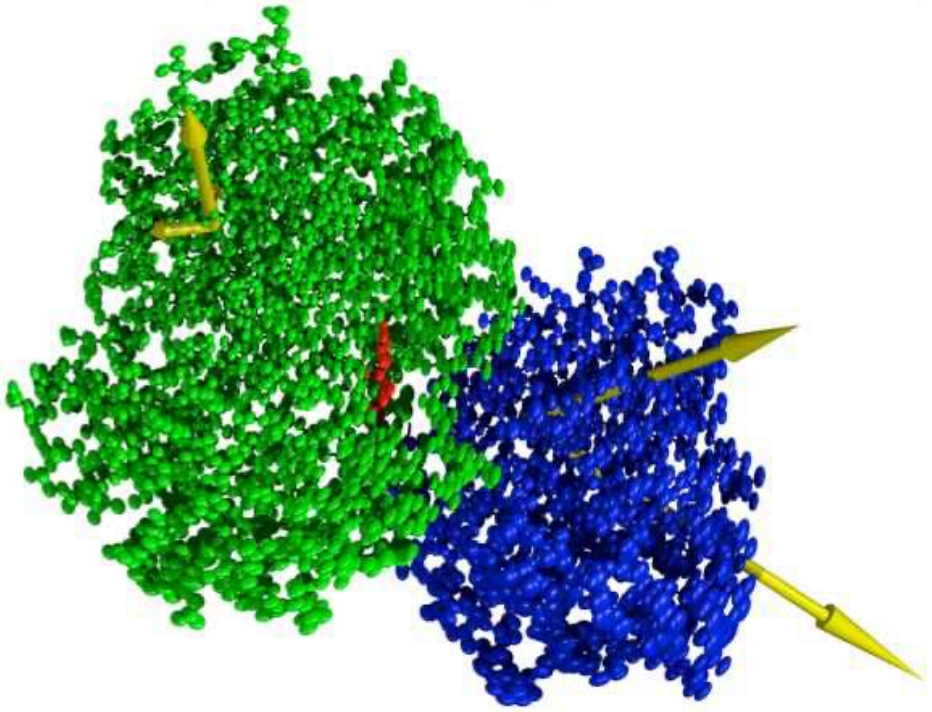
- Overall B-factor (and anisotropic U)
- Solvent treatment

Note – different to data anisotropy (which is dealt with during data processing)

TLS Groups

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

Suitable for medium/low resolutions, when full anisotropy is impossible



Per group (20 parameters):

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

Refined as a separate step

- Auto: one group per chain
- Define groups manually
- TLSMD webserver: http://skuld.bmsc.washington.edu/~tlsmd/

Model Refinement

We now have:

- Data – to refine our model against
- Parameters to refine – describing the model

How do we refine the model?

Model Refinement

REFMAC5 uses a Maximum Likelihood approach

Crystallographic target functions have two components:

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

likelihood of the data

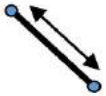
probability of the model

We also need prior knowledge (restraints)

These help ensure chemical and structural integrity

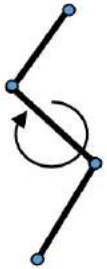
Restraints

Standard restraints (used by default) include:

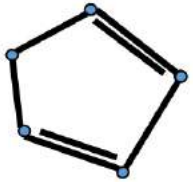


Bond lengths

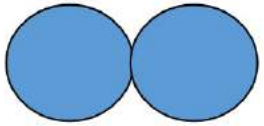
Angles



Torsion-angles



Planes



VdW repulsions

B-values

These help to ensure that the model is chemically sensible

Note – we generally deal with restraints, not constraints

Restraints

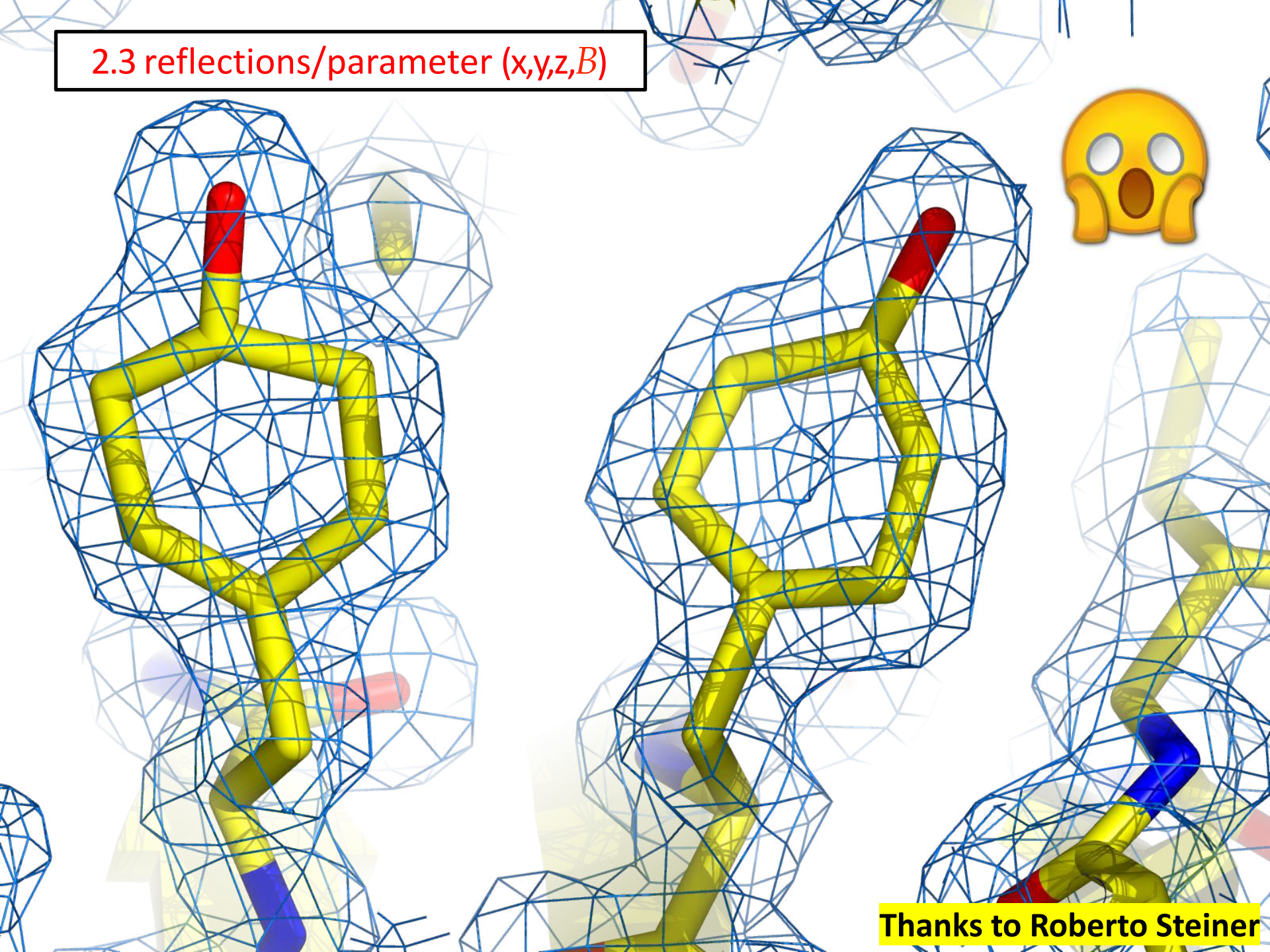
Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.

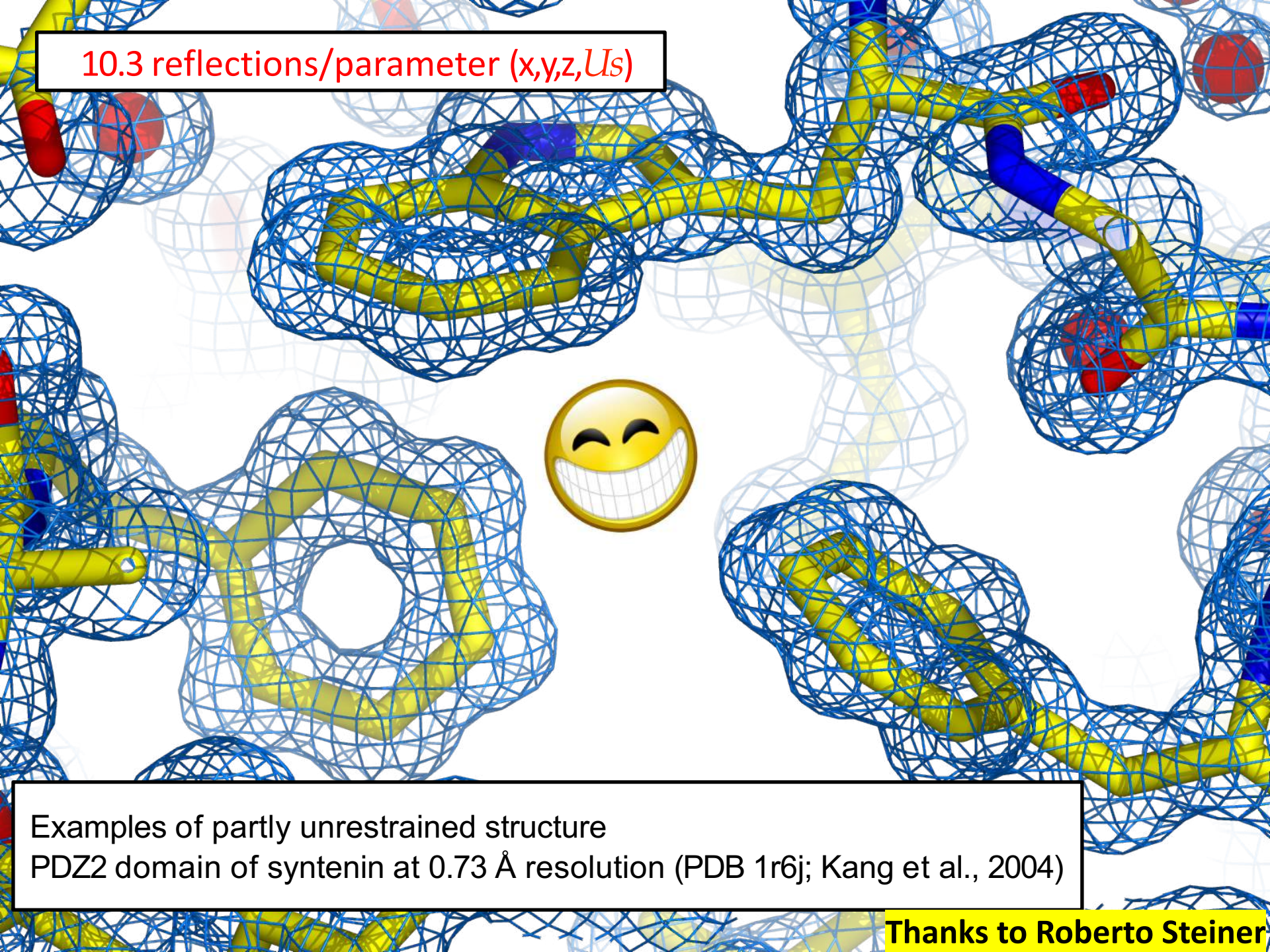
1.8 Å / 164 aa / 1540 non-H atoms / 14217 reflections

≈2.3 reflections/parameter (x,y,z,B)

2.3 reflections/parameter (x,y,z,B)



Thanks to Roberto Steiner



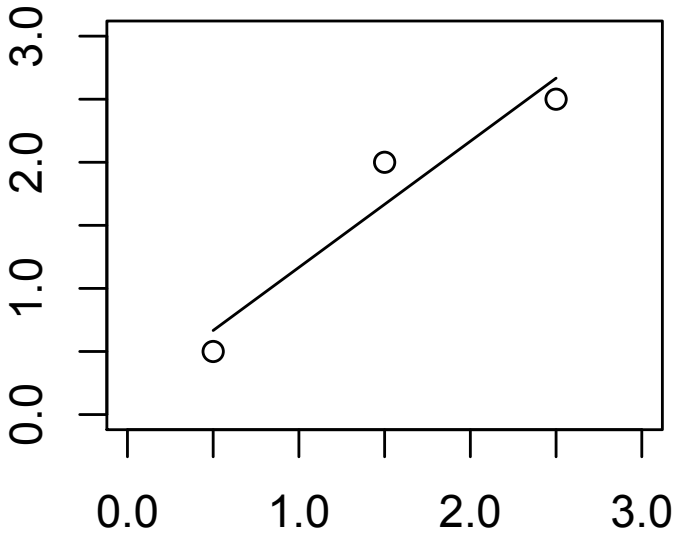
Examples of partly unrestrained structure
PDZ2 domain of syntenin at 0.73 Å resolution (PDB 1r6j; Kang et al., 2004)

Thanks to Roberto Steiner

Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.

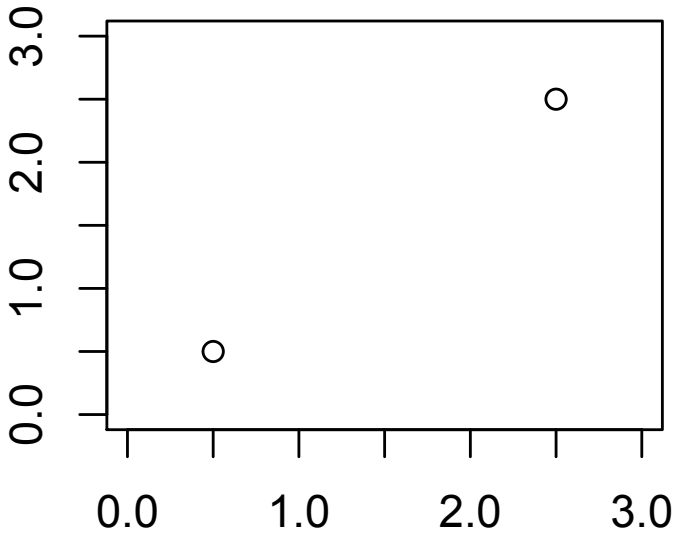


Example: Fitting a line $y = a + bx$

Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.



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

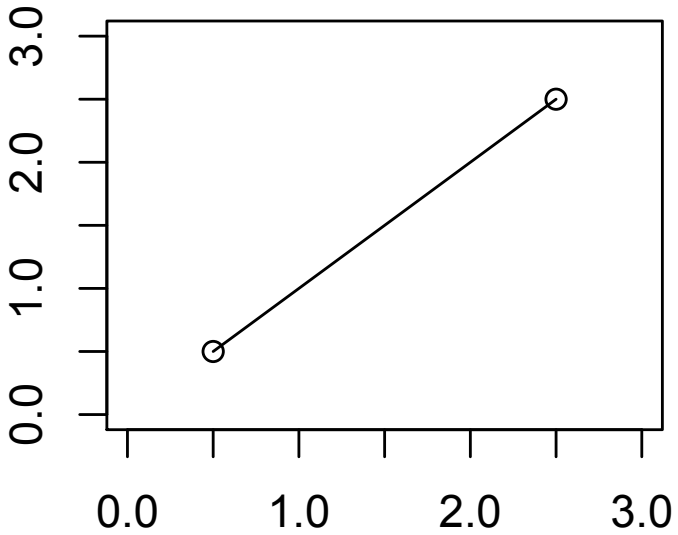
Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.

Can fit a line

Line is unreliable



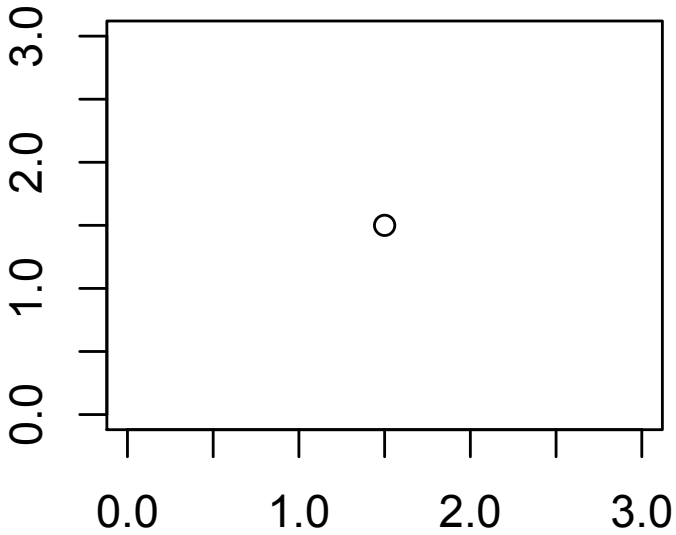
Overfitting
Model Bias

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

Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.



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

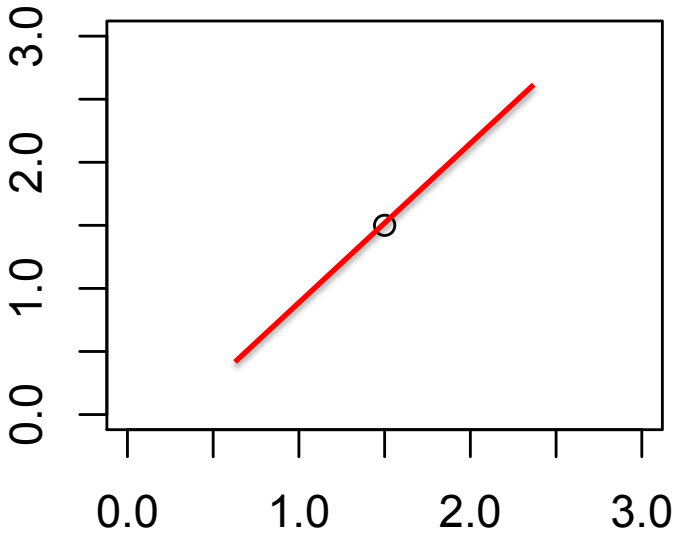
Restraints

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

Med-low resolution	:	Isotropic ADP	– 4 params per atom
High resolution	:	Anisotropic ADP	– 9 params per atom

- TLS – 20 additional parameters per group
- Rigid body refinement – 9 parameters per body

Restraints

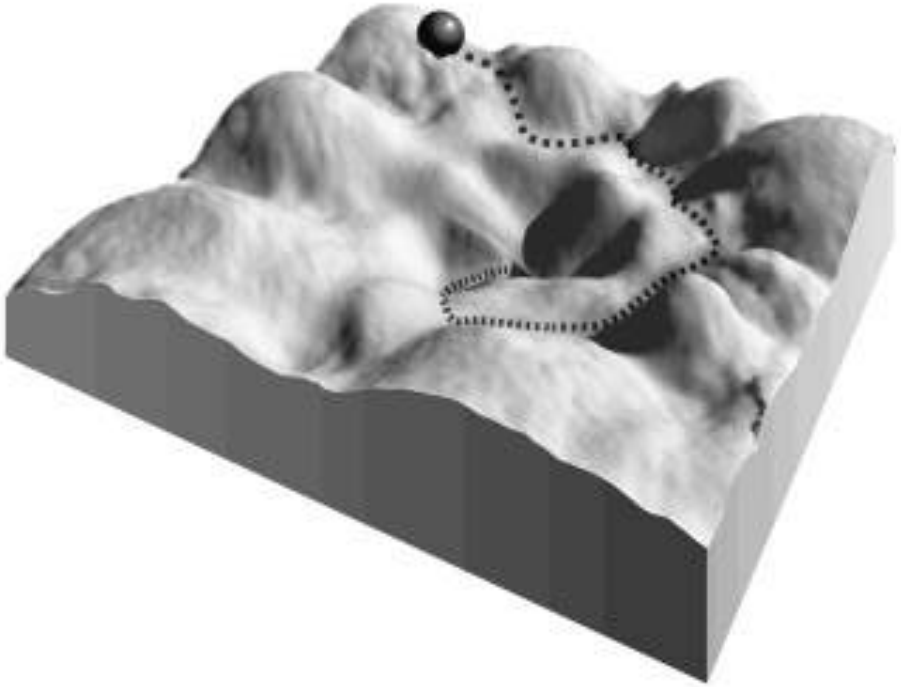
How to improve the observation:parameter ratio.

2. Increase number of restraints

Particularly useful at low-resolution:

- Reflection intensities often noisy
- Limited data – poor observation:parameter ratio
- Refinement becomes unstable
- Overfitting – R-factors diverge

‘Refinement problem’

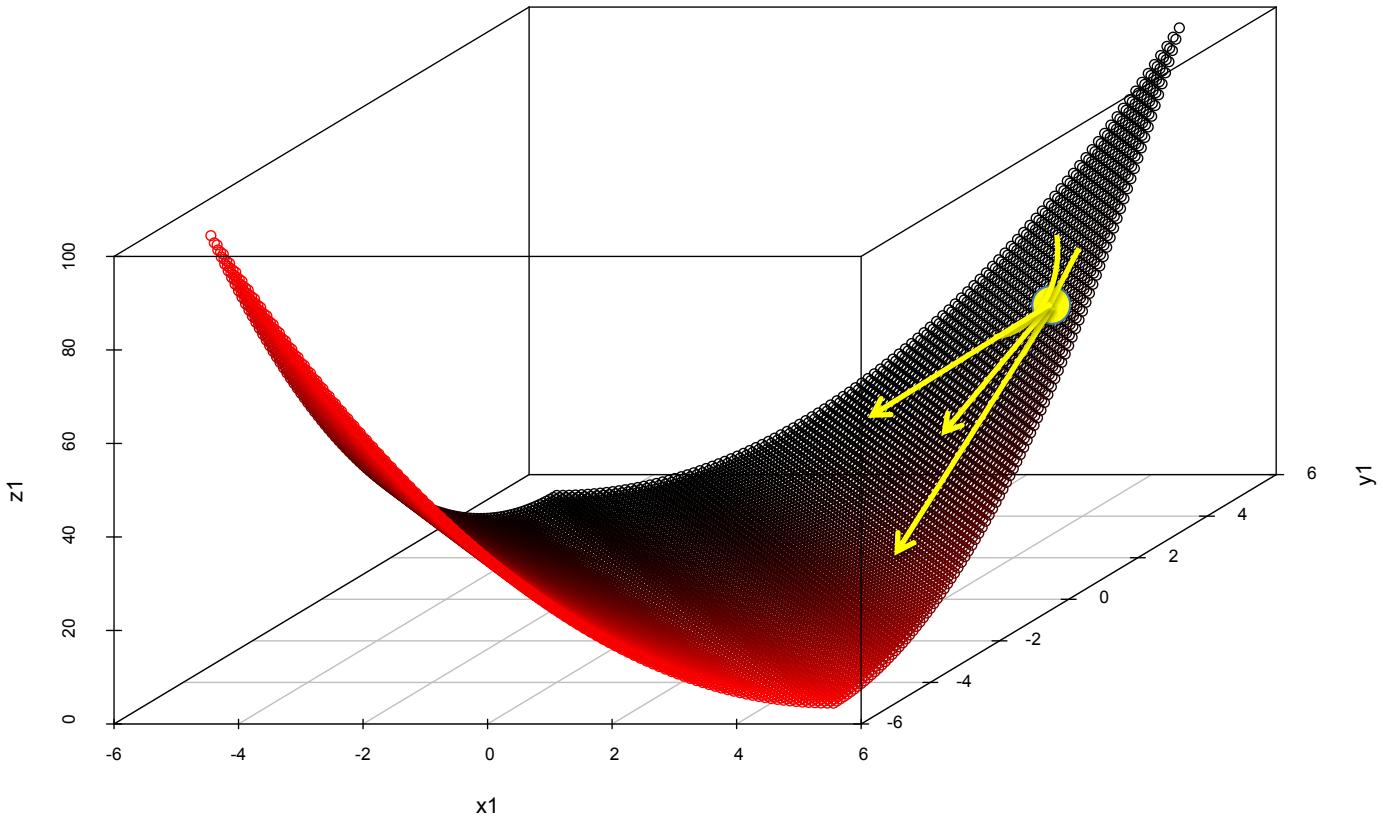


- 2D representation of refinement
- The ball represent the current set of parameters
- We want to minimize a function i.e. find the best set of parameters
- The ball could get stuck or go in the wrong way

Regularisation

Example:

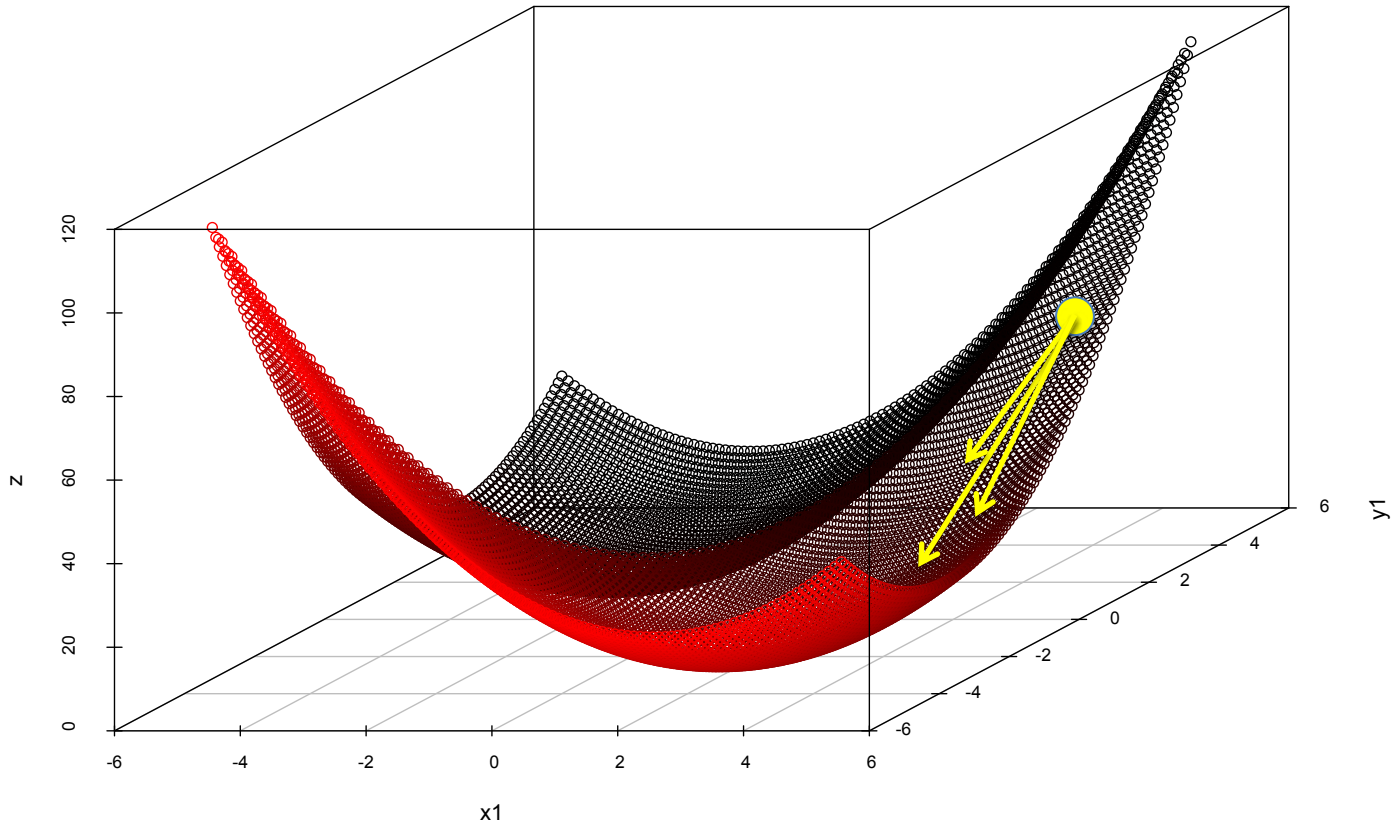
$$z = (x + y)^2$$



Regularisation

Example:

$$z = (x + y)^2 + (|x - y| - 4)^2$$



Regularise using prior information: $|x - y| = 4$

Regularisation

Use of available knowledge (prior information):

High-low resolution:

- Geometry restraints (chemical information)

Medium-low resolution:

- Local NCS restraints
- B-value restraints
- Jelly body restraints

Low resolution (and medium-low resolution model building):

- External restraints

Regularisation

Use of available knowledge (prior information):

High-low resolution:

- **Geometry restraints (chemical information)**

Medium-low resolution:

- **Local NCS restraints**
- **B-value restraints**
- Jelly body restraints

Low resolution (and medium-low resolution model building):

- **External restraints**

Regularisers with a target value

Regularisation

Use of available knowledge (prior information):

High-low resolution:

- Geometry restraints (chemical information)

Medium-low resolution:

- Local NCS restraints
- B-value restraints
- **Jelly body restraints**

Low resolution (and medium-low resolution model building):

- External restraints

Regularisers without an external target value

Jelly Body Restraints

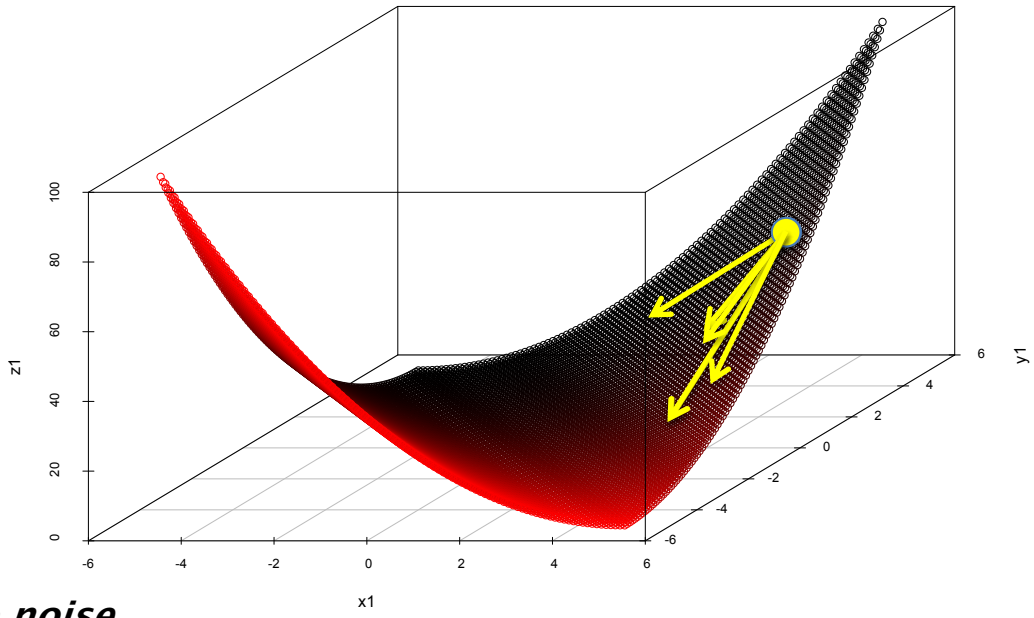
Regularisers without a target:

$$f = \sum_{\text{close atom pairs}} \frac{1}{\sigma^2}\left(d - d_{\text{current}}\right)^2$$

d : interatomic distance

d_{current} : current interatomic distance

σ : restraint standard deviation



Model should be less prone to fitting into noise

Should only work if parameters are near the minima (model is good)

Typical: σ = 0.01–0.02

Distance threshold: 4.2Å

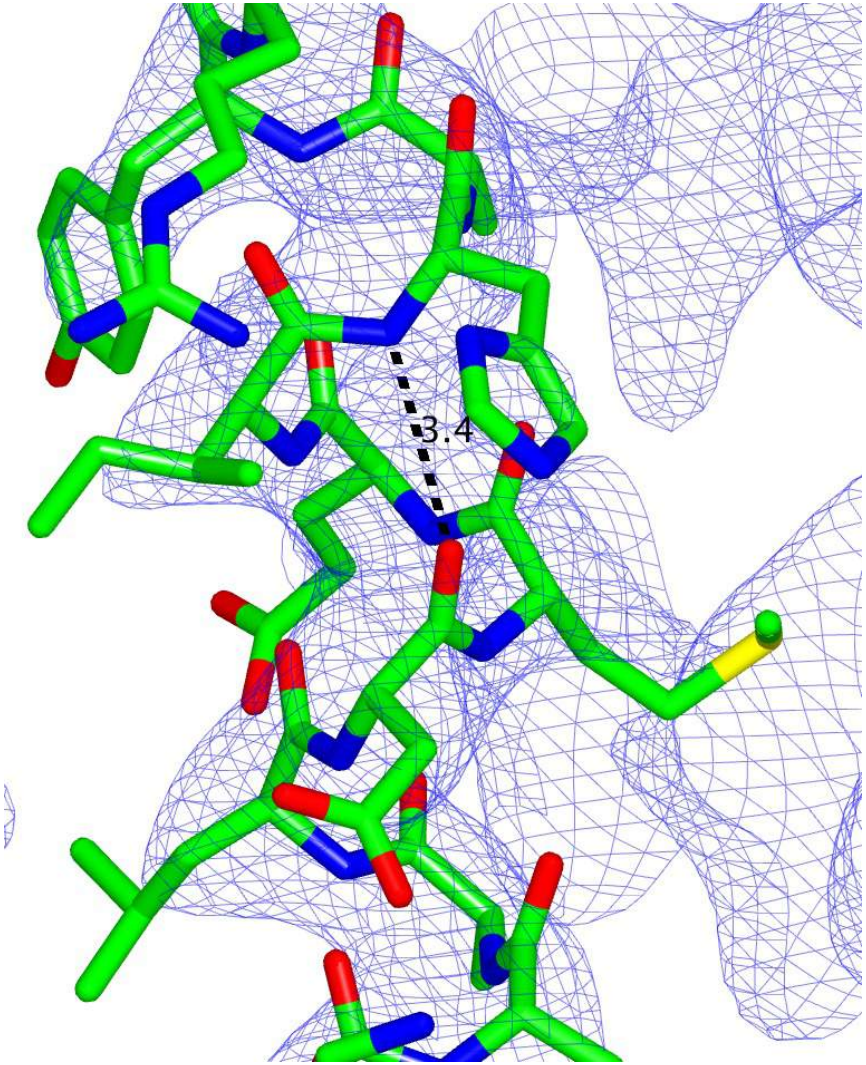
ProSMART

Injection of prior knowledge to aid new structure determination

- **External Restraints from homologous structures**
  - Protein or nucleic acid chains
- **Hydrogen bond restraints**
  - Protein backbone
- **Generic self-restraints**
  - Everything – protein, nucleic acid, ligand, water
- **Structure analysis**
  - Alignment & comparison - helps analyse differences between models

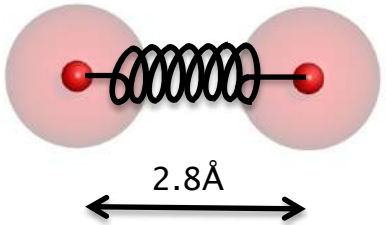
Independent of global conformation

ProSMART External Restraints



3g4w – 3.7Å

Prior information:

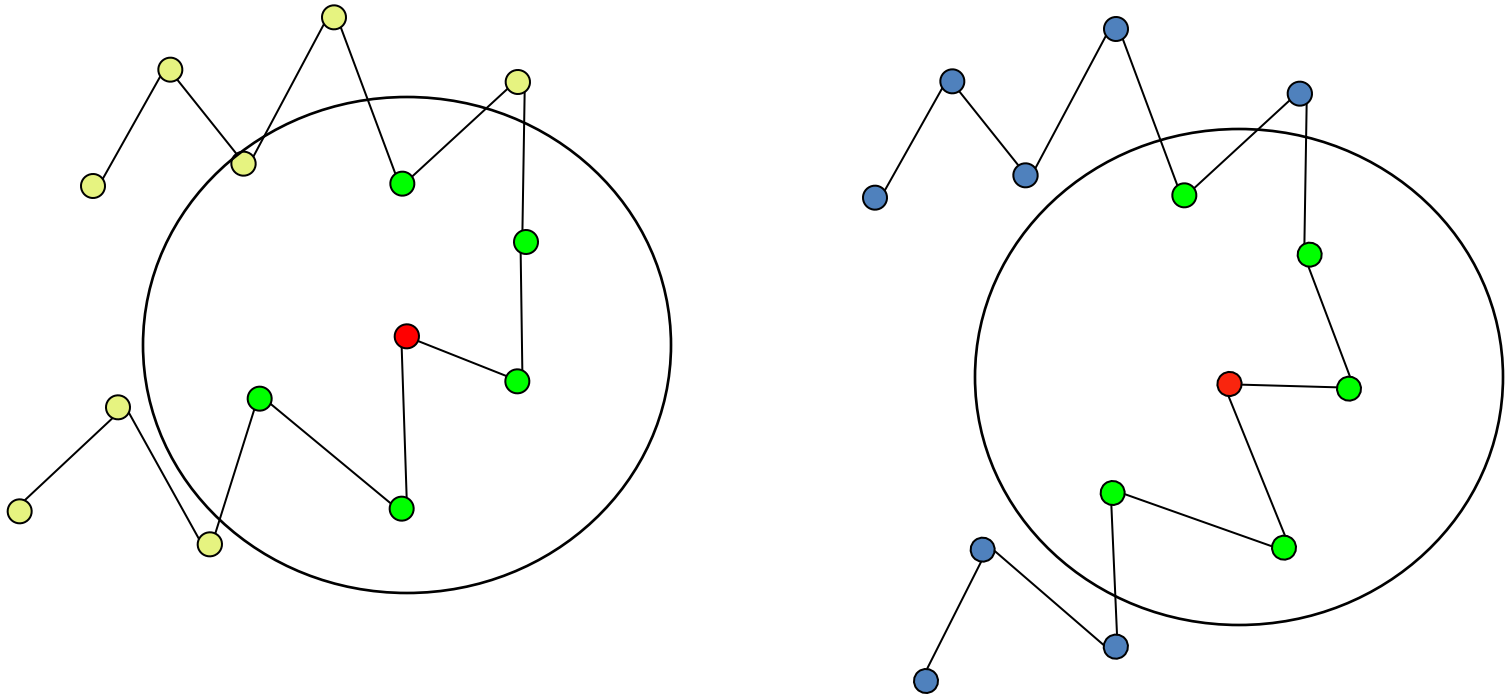


Stabilises structural features

External Restraint Generation

structure to be refined

known similar structure (prior)

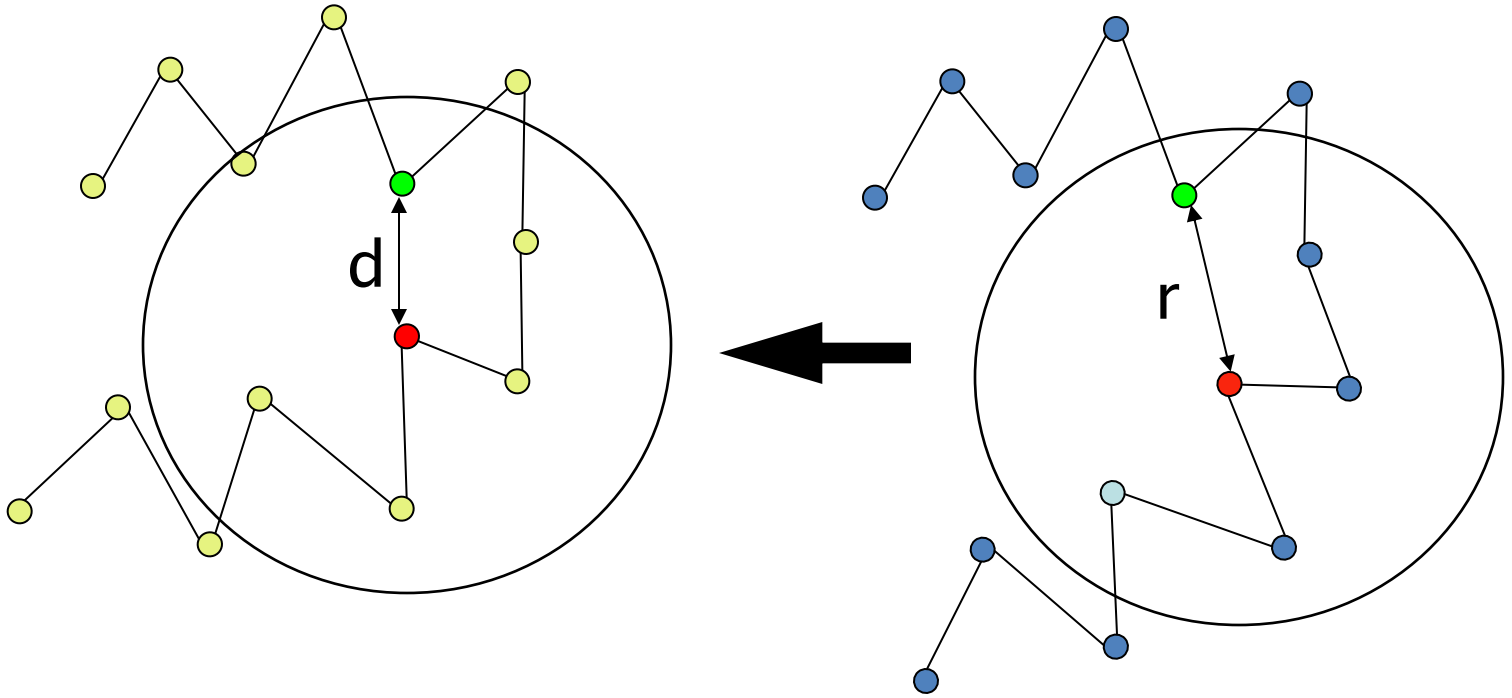


(abstract representation of an atomic model; circles = atoms)

External Restraint Generation

structure to be refined

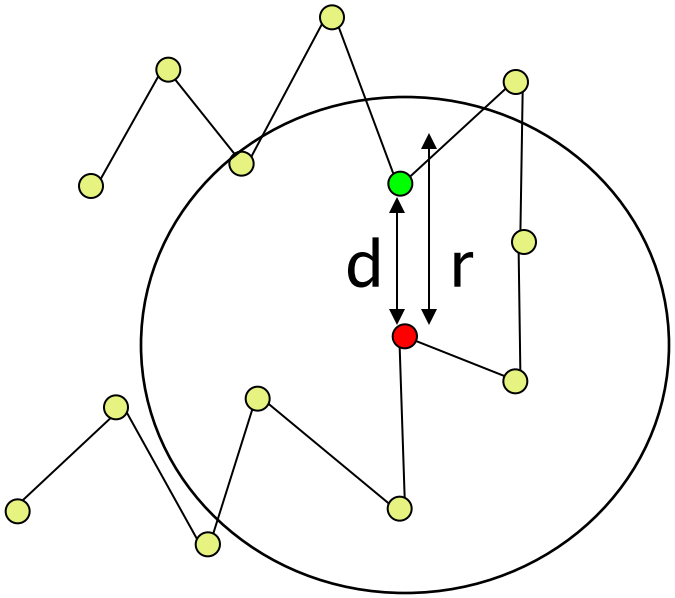
known similar structure (prior)



(abstract representation of an atomic model; circles = atoms)

External Restraint Generation

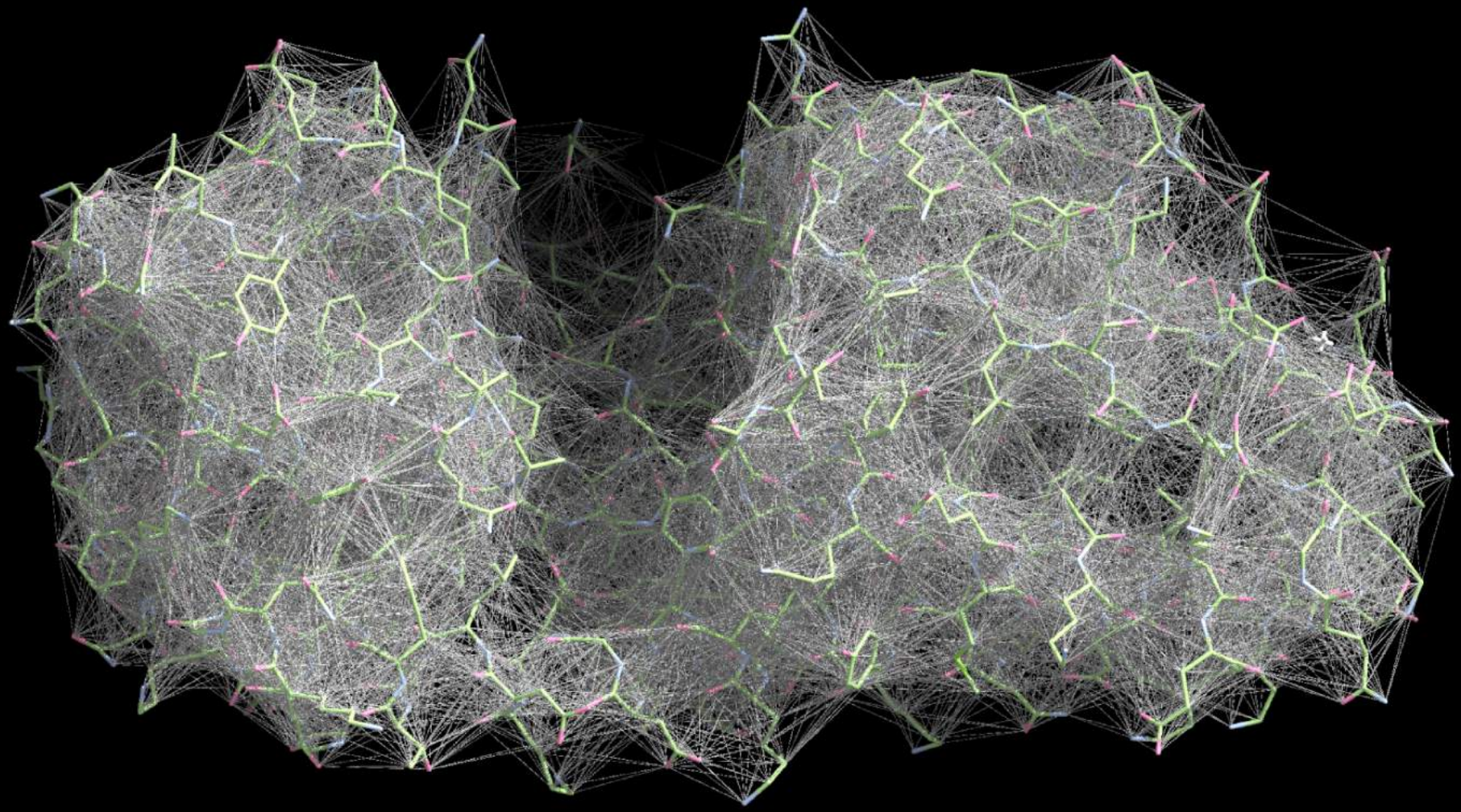
structure to be refined



$d \sim N(r,\sigma^2)$

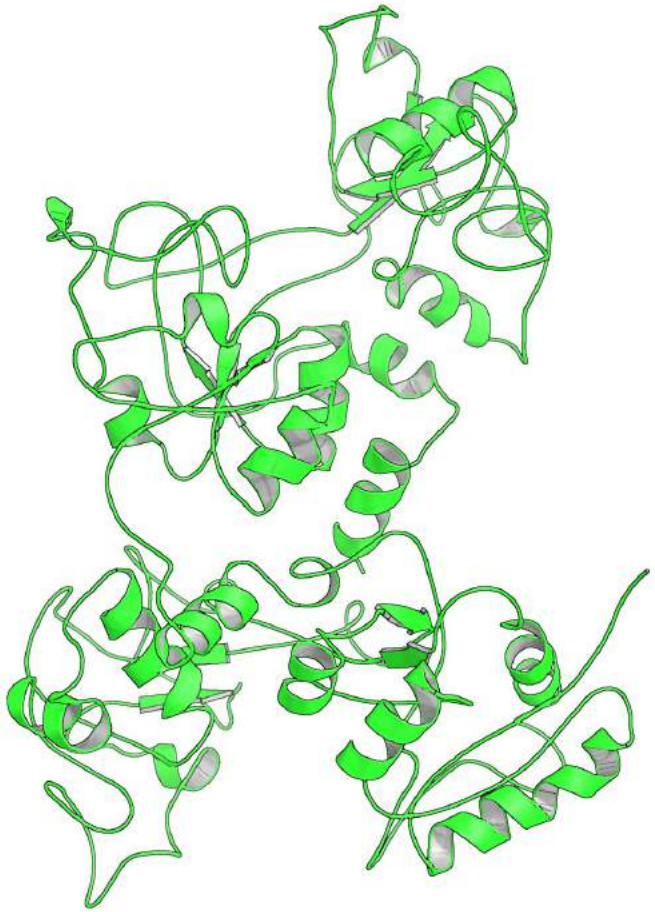
(abstract representation of an atomic model; circles = atoms)

ProSMART Restraints



Motivational Example

Ovotransferrin



1ryx – 3.5Å

Low-resolution refinement:

Weak signal
Noisy data

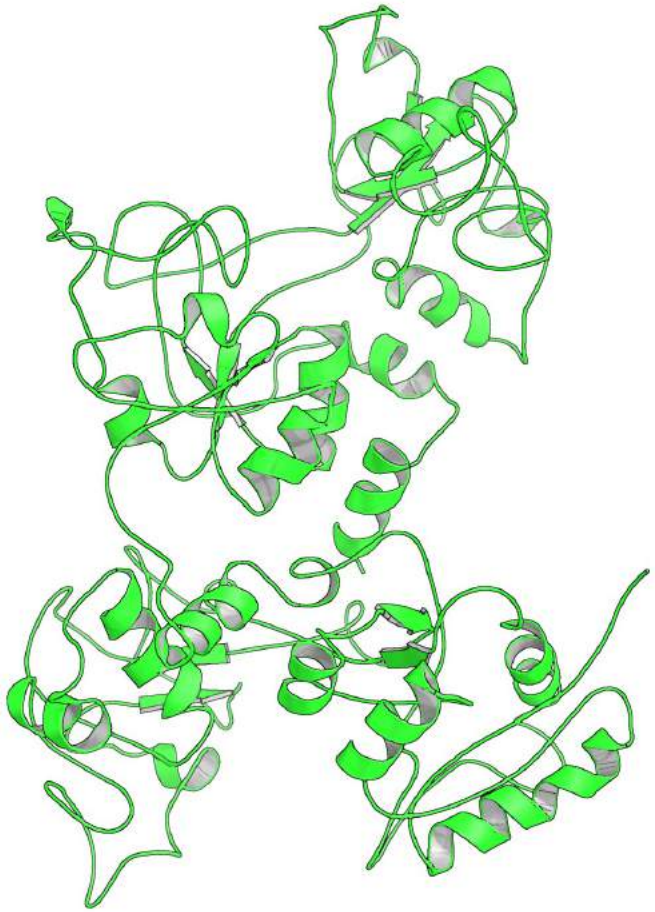
↓

Unstable refinement

Result:
Poor quality model

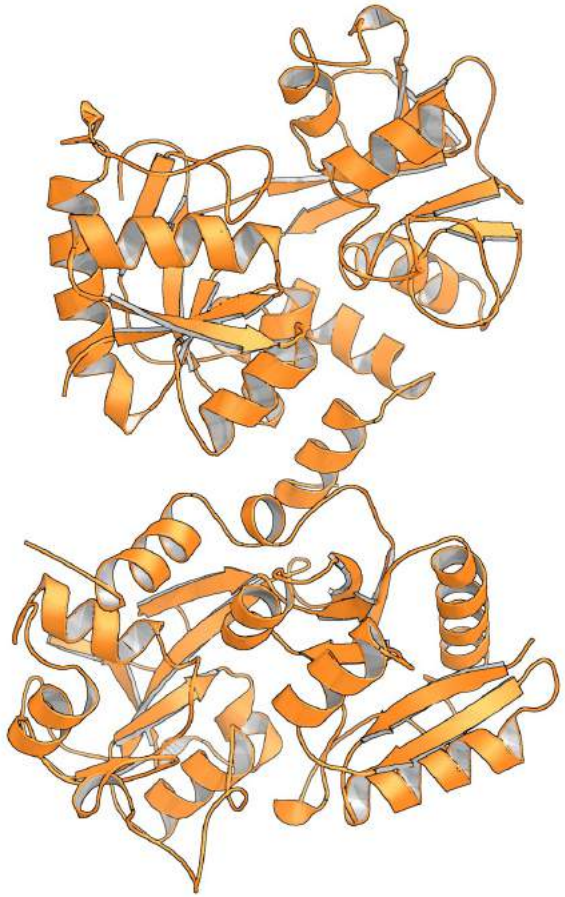
Motivational Example

Ovotransferrin



1ryx – 3.5Å

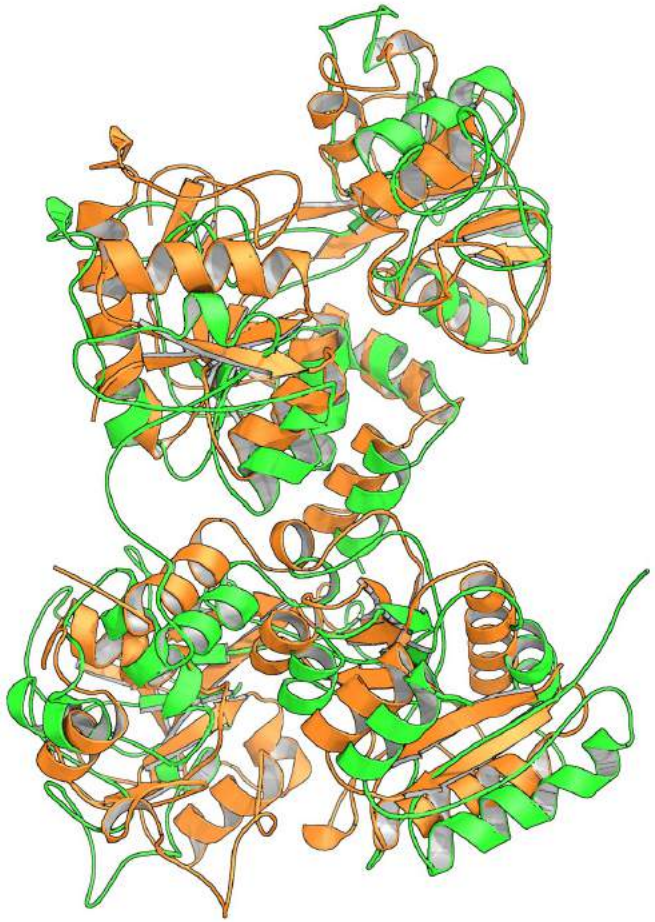
High-resolution homologue



2d3i – 2.15Å

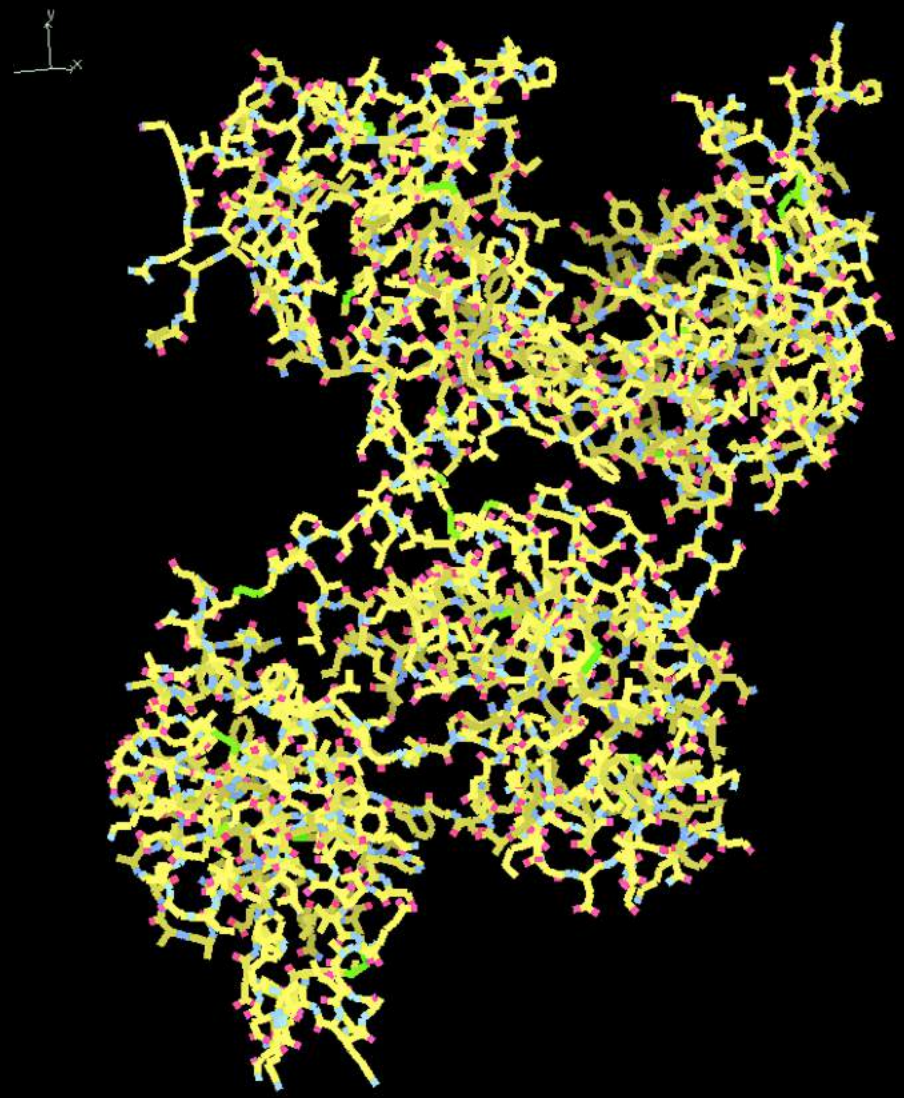
Motivational Example

Ovotransferrin



Models don’t superpose well

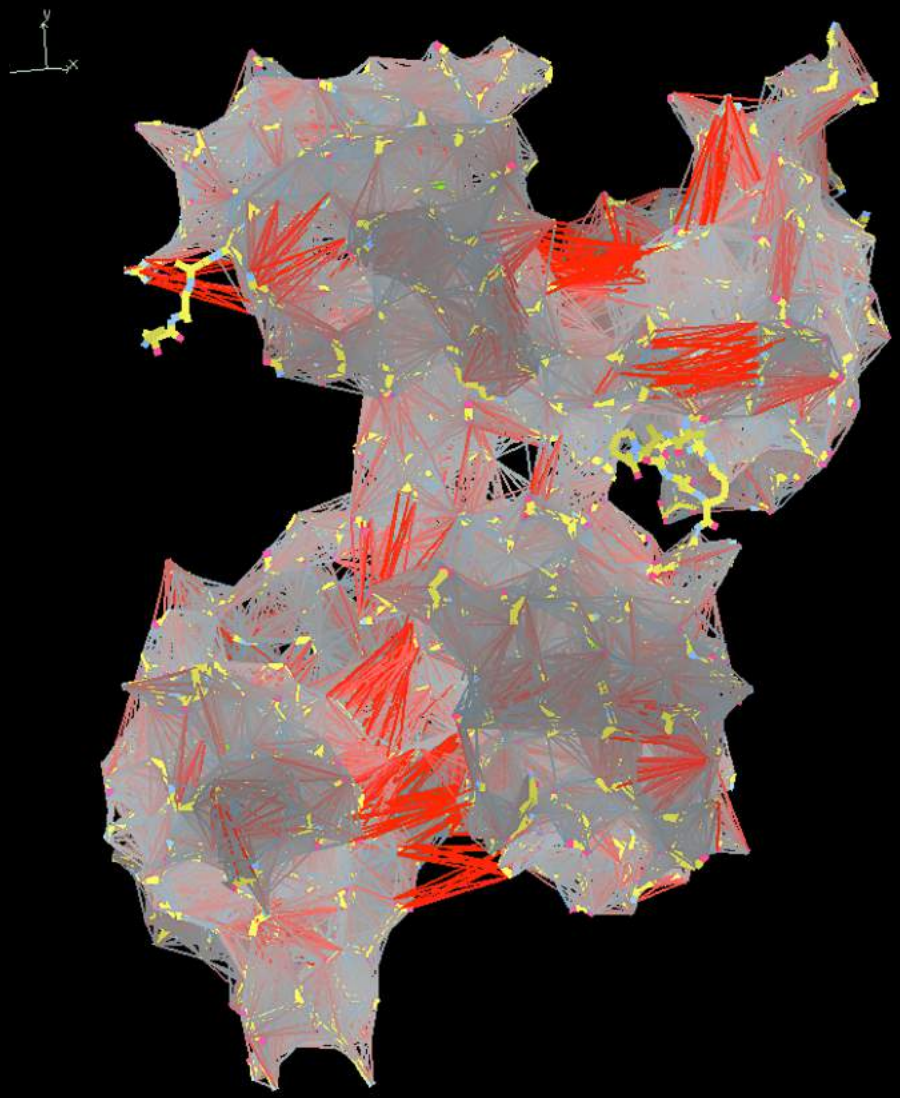
Example: Ovotransferrin



1ryx (3.5Å)

Example: Ovotransferrin

Restraints:
Backbone
Side chains



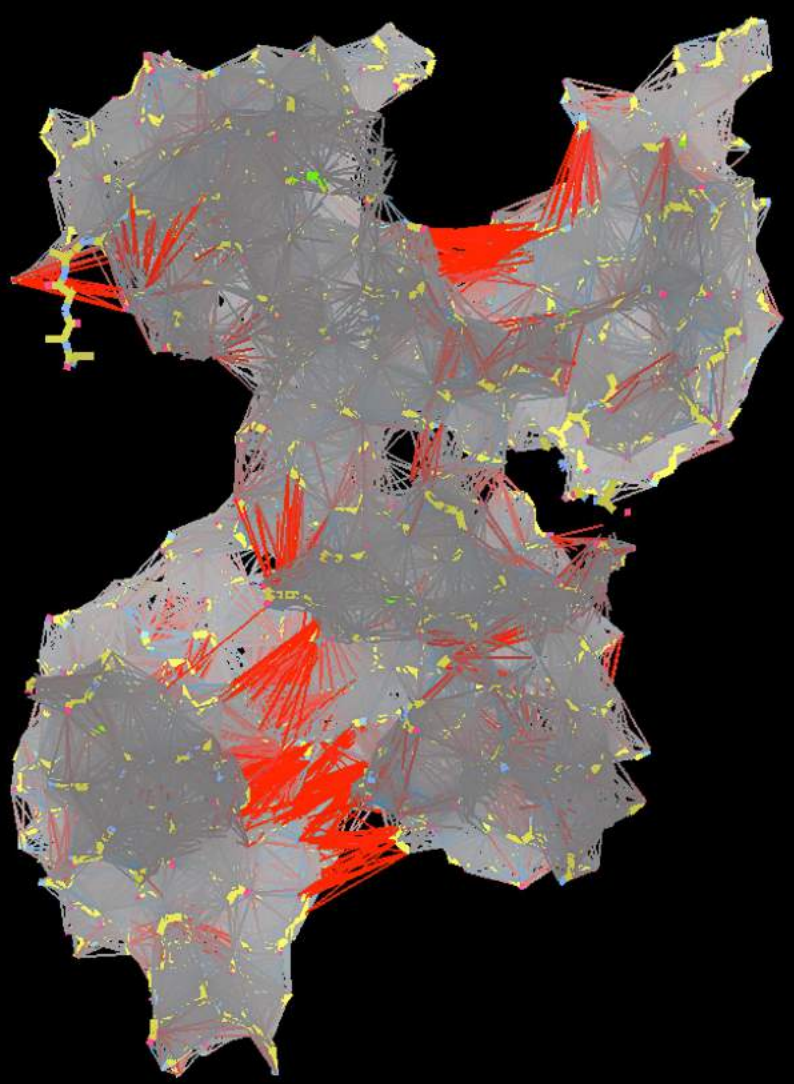
1ryx (3.5Å)
restrained to
2d3i (2.15Å)

Red: long
Grey: similar
Blue: short

Example: Ovotransferrin

Restraints:
- Backbone
- Side chains

After re-refinement

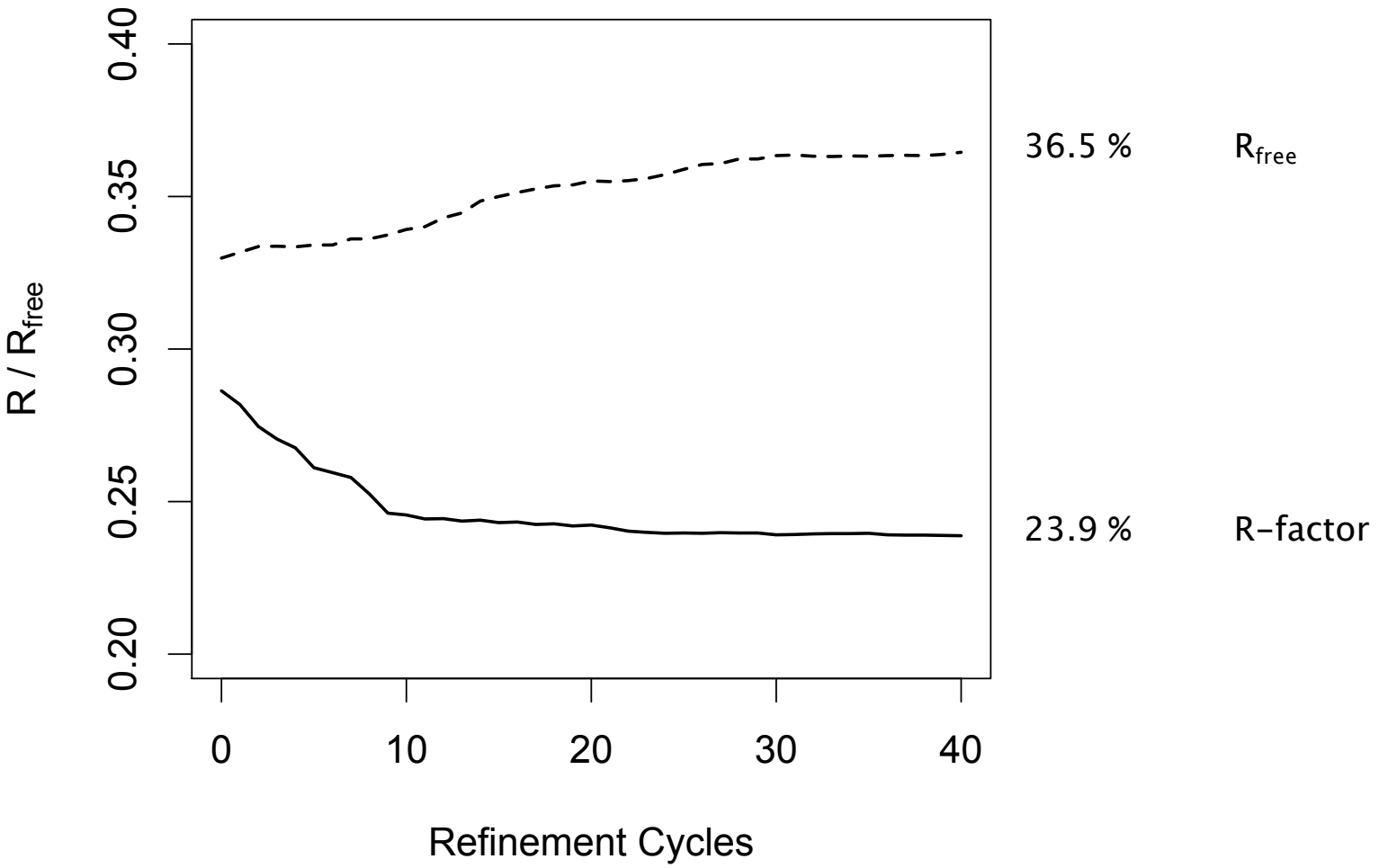


1ryx (3.5Å)
restrained to
2d3i (2.15Å)

Red: long
Grey: similar
Blue: short

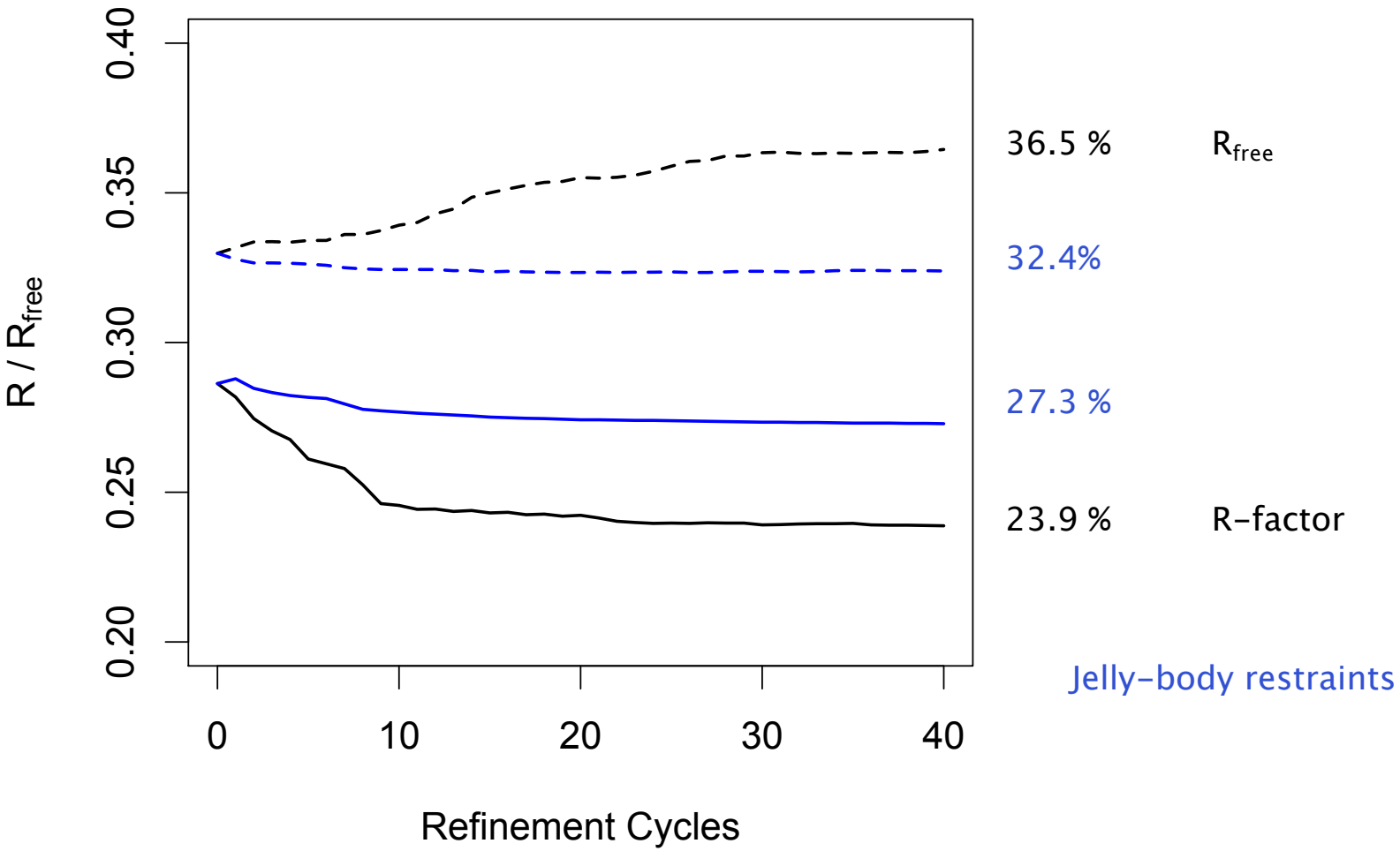
External Restraints

Ovotransferrin



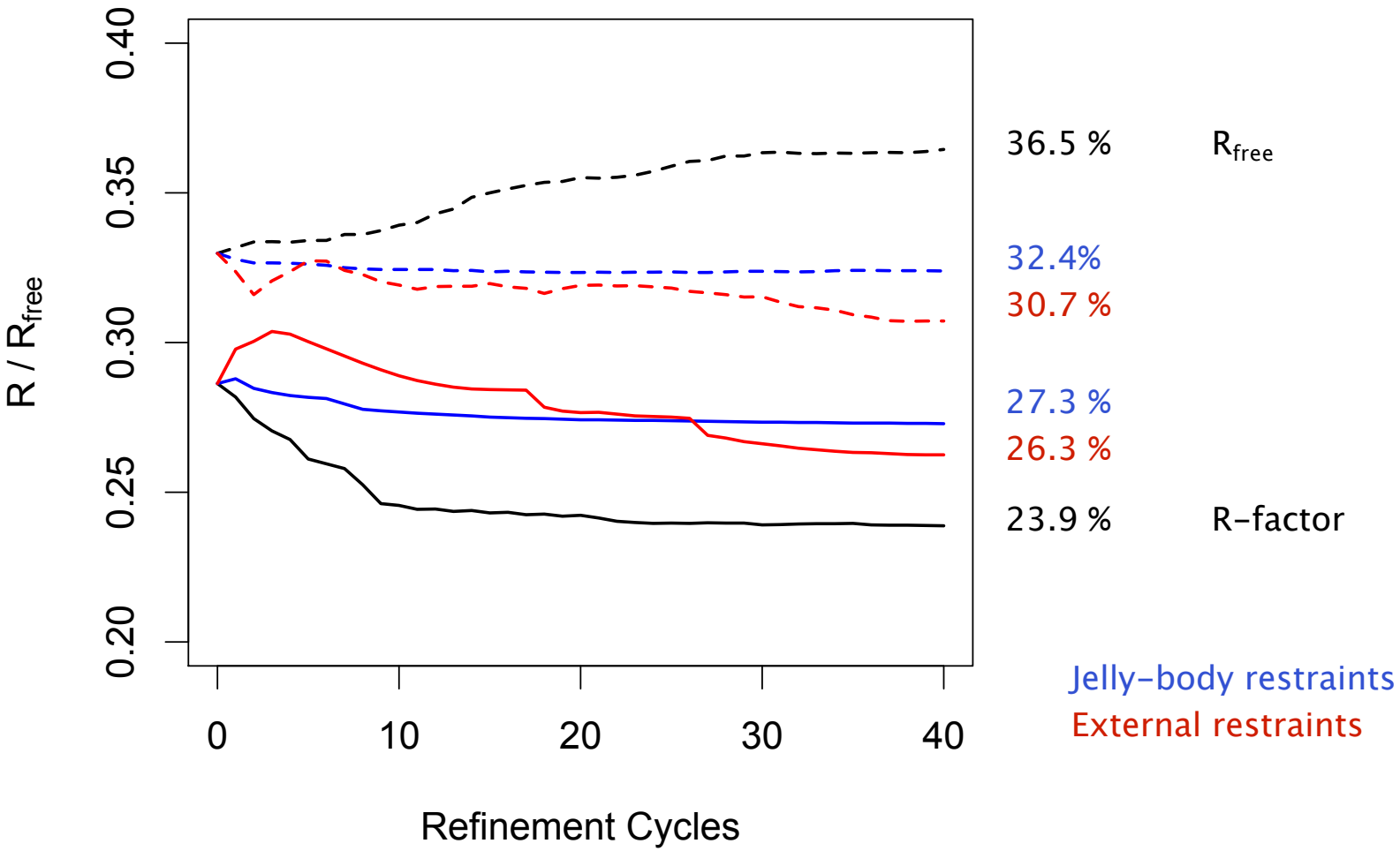
External Restraints

Ovotransferrin



External Restraints

Ovotransferrin

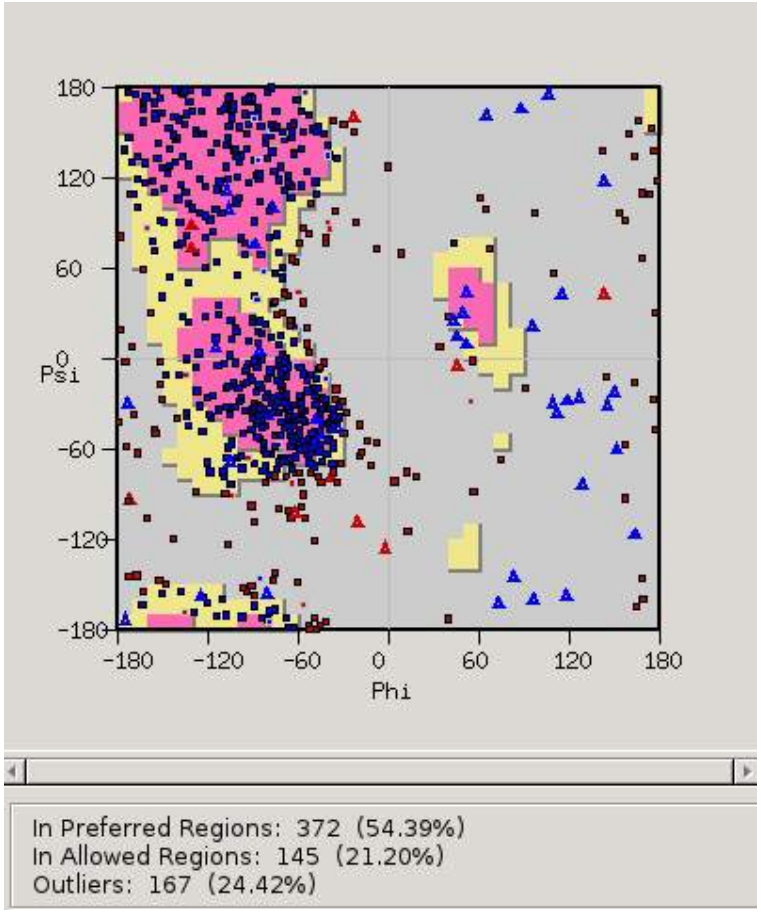


External Restraints

Ovotransferrin

Original Structure

R/R_{free} : 0.286/0.330

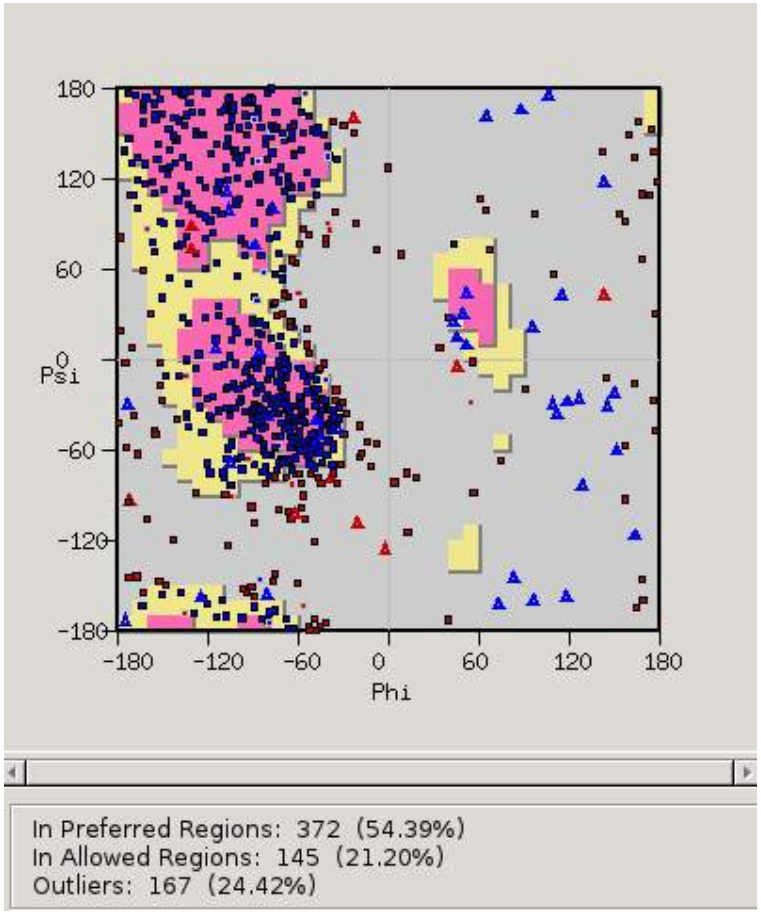


External Restraints

Ovotransferrin

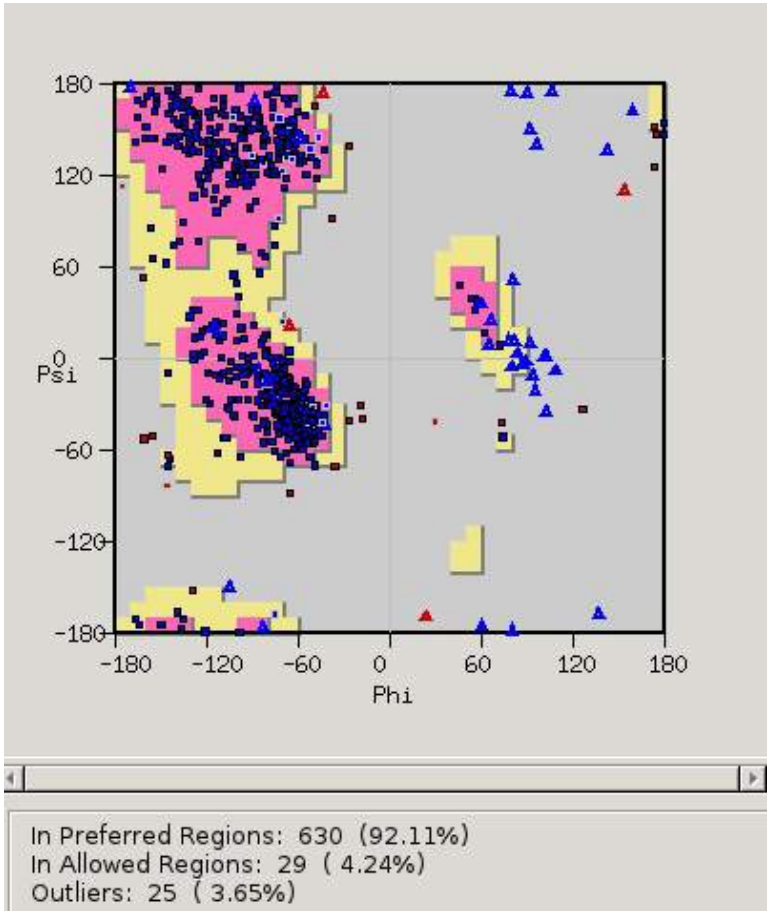
Original Structure

R/R_{free} : 0.286/0.330

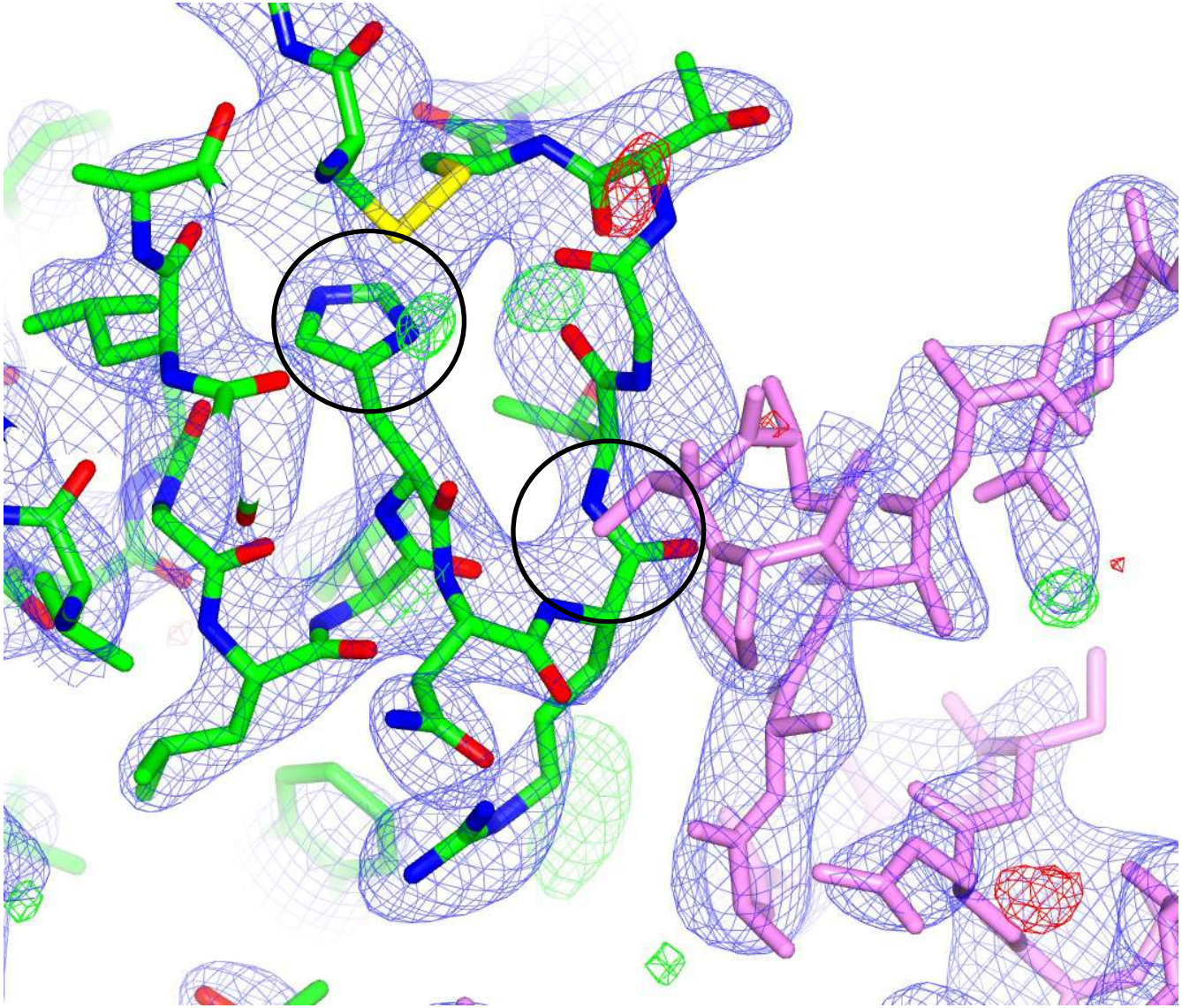


Re-refined with External Restraints

R/R_{free} : 0.263/0.307



External Restraints



1.3σ

Original Structure

R/R_{free} : 0.286/0.330

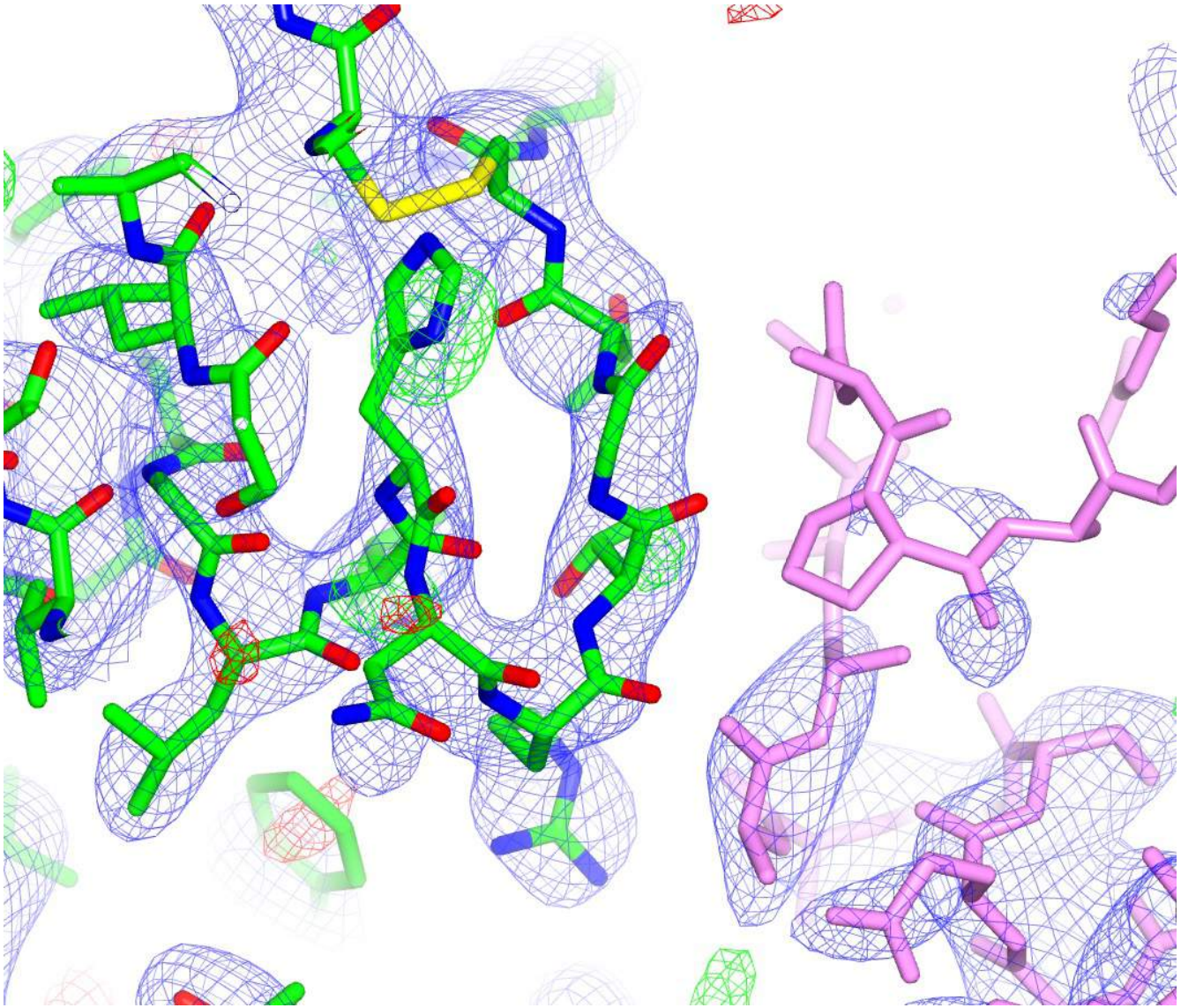
↓

External restraints

(40 cycles)

R/R_{free} : 0.263/0.307

External Restraints



1.3σ

Original Structure

R/R_{free} : 0.286/0.330

↓

External restraints

(40 cycles)

R/R_{free} : 0.263/0.307

↓

Modify

Real Space Refine

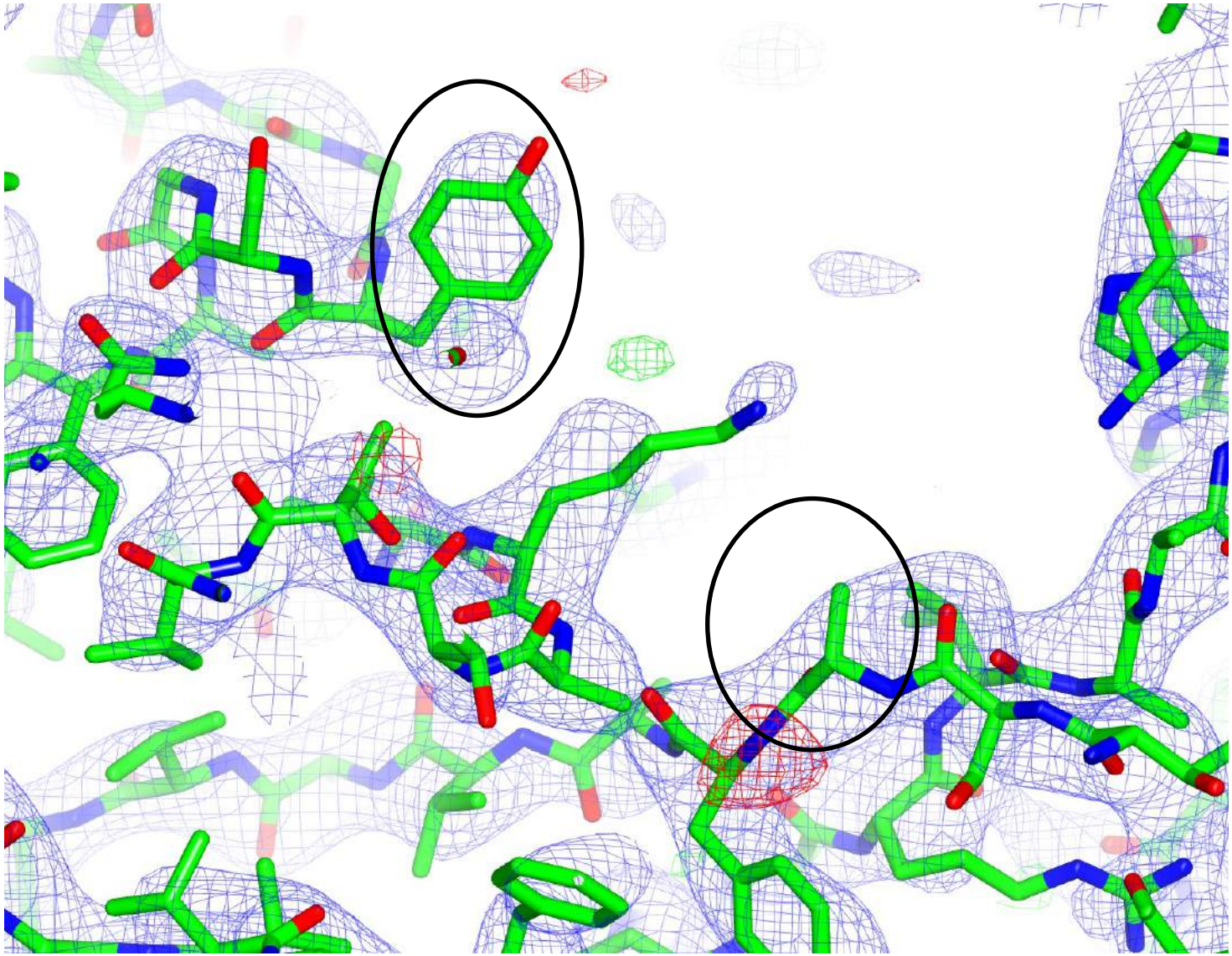
↓

Jelly body

(40 cycles)

R/R_{free} : 0.253/0.304

External Restraints



1.3σ

Original Structure

R/R_{free} : 0.286/0.330

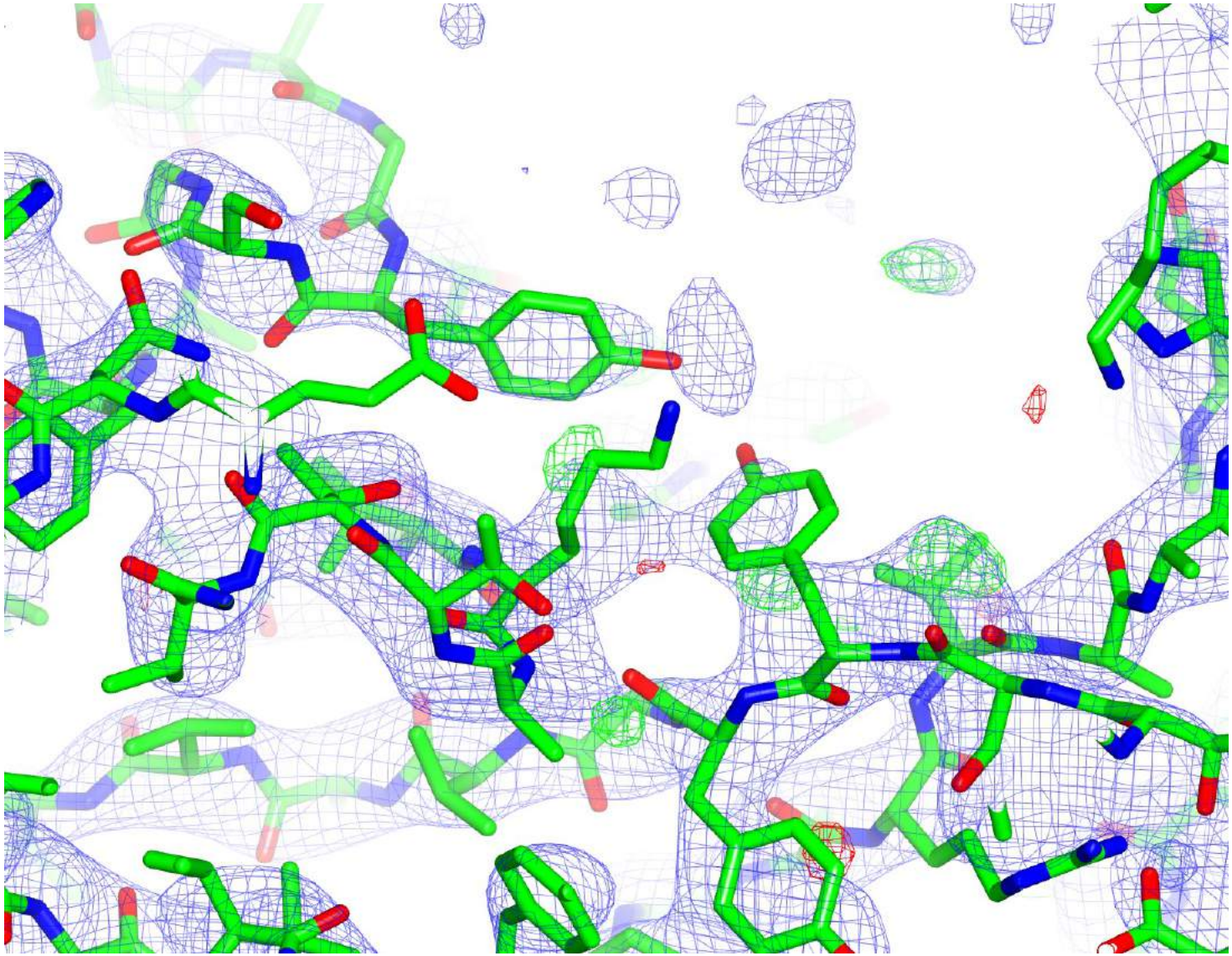
↓

External restraints

(40 cycles)

R/R_{free} : 0.263/0.307

External Restraints



1.3σ

Original Structure

R/R_{free} : 0.286/0.330

↓

External restraints

(40 cycles)

R/R_{free} : 0.263/0.307

↓

Build TYR92

Modify LYS209

↓

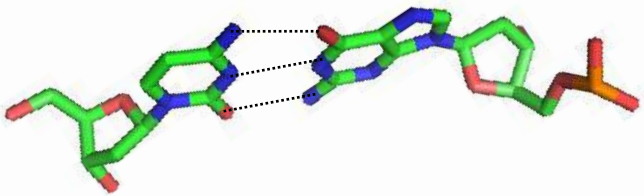
Jelly body

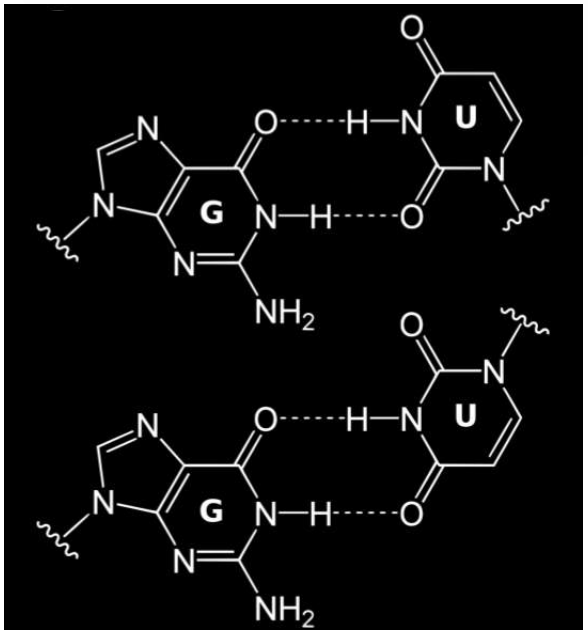
(40 cycles)

R/R_{free} : 0.252/0.307

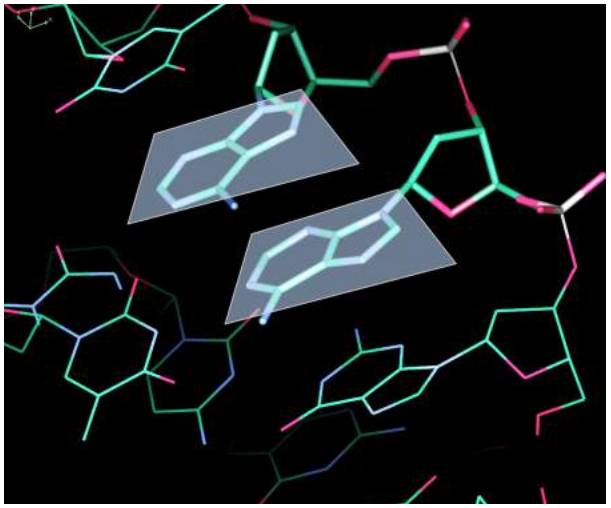
LibG Nucleic Acid Restraints

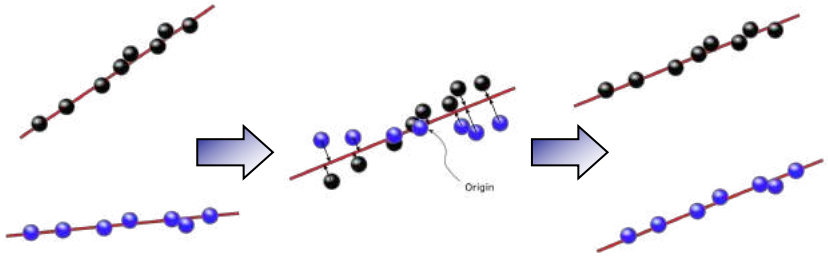
Base-pair restraints



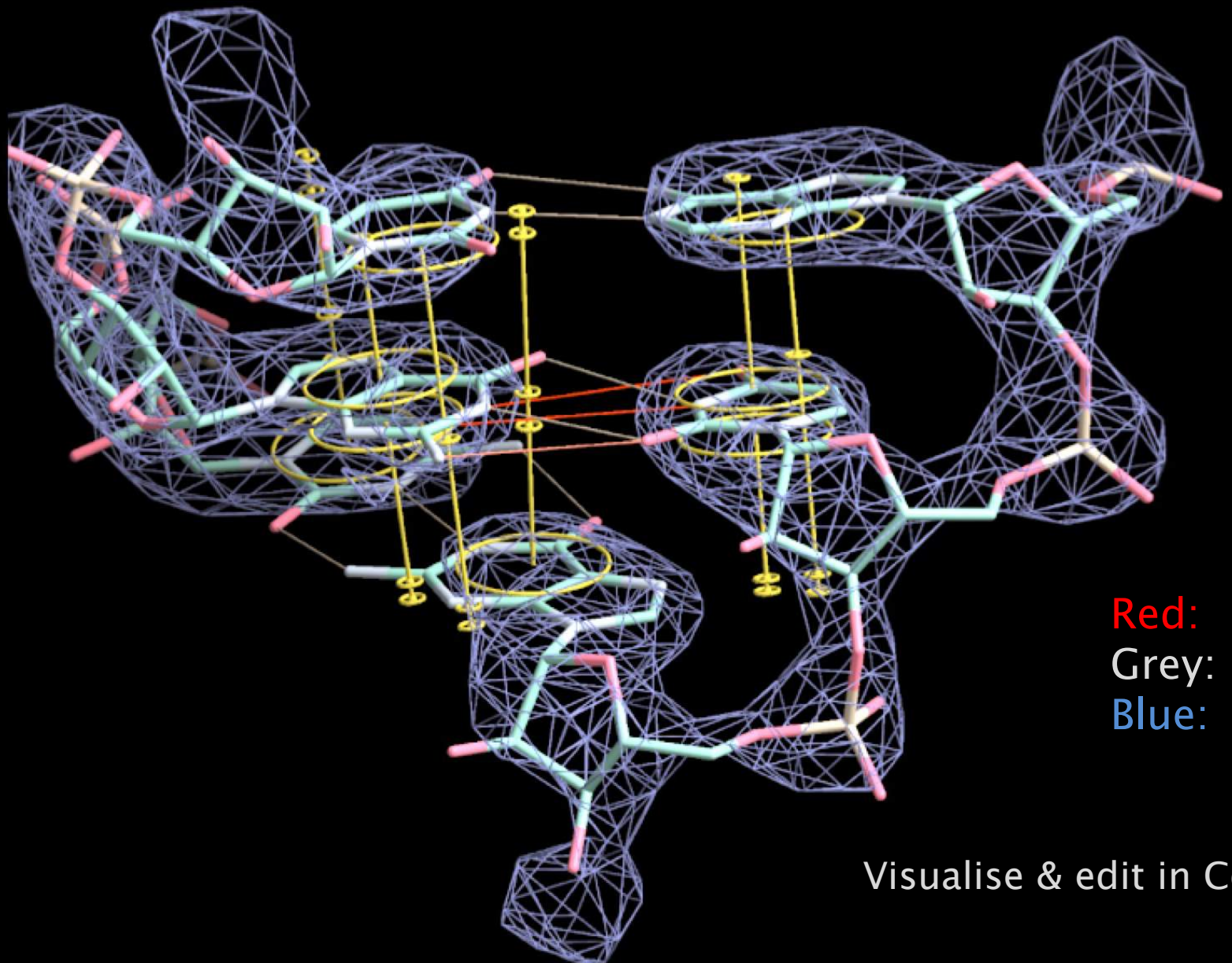


Parallel plane restraints





LibG Nucleic Acid Restraints



Red: long
Grey: similar
Blue: short

Visualise & edit in Coot

External Restraints

When refining at low resolution, check:

- Refinement statistics – ***Not always conclusive***
- Geometry – ***Not always conclusive***
- Electron density – ***Not always reliable***

Conclusion: **At low resolution, everything has to add up!**

Take care; reflect

Quality of prior information is important – **consider manual re-refinement**

– **PDB-REDO is useful**


DRINKS

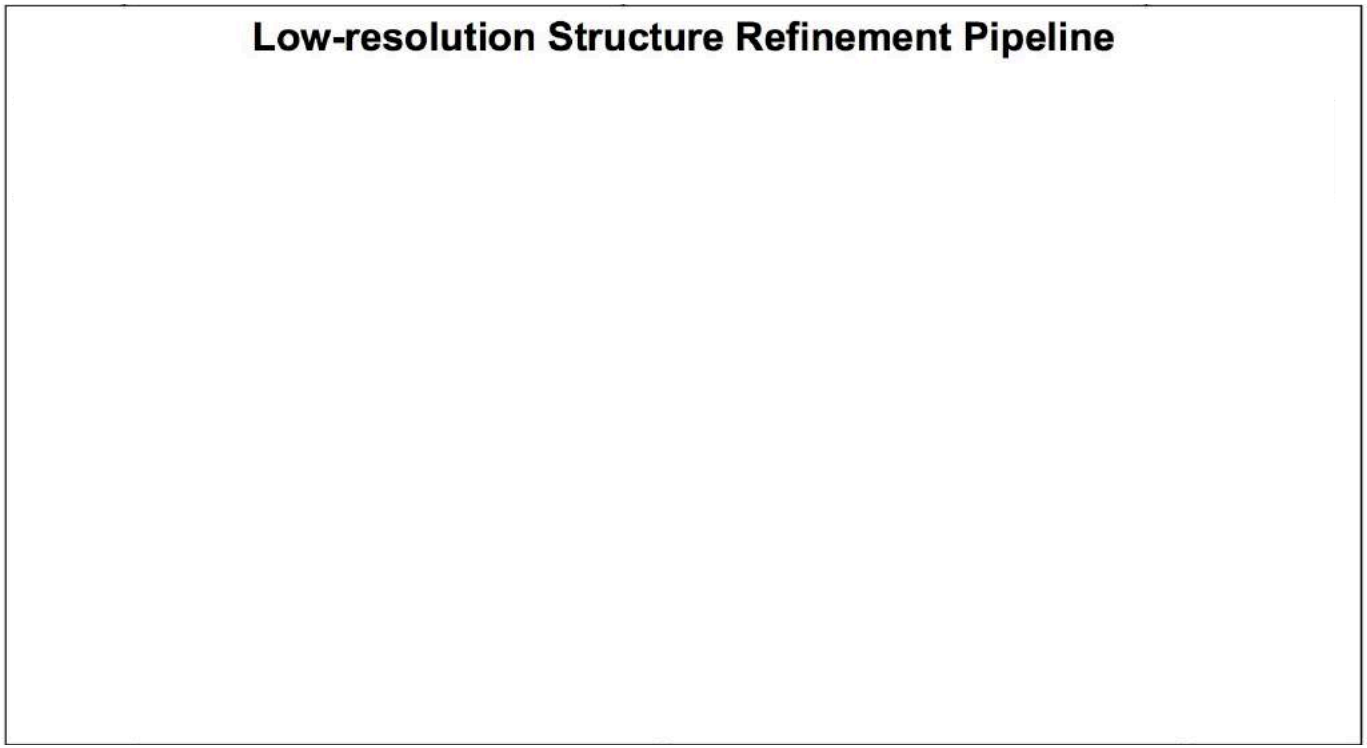
Automated pipeline - LORESTR

- Efficiency of ProSMART-generated restraints greatly depends on the homologues used
- If several homologues are available, substantial manual effort is required to find their optimal combination
- Other refinement parameters (scaling, solvent, etc) also affect efficiency of the process

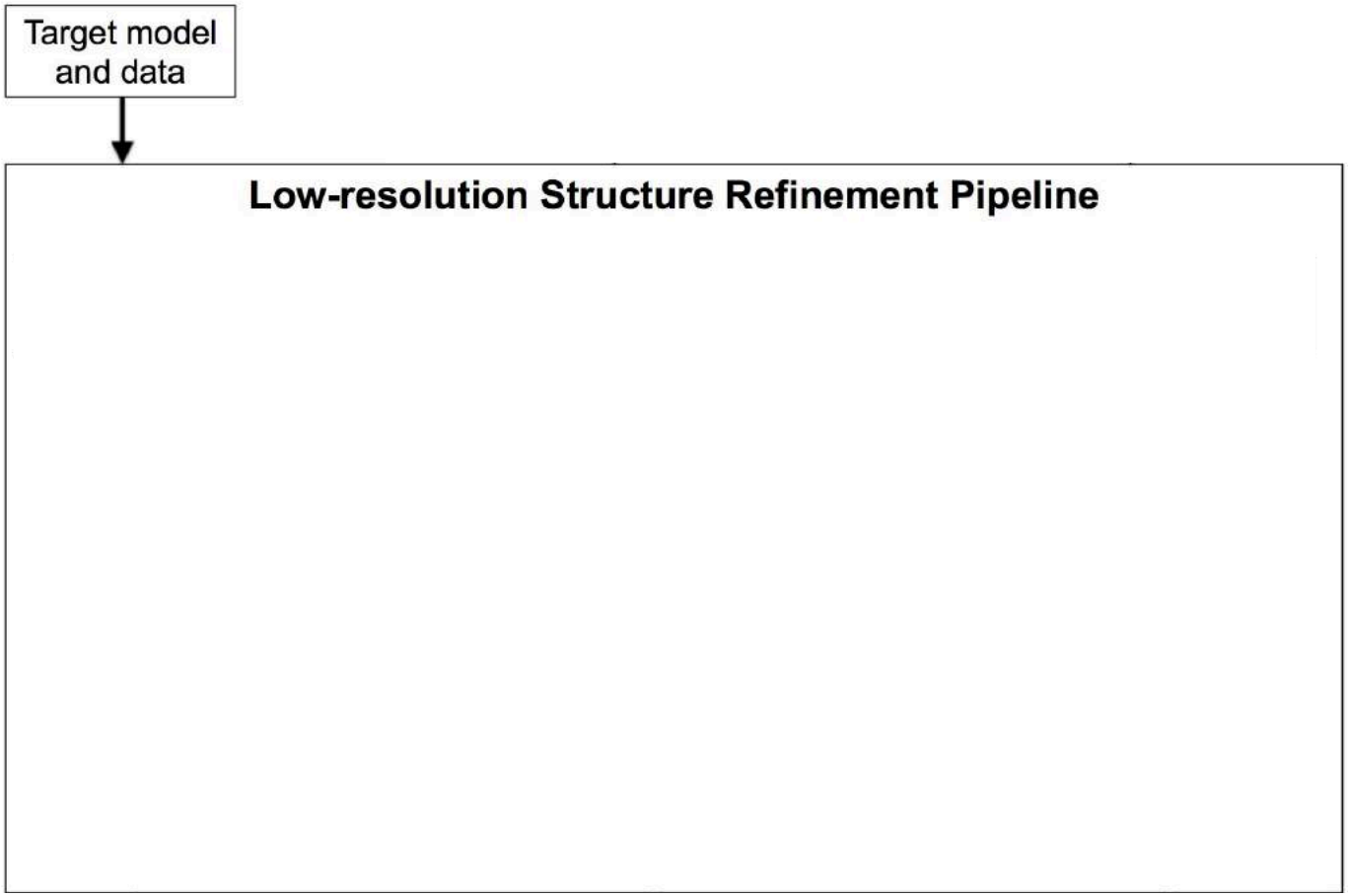
Solution:

LOw-**RE**solution **ST**ructure **R**efinement

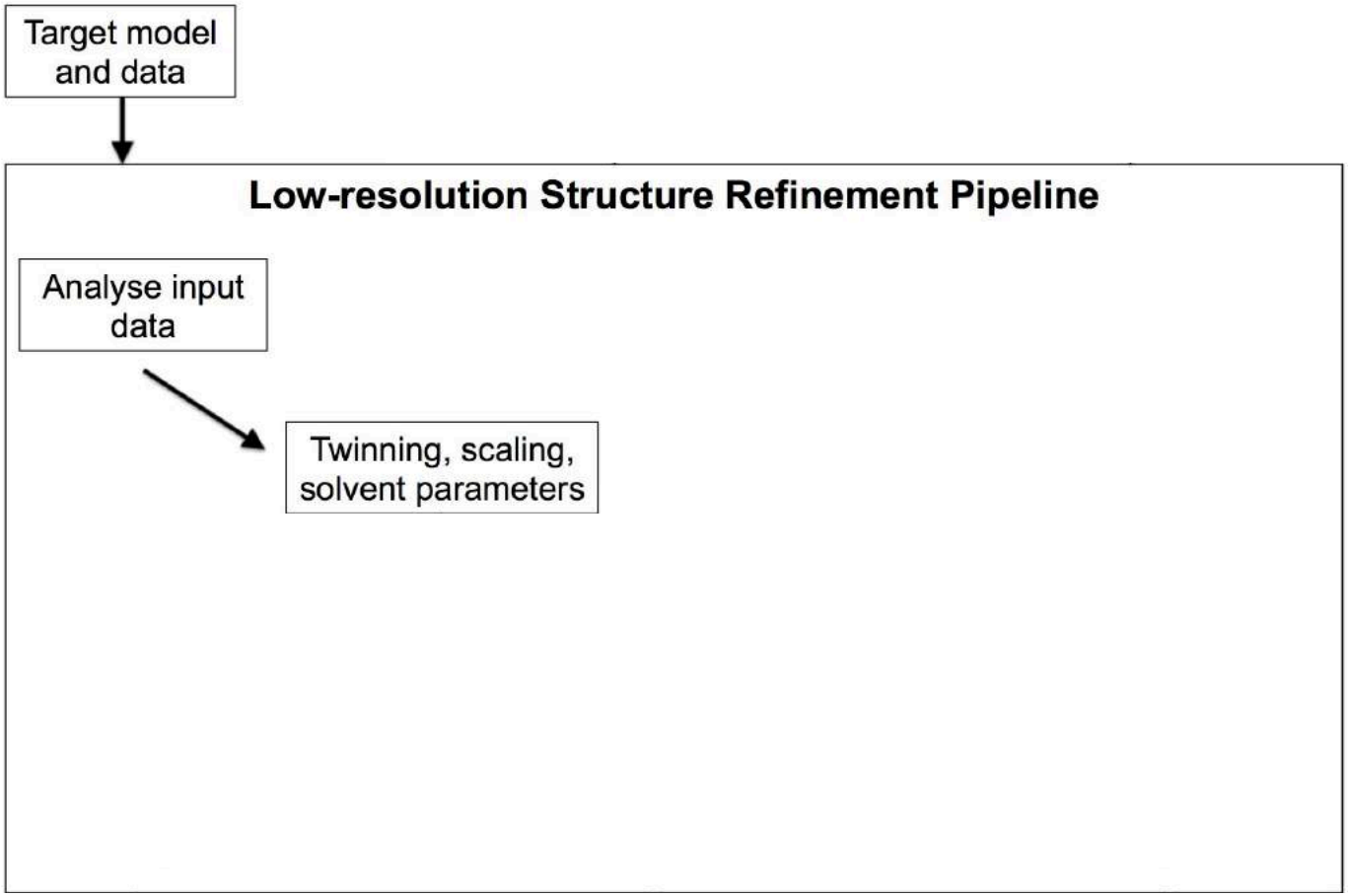
Automated pipeline - LORESTR



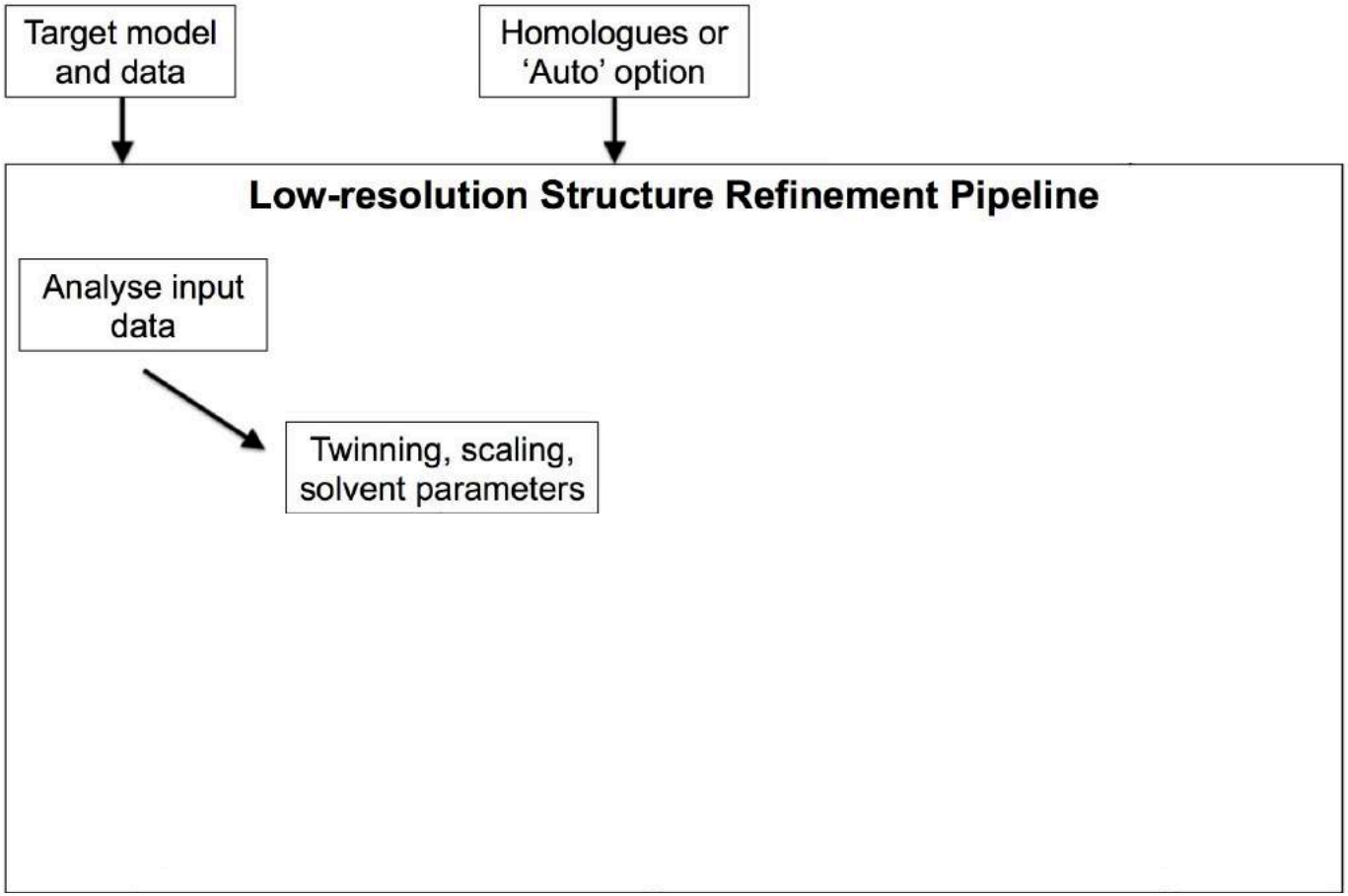
Automated pipeline - LORESTR



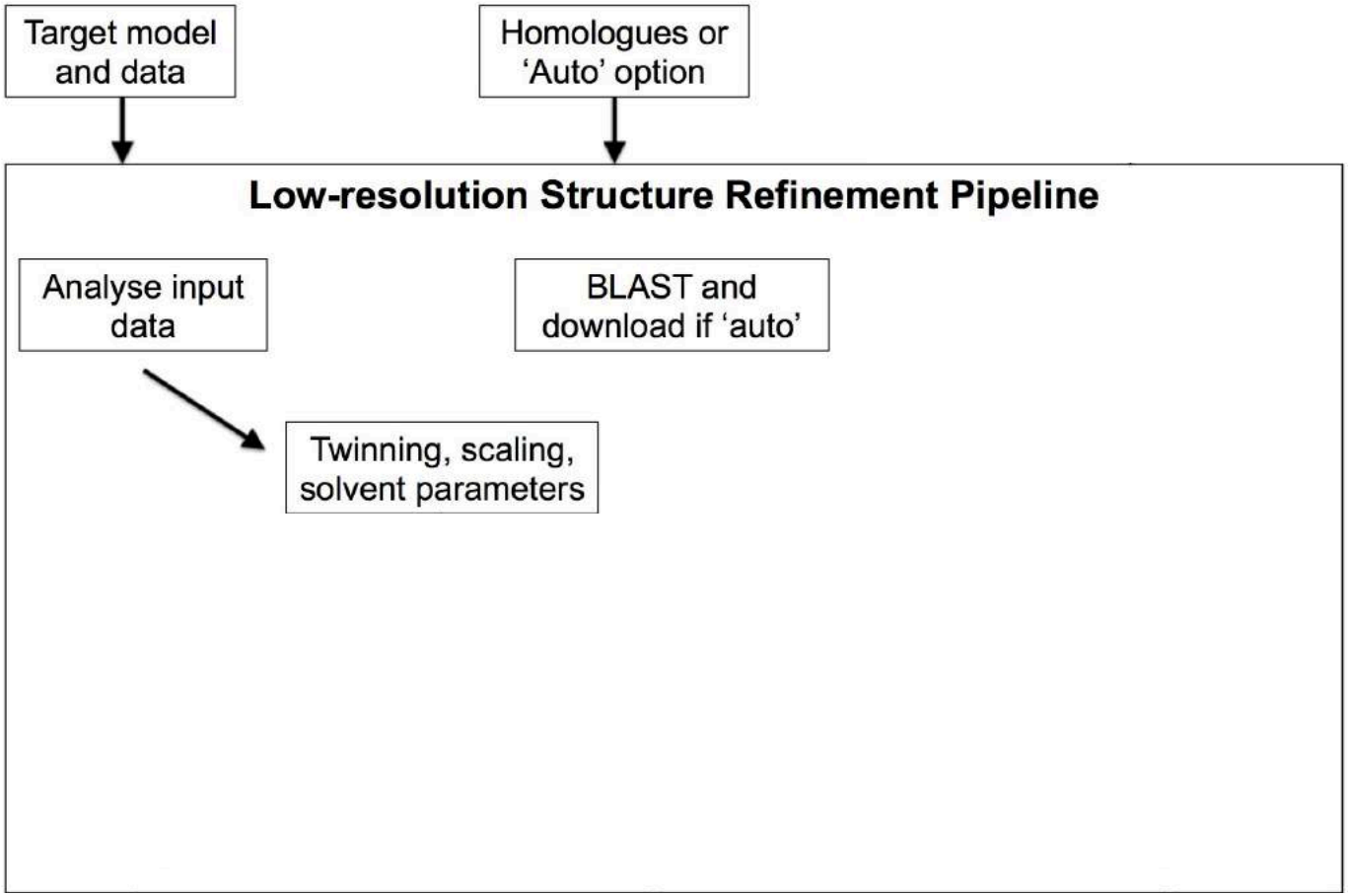
Automated pipeline - LORESTR



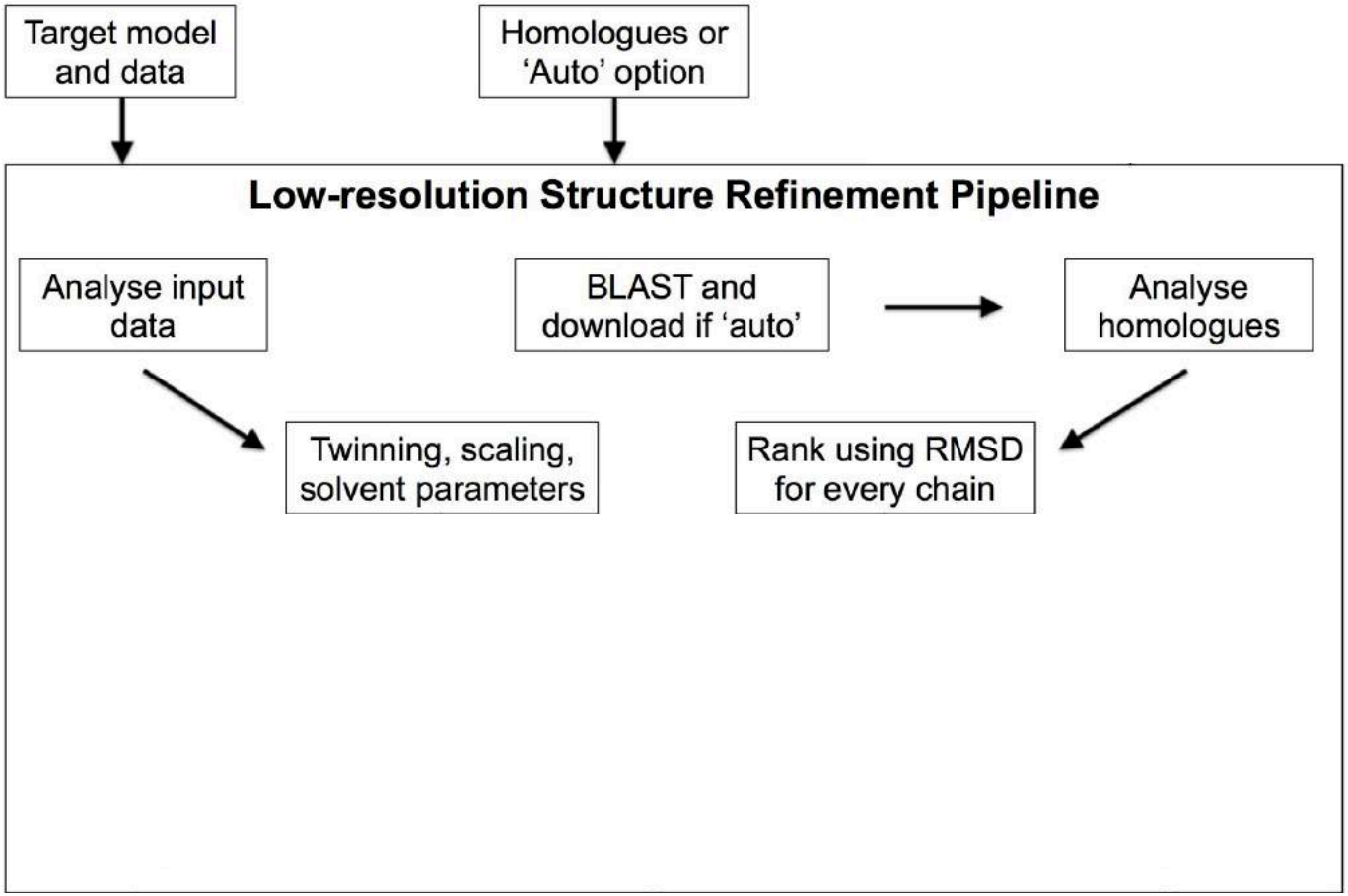
Automated pipeline - LORESTR



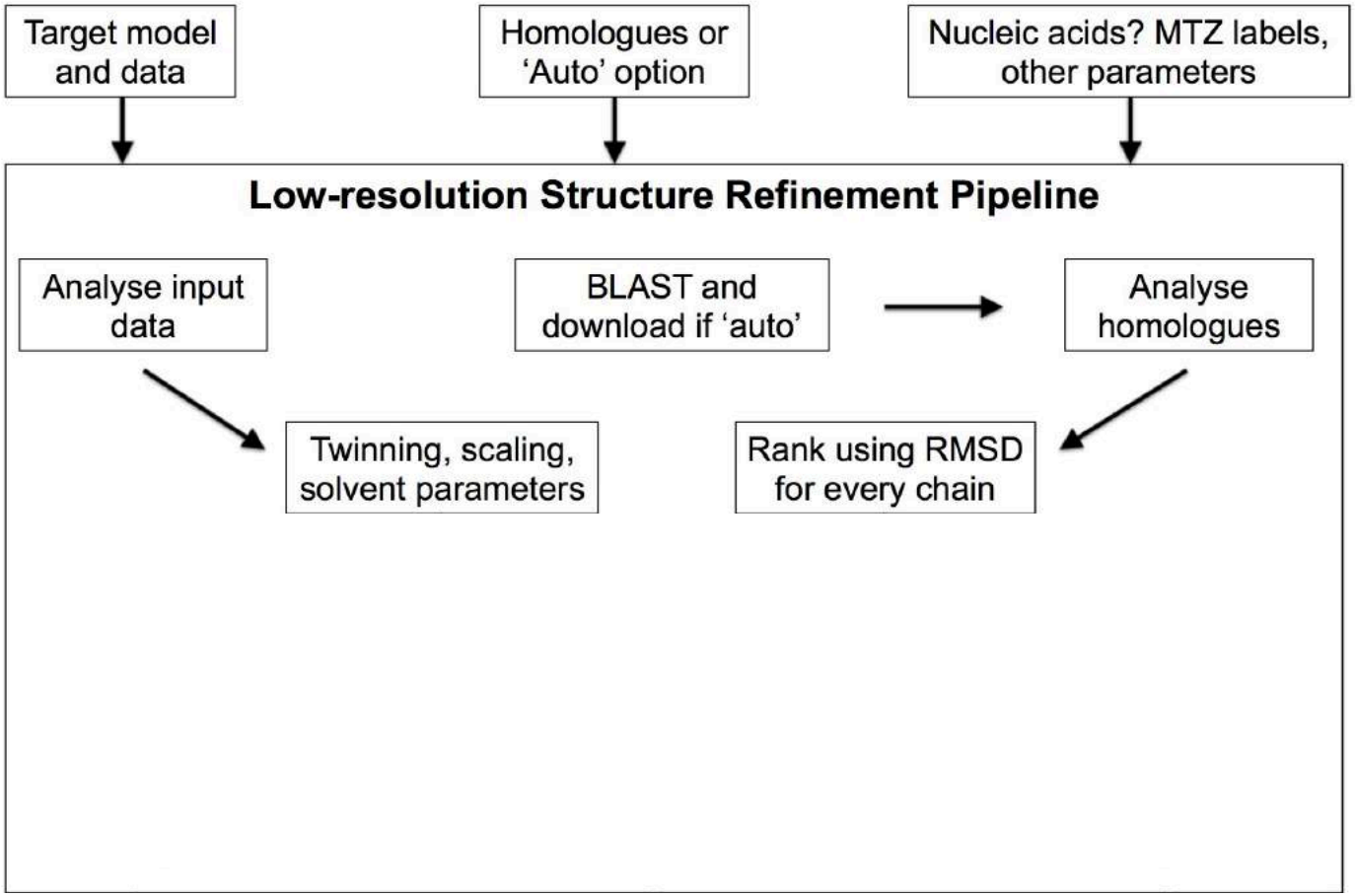
Automated pipeline - LORESTR



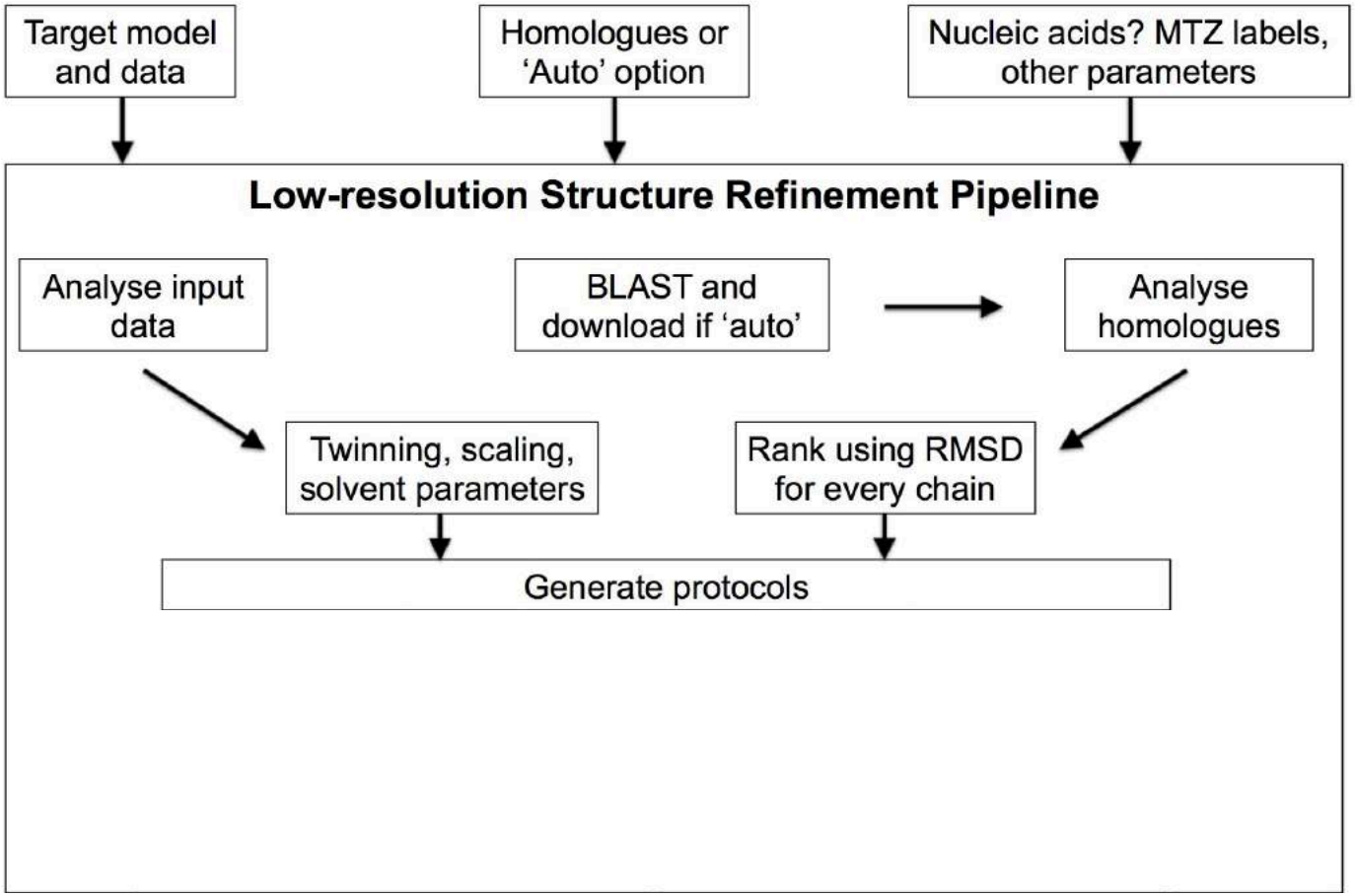
Automated pipeline - LORESTR



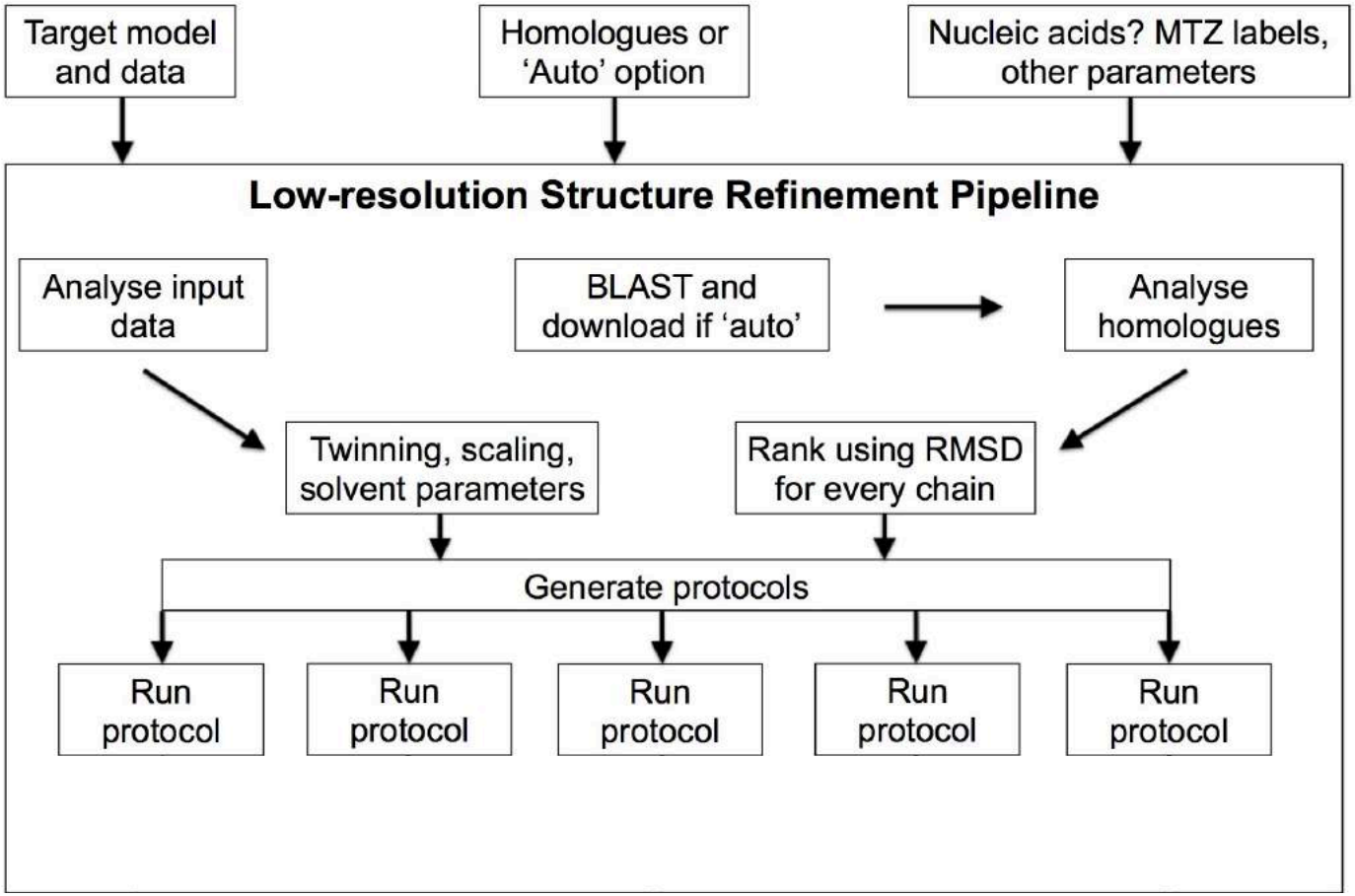
Automated pipeline - LORESTR



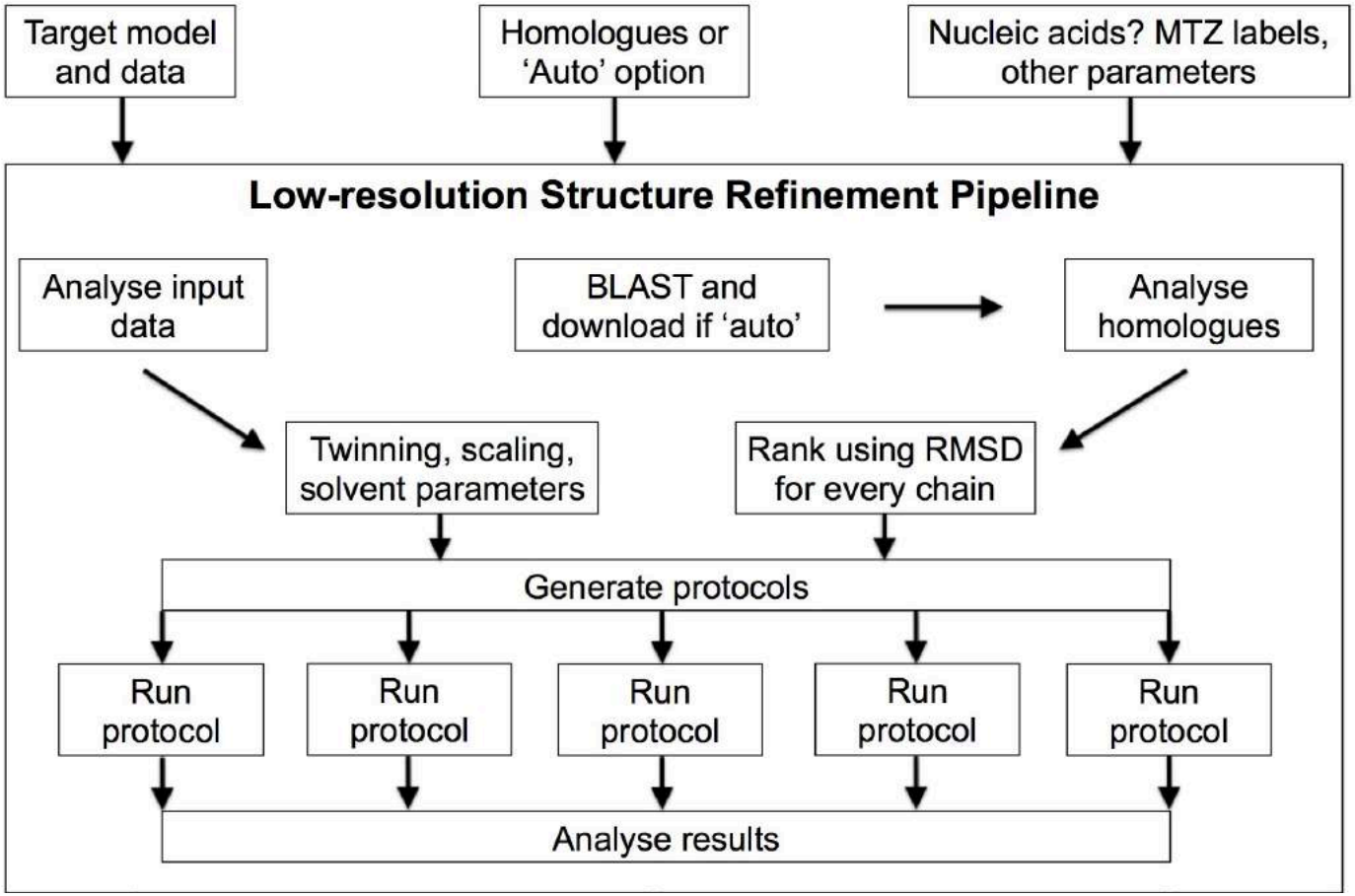
Automated pipeline - LORESTR



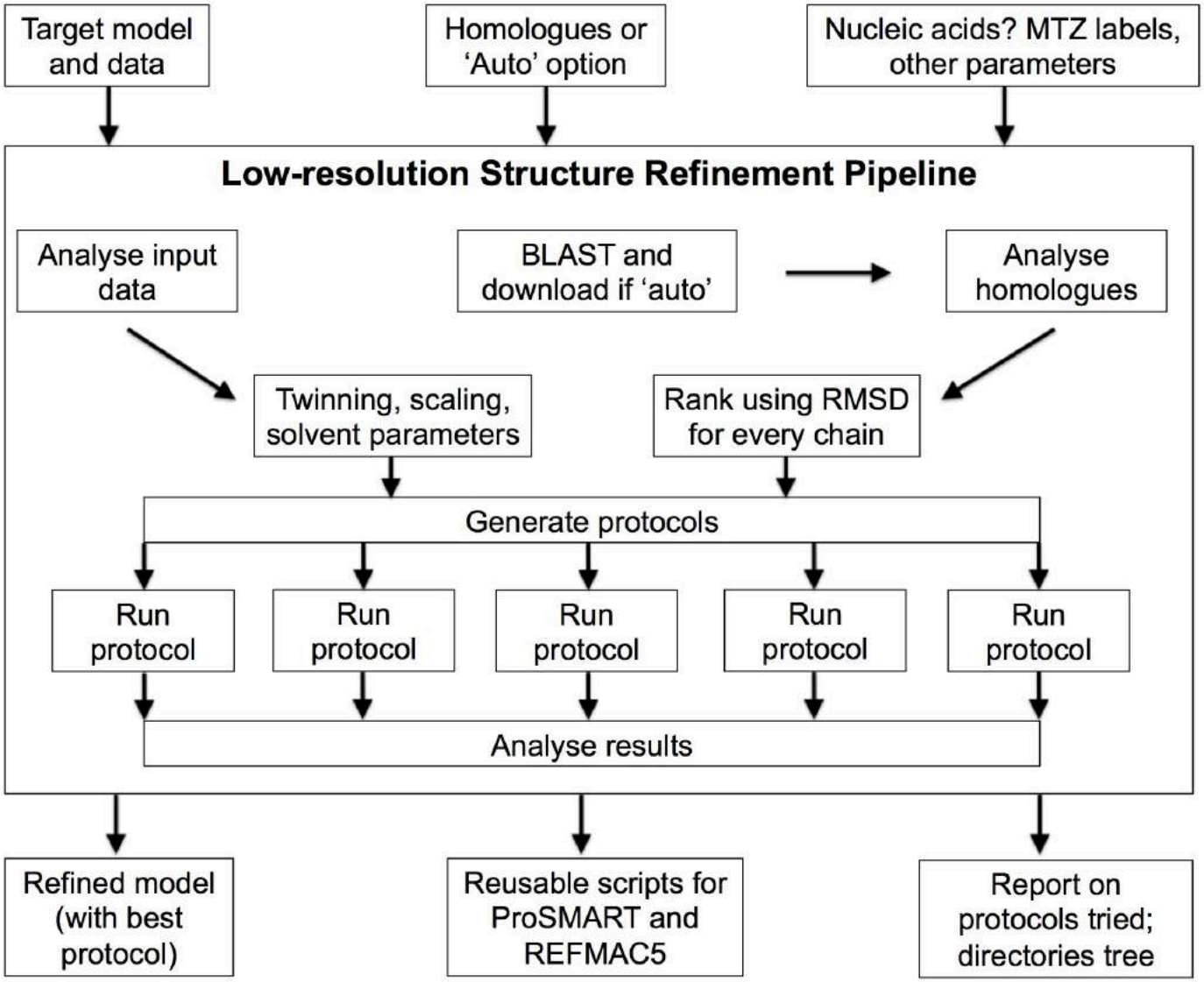
Automated pipeline - LORESTR



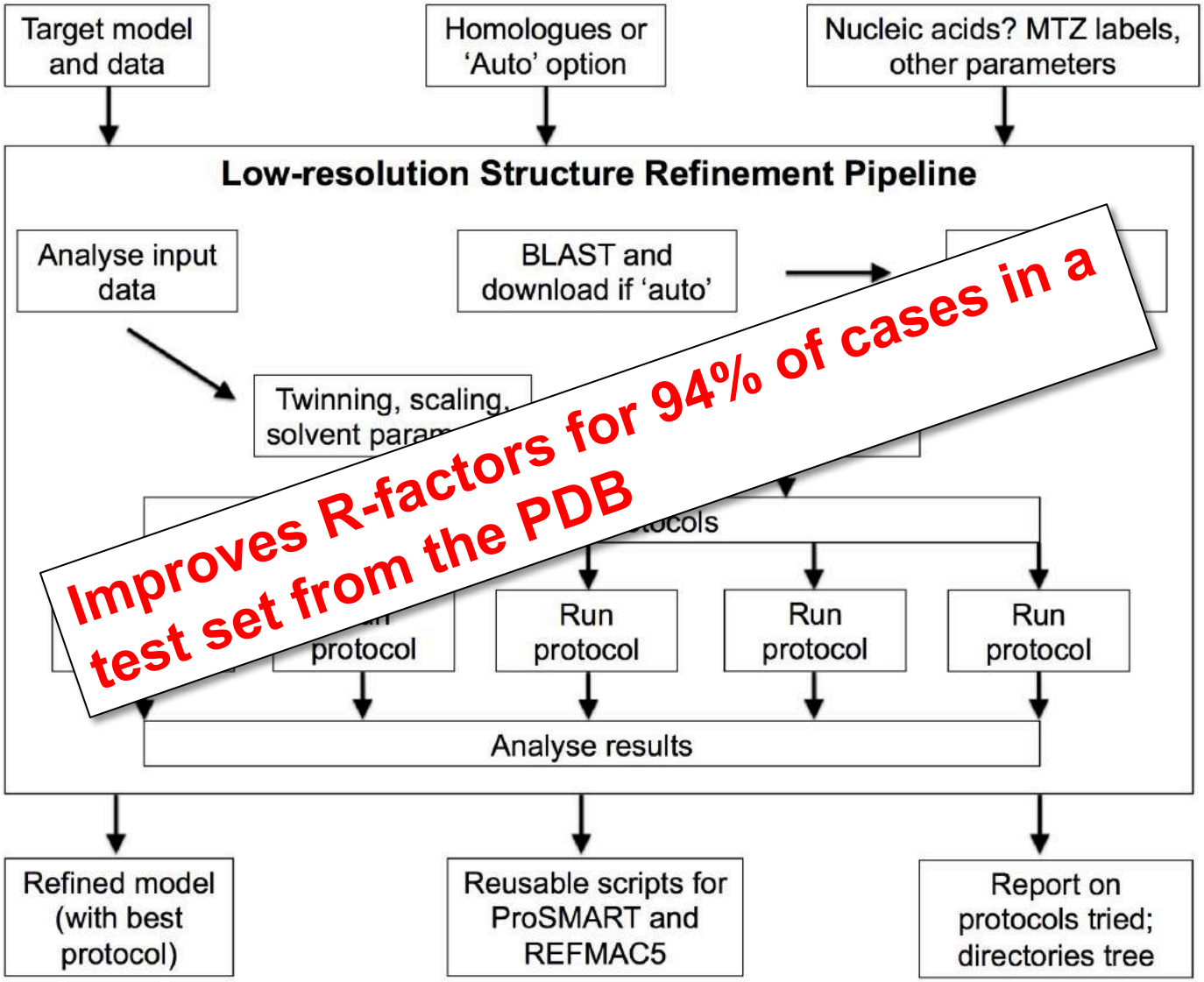
Automated pipeline - LORESTR



Automated pipeline - LORESTR



Automated pipeline - LORESTR



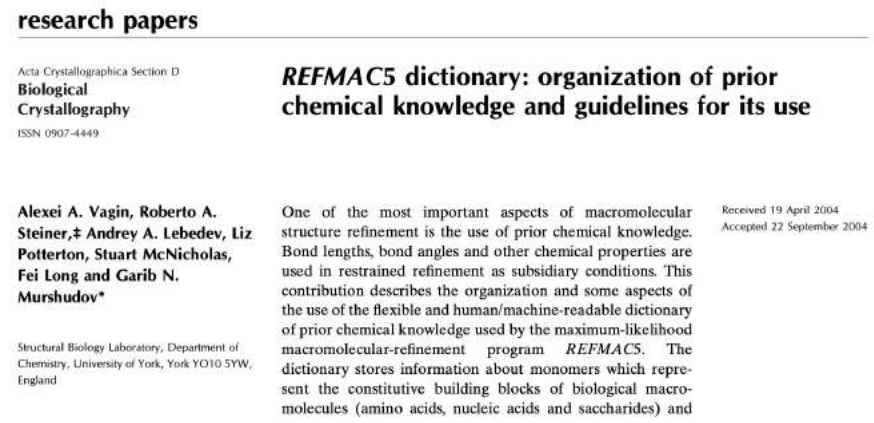
Restraints for Ligands

Geometric restraints for protein / nucleic acids are pre-tabulated

- **Common/known structures are dealt with automatically**
  - CCP4/REFMAC monomer library has pre-computed descriptions

- **New ligands require description (CIF file)**
  - AceDRG

CCP4 Monomer Library



research papers

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

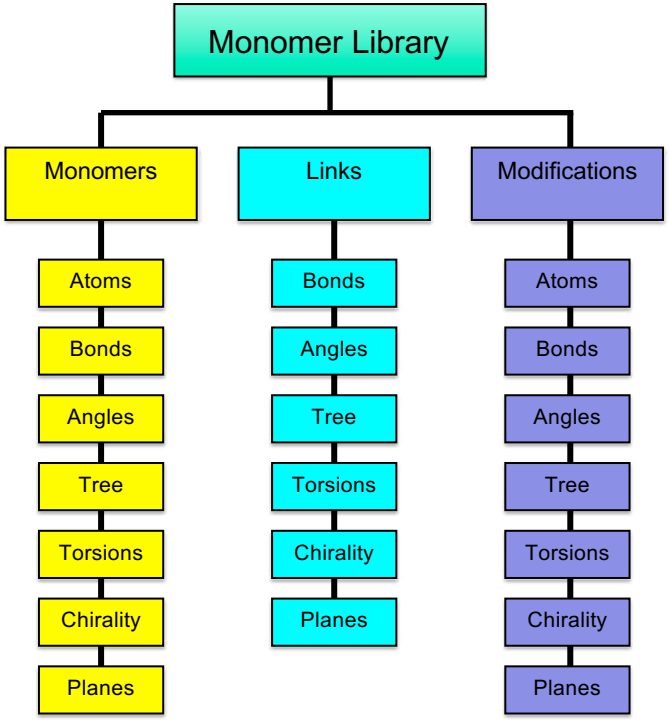
REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use

Alexei A. Vagin, Roberto A. Steiner,‡ Andrey A. Lebedev, Liz Potterton, Stuart McNicholas, Fei Long and Garib N. Murshudov*

Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5YW, England

One of the most important aspects of macromolecular structure refinement is the use of prior chemical knowledge. Bond lengths, bond angles and other chemical properties are used in restrained refinement as subsidiary conditions. This contribution describes the organization and some aspects of the use of the flexible and human/machine-readable dictionary of prior chemical knowledge used by the maximum-likelihood macromolecular-refinement program *REFMAC5*. The dictionary stores information about monomers which represent the constitutive building blocks of biological macromolecules (amino acids, nucleic acids and saccharides) and

Received 19 April 2004
Accepted 22 September 2004



- **30,347 monomers** in the dictionary
- > **100** entries of **modifications and links**
- **CCP4-ML entries** are generated by **AceDRG**

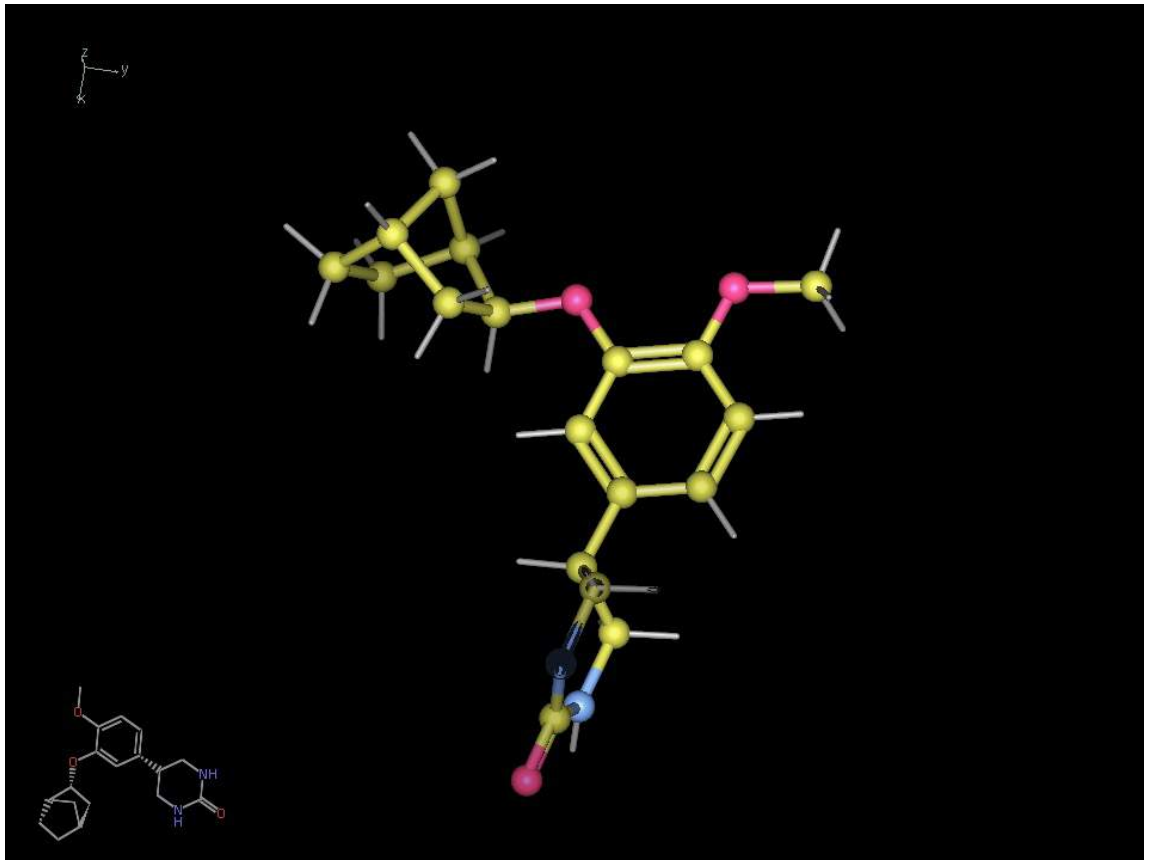
Vagin, A. A. et al., (2004). *Acta Cryst.* D**60**, 2184-2195

AceDRG

New atom types

Full 2nd order neighbour-based atom description

(3rd order in some cases)



AceDRG derives atom types from small molecular databases
These are tabulated, distributed as part of CCP4 for quick lookup

AceDRG

Functionalities:

(1) Restraint Dictionary Generator

- Uses restraint tables to generate restraints for given molecule
- Inputs – mmCIF, SMILES, MDL/SDF, SYBIL/MOL2
- Output – CIF – bond lengths, angles, torsions, planes, chiralities

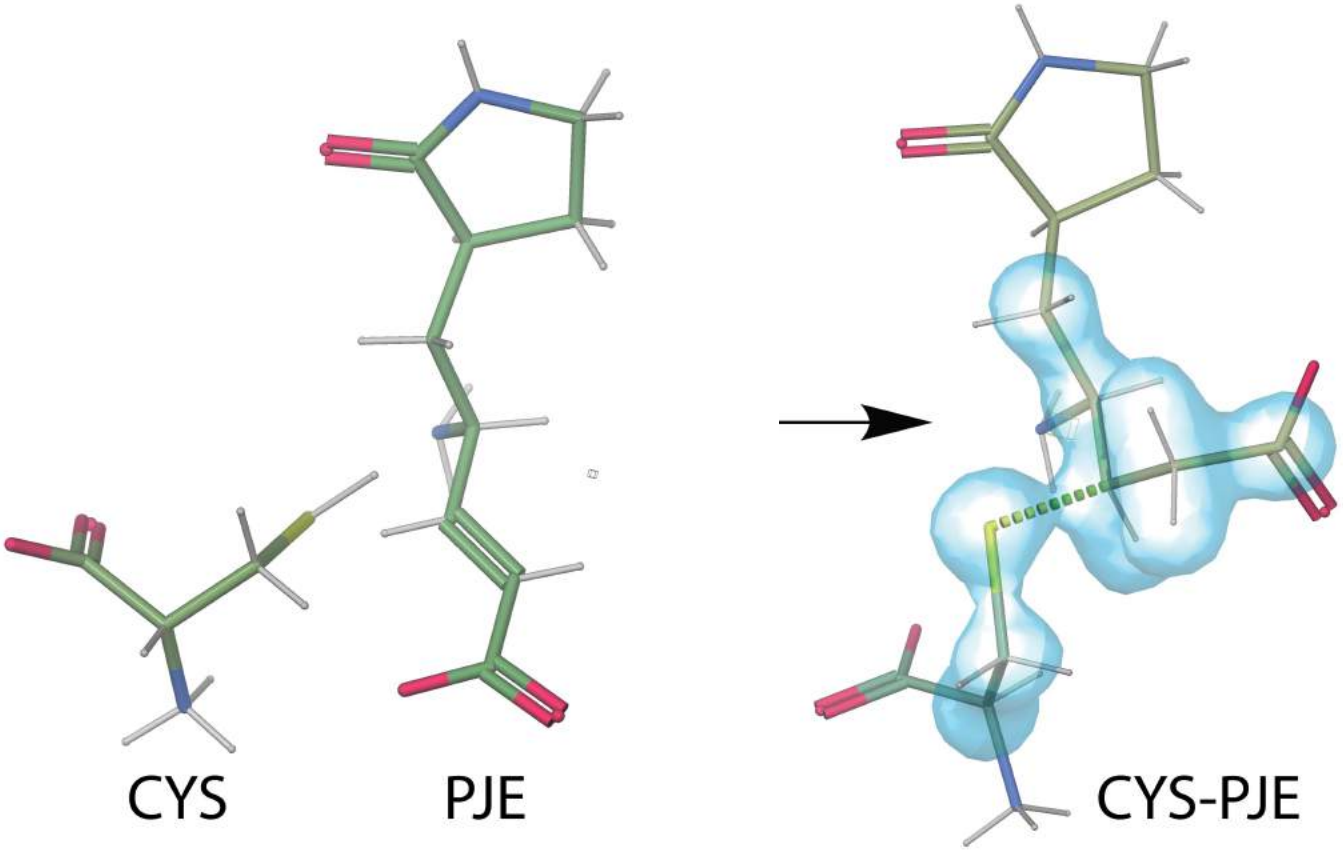
(2) Conformer Generator

- Generates coordinates from graph-based molecule description
- Generates one of the lower-energy conformations
- Refines conformation using Refmac
- Output – PDB

(3) **Link Creation**

- Creates link between two components
- May be from the CCP4 monomer library, or custom CIFs
- Separate operational mode – requires separate execution

Modelling Covalent Linkages



- Link between Cys[SG] and PJE[C20]
- PJE[C20]–[C21] bond order changed from double to single
- Various restraints in the local environment are changed – modifications

Summary

Tools to help with model building and refinement:

REFMAC5: Refinement, jelly body restraints, map sharpening/blurring

ProSMART: External restraints, comparative analysis

LibG: Nucleic acid restraints

LORESTR: Automated low-resolution pipeline

AceDRG: Ligand dictionary and conformer generation

Coot: Visualisation & manipulation of restraints, map blurring
…also morphing, jiggle-fit, backrub rotamers…

Many tools are applicable to cryo-EM as well as MX

What and When

Early stages (e.g. straight after MR)

- Rigid body refinement
- Jelly body – sometimes up to 200 cycles

Medium stages – during model building

- Auto local NCS – wherever possible
- External restraints (40 cycles) – homologue available
- Otherwise, jelly body… but not together
- H-bond and DNA/RNA restraints – no homologue available
- Secondary structure conformation restraints – model building tool
- Add hydrogens (?)

Medium-final stages

- TLS – at medium resolutions
- Anisotropic B-factors – only at high resolution
- Twin refinement – only if you are sure

Final stages of refinement

Jelly body – around 20 cycles

Relevant Publications

Low-resolution refinement with REFMAC5, ProSMART, LibG & LORESTR:

Nicholls *et al.* (2017)) Low Resolution Refinement of Atomic Models Against Crystallographic Data. *Protein Crystallography*, 565-93.

Nicholls *et al.* (2013) Recent Advances in Low Resolution Refinement Tools in REFMAC5. *Adv. Methods for Bio. Xtallography*, 231-58.

Nicholls *et al.* (2012) Low Resolution Refinement Tools in REFMAC5. *Acta Cryst.* D68, 404-17.

Tools for cryo-EM model fitting & refinement:

Nicholls *et al.* (2018) Current approaches for the fitting and refinement of atomic models into cryo-EM maps using CCP-EM. *Acta Cryst.* D74, 492-505.

Murshudov (2016) Refinement of atomic structures against cryo-EM maps. *Methods in Enzymology*, 277-305.

Brown *et al.* (2015) Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Cryst.* D71, 136-53.

Tools for ligand fitting & validation:

Nicholls (2017) Ligand fitting with CCP4. *Acta Cryst.* D73, 158-170.

Emsley (2017) Tools for ligand validation in Coot. *Acta Cryst.* D73, 203-10.

Debreczeni & Emsley (2012) Handling ligands with Coot. *Acta Cryst.* D68, 425-30.

Cooperative utilisation of information from Xtal and NMR:

Kovalevskiy et al. (2018) Overview of refinement procedures within REFMAC5: Utilising Data from Different Sources. *Acta Cryst*. D74, 215-27.

Carlon et al. (2016) How to tackle protein structural data from solution and solid state: An integrated approach. *Progress in nuclear magnetic resonance spectroscopy*. 92, 54-70.

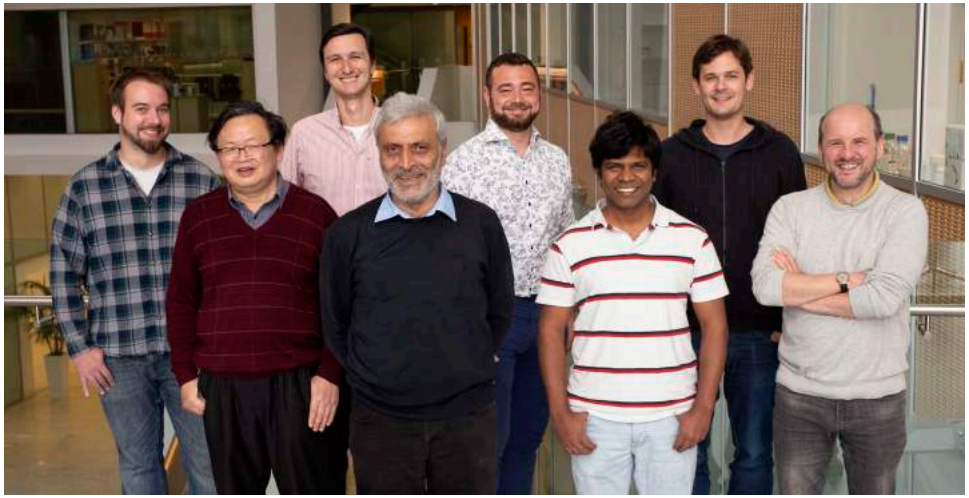
Effect of Twinning on R-factors:

Murshudov GN (2011) Some properties of Crystallographic Reliability Index – Rfactor: Effect of Twinning. *Appl. & Comp. Math*., 10, 250-61.

Acknowledgements

Contact: **nicholls@mrc-lmb.cam.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
Keitaro Yamahita

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

Science & Technology Facilities Council