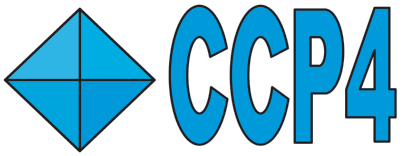
Macromolecular Refinement with CCP4

DLS-CCP4 Data Collection and Structure Solution Workshop

1st December 2024

Rob Nicholls

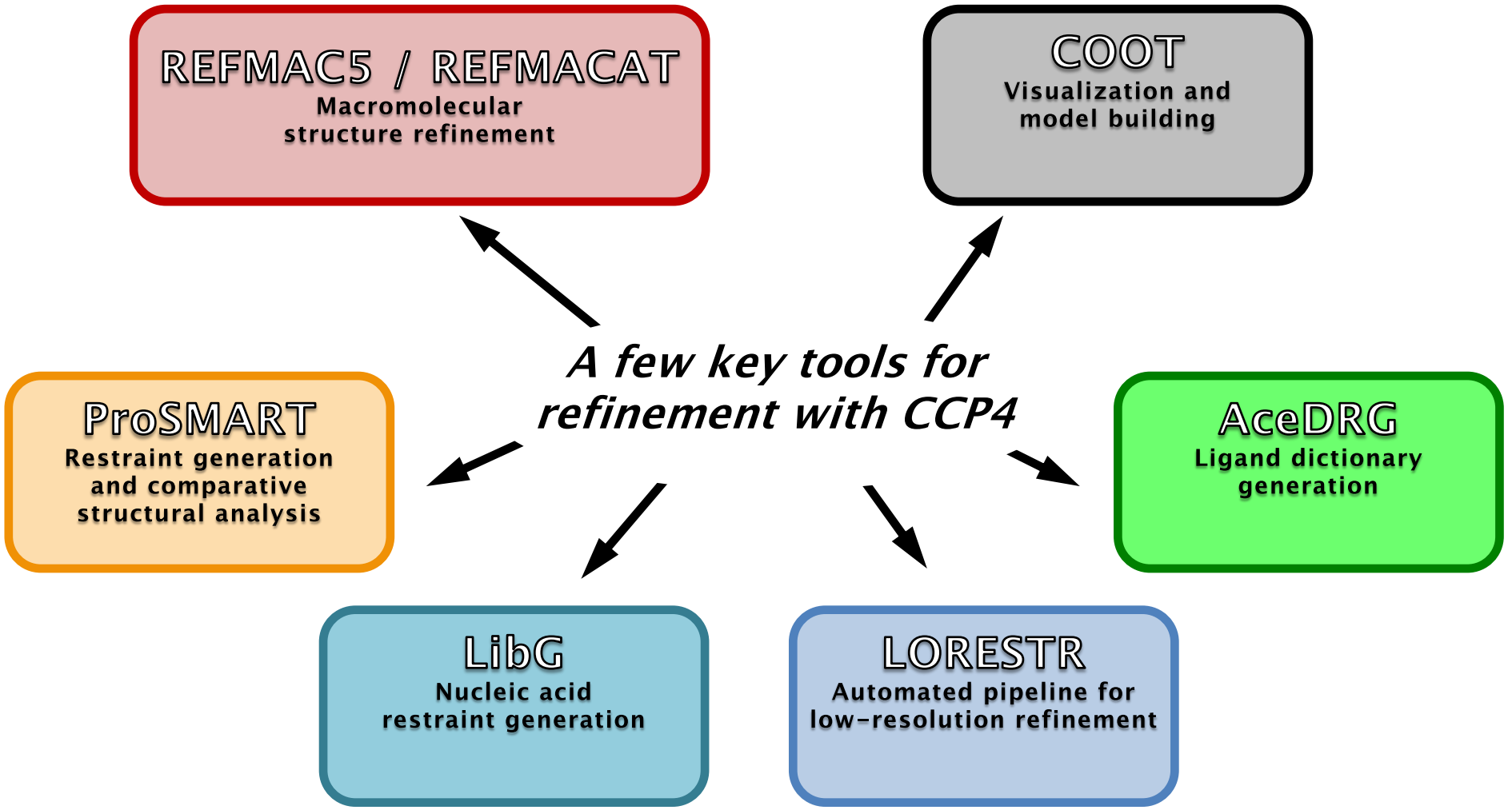
robert.nicholls@stfc.ac.uk



Research Complex at Harwell

MRC Laboratory of Molecular Biology

Science and Technology Facilities Council


REFMAC5 / REFMACAT
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 dictionary generation
LibG
Nucleic acid restraint generation
LORESTR
Automated pipeline for low-resolution refinement

Purpose of Refinement

Crystallographic refinement has two purposes:

1. Fit atomic model into observed X-ray crystallographic data

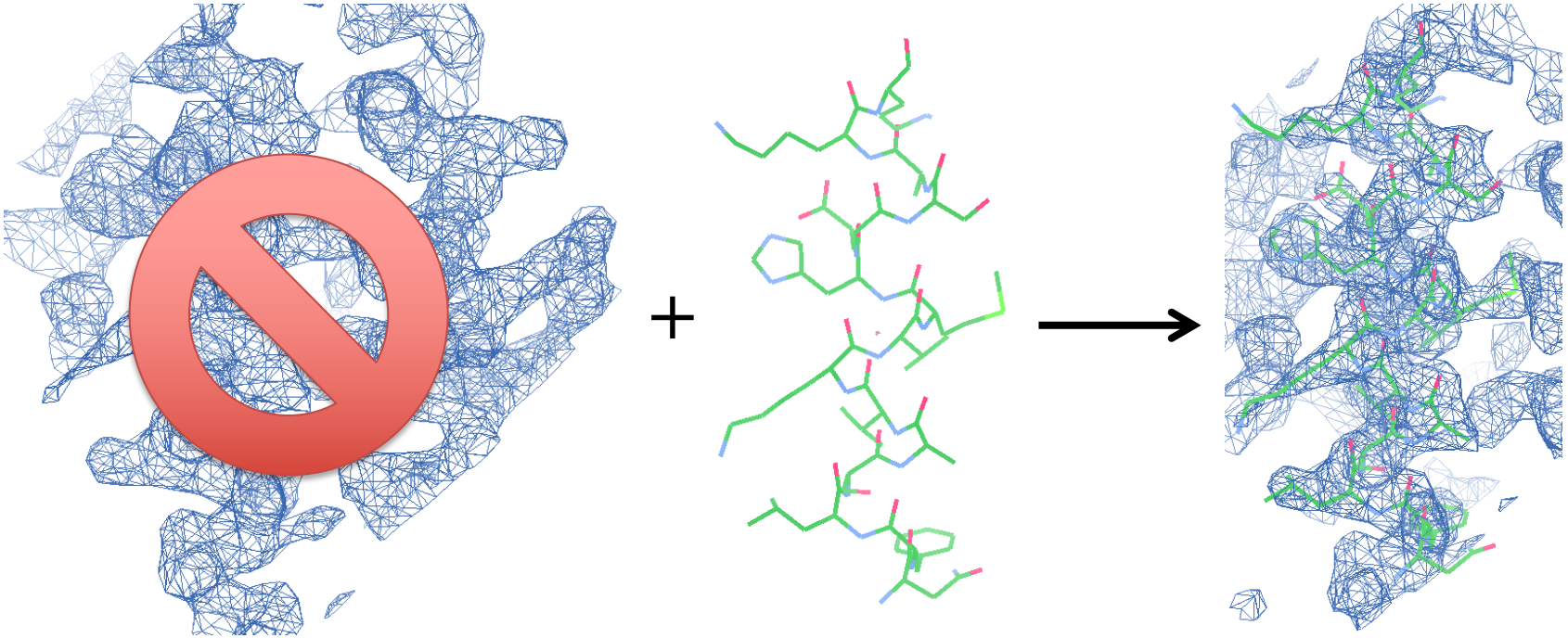
   Model should agree with the observed data

   Model must be chemically and structurally sensible

2. Calculate best possible electron density map

   Allowing the atom model to be visualised, criticised and analysed

Purpose of Refinement



Data

Atomic model

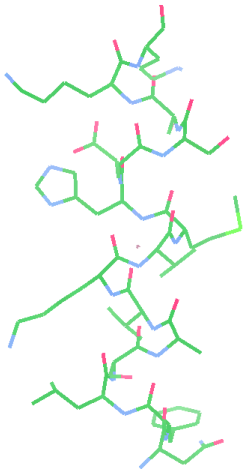
Fit and refine

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 $|F_{calc}|$

So refinement essentially tries to minimise the R-factor:

(note: the exact function to optimize is a bit more complicated)

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

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

How do we know that the model is reliable?

How to validate?

- $\mathbf{R_{free}}$ – 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

Refinement programs output coefficients for (at least) two types of maps:

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

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

Warning:

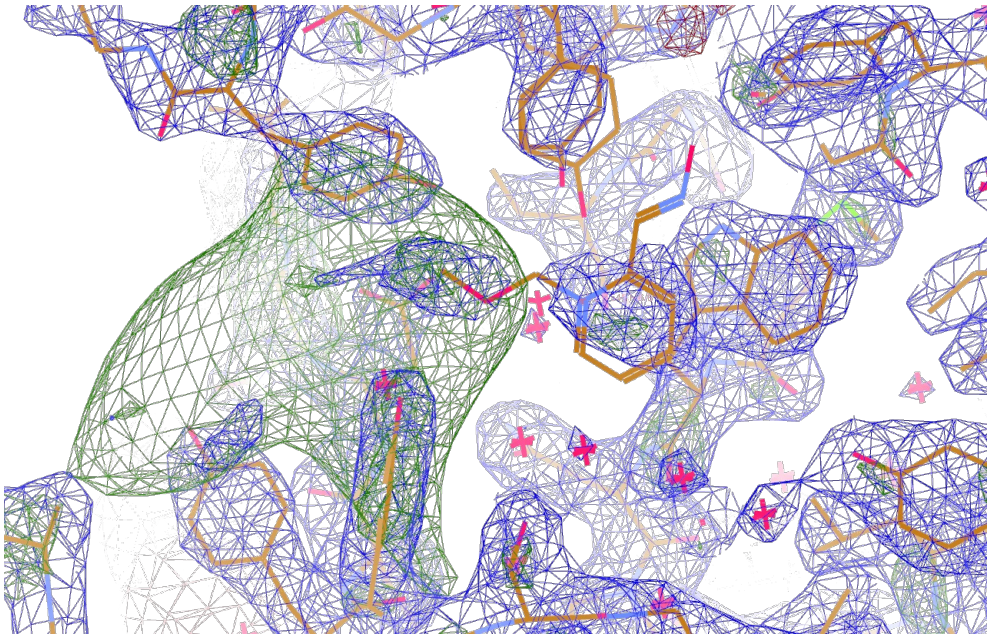
Model updated

↓

φ_{calc} updated

↓

Density changes



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

Model Parameterisation

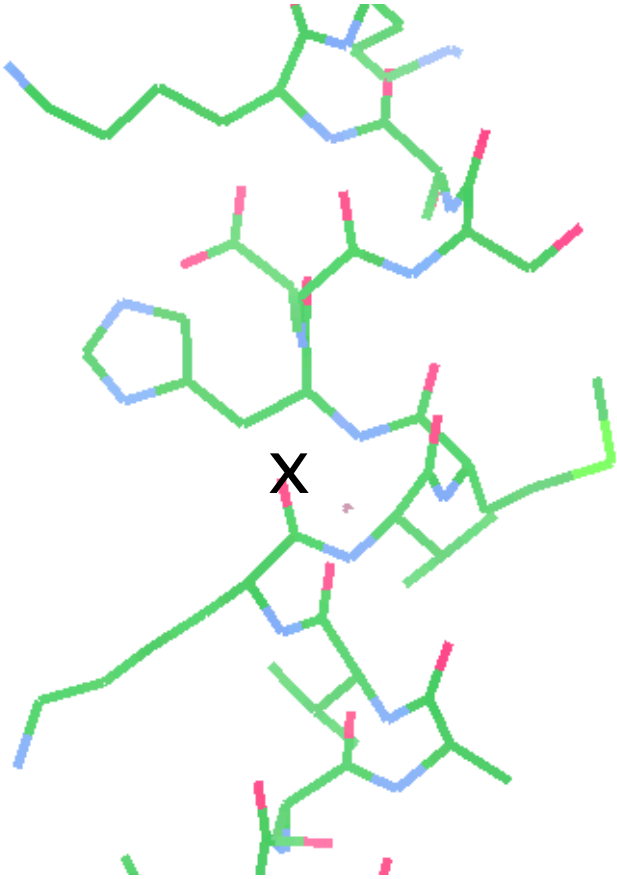
Standard refinable parameters:

Atomic model:

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

Overall parameters (scaling)

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



Model Parameterisation

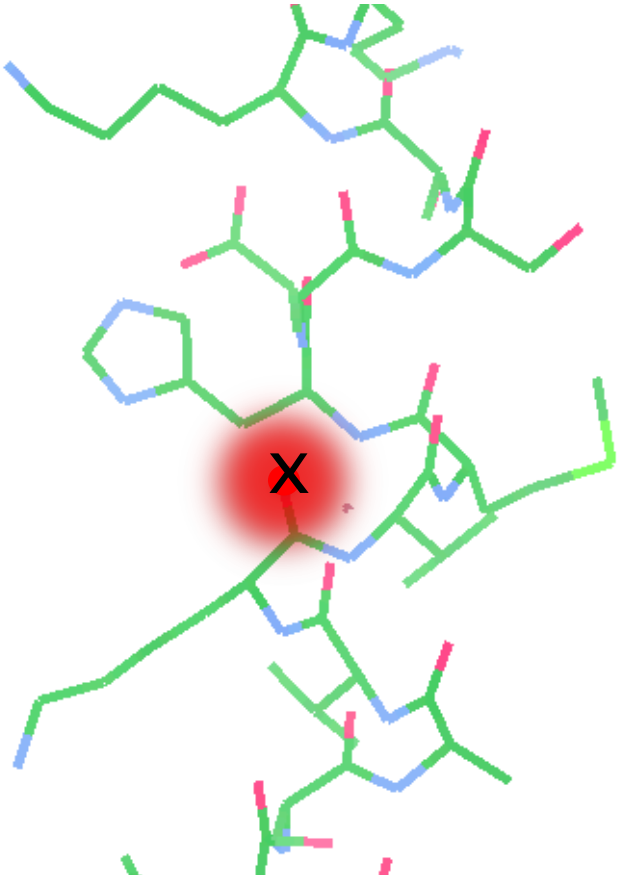
Standard refinable parameters:

Atomic model:

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

Overall parameters (scaling)

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



Model Parameterisation

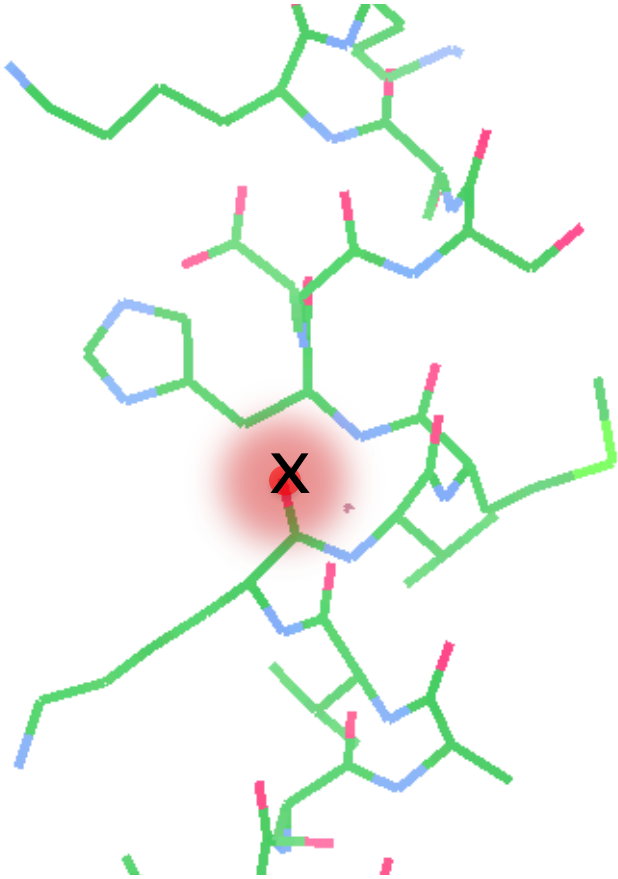
Standard refinable parameters:

Atomic model:

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

Overall parameters (scaling)

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



Model Parameterisation

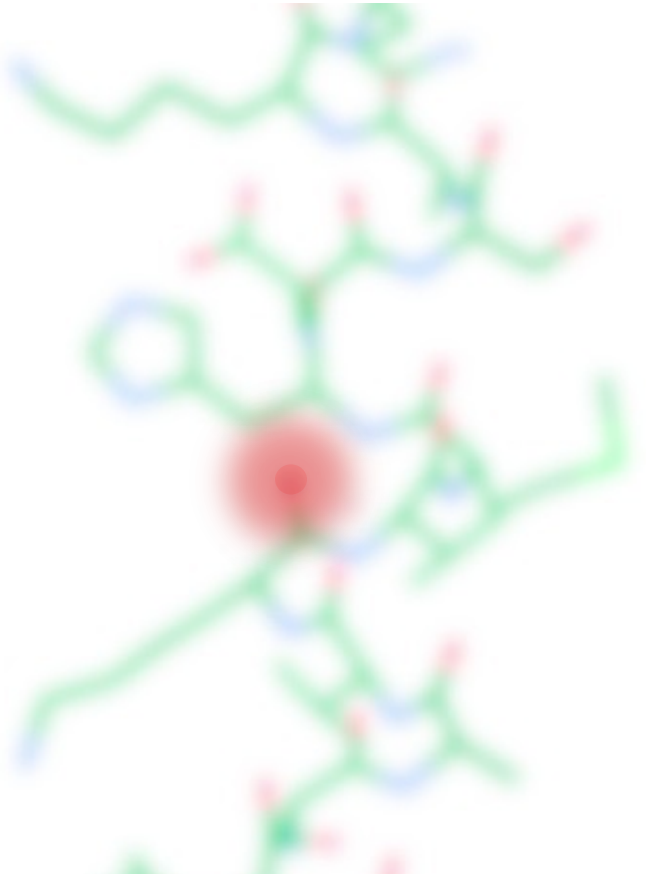
Standard refinable parameters:

Atomic model:

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

Overall parameters (scaling)

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



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 F_{calc} to be on the same scale as F_{obs}
- Allows parameterisation of solvent
    1. Simple scaling / Babinet's bulk solvent model
    2. Explicit solvent mask / no mask

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

Model Refinement

Modern refinement programs use Likelihood-based approaches

Maximum A Posteriori (MAP) target:

likelihood: $P(|F_{obs}|;F_{calc})$

prior

$P(model;obs) \rightarrow max$

$P(model;obs) \propto P(obs;model)\ P(model) \rightarrow max$

$\log[P(obs;model)] + \log[P(model)] \rightarrow max$

$-\log[P(obs;model)] - \log[P(model)] \rightarrow min$

Objective – minimise the negative log-likelihood,

subject to our prior knowledge

Model Refinement

Modern refinement programs use Likelihood-based approaches

Crystallographic target functions have two components:

$$f_{total} = wf_{data} + f_{geometry}$$

likelihood of the data

probability of the model

We have:

- Data – to refine our model against
- Parameters to refine – describing 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
- Chirals
- Planes
- Some torsion angles
- B-values
- VDW repulsions

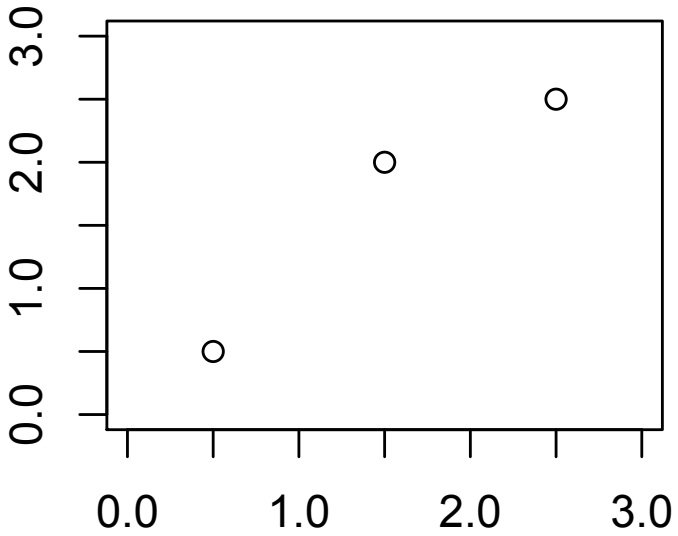
These help to ensure that the model is chemically sensible

Note – we generally deal with restraints, not constraints

Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.

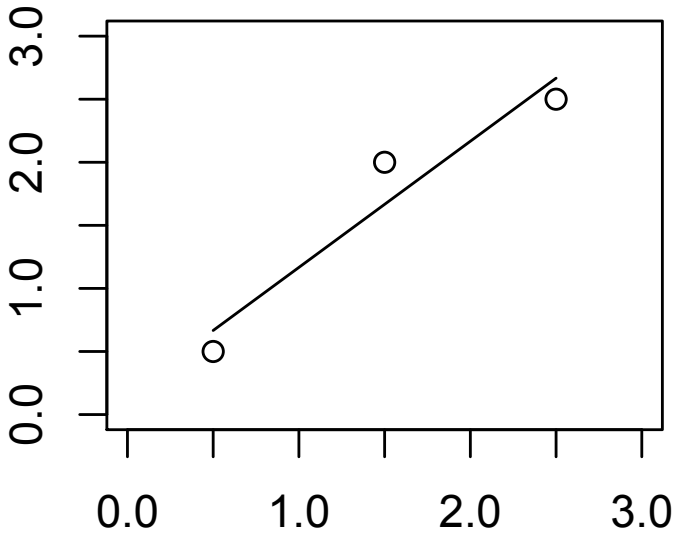


Example: Fitting a line $y = a + bx$

Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.

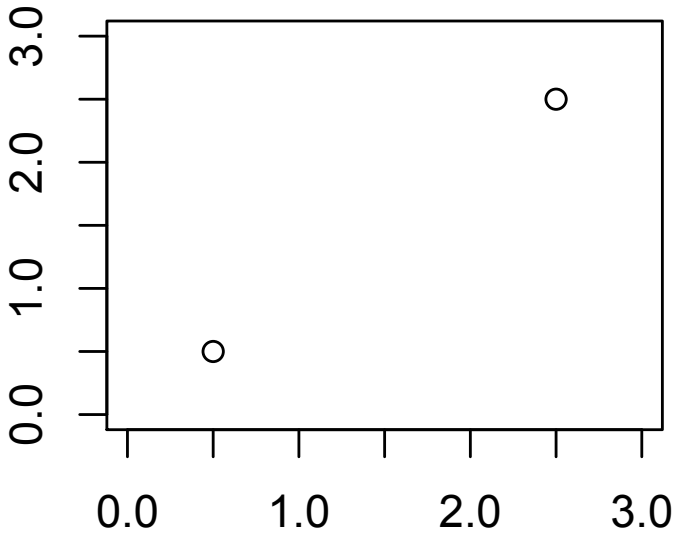


Example: Fitting a line $y = a + bx$

Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.



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

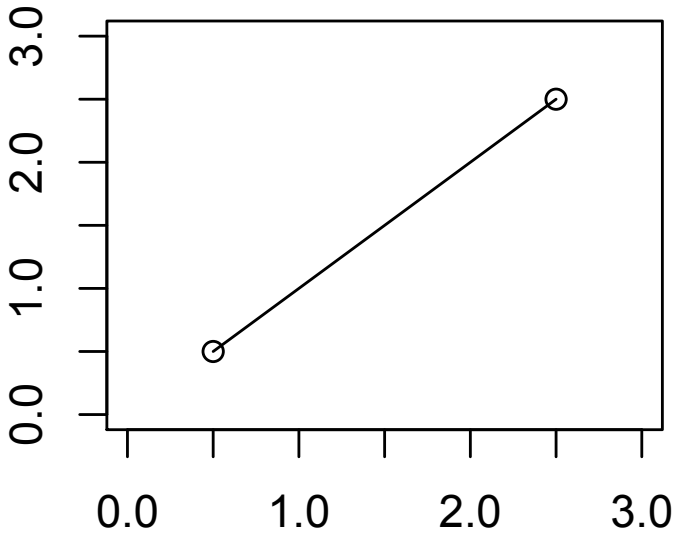
Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.

Can fit a line

Line is unreliable



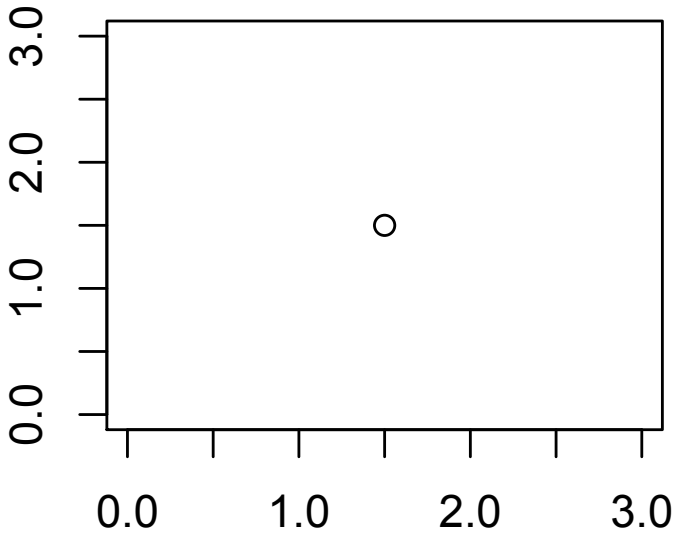
Overfitting
Model Bias

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

Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.



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

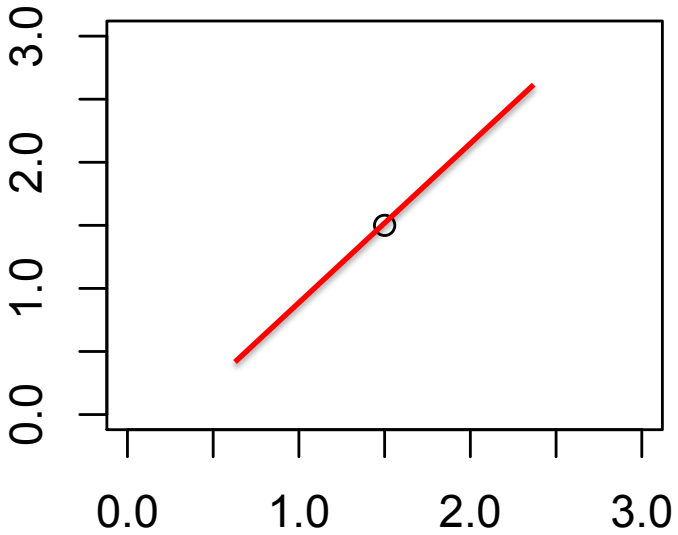
Restraints

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

For each atom:

- Coordinates – 3 parameters
- B-factor – 1 parameter
- Anisotropic U – 5 additional parameters

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

1. Reduce number of parameters

2. Increase number of restraints

Particularly useful at low-resolution:

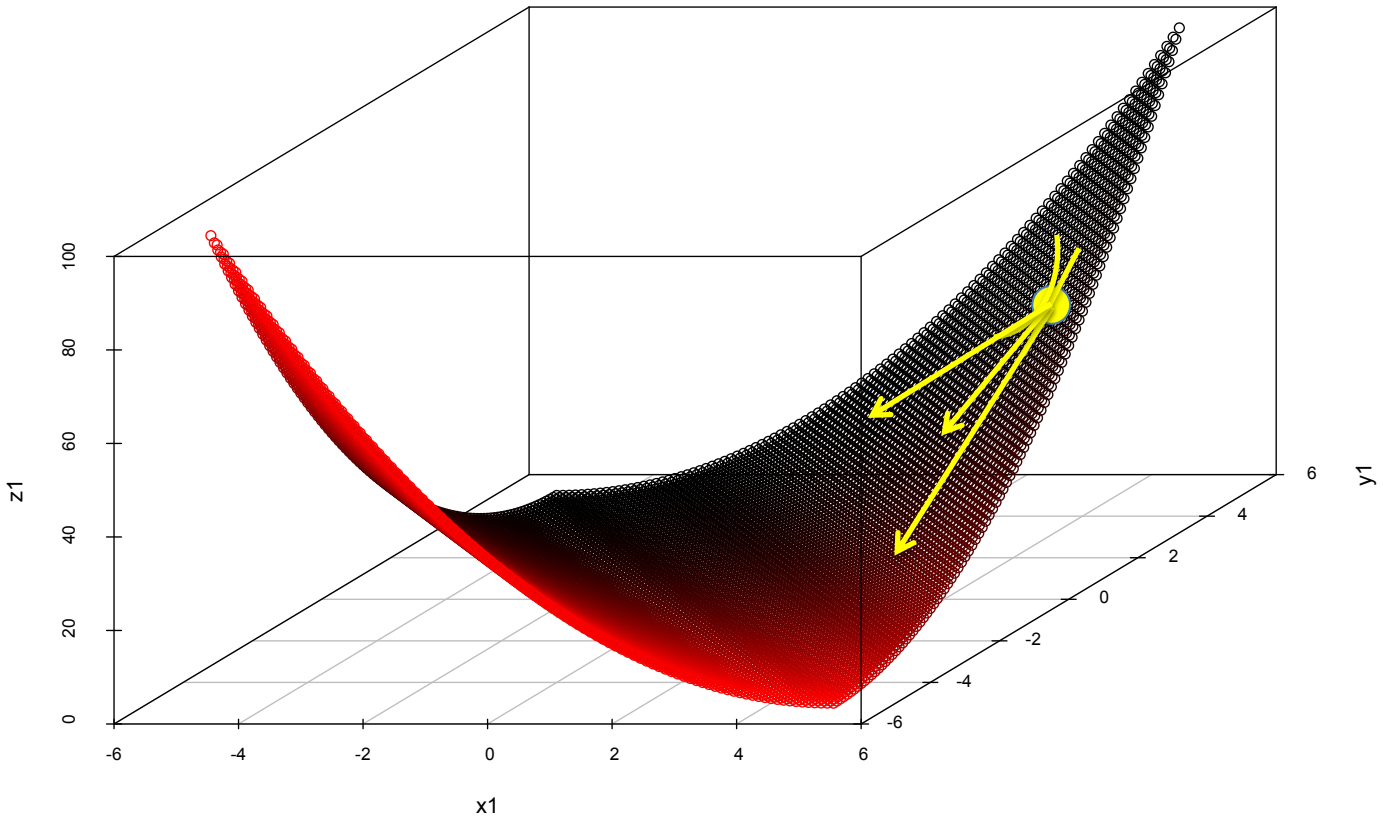
- Reflection intensities often noisy
- Limited data – poor observation:parameter ratio
- Refinement becomes unstable
- Overfitting – R-factors diverge

Increase number of restraints to regularise refinement

Regularisation

Example:

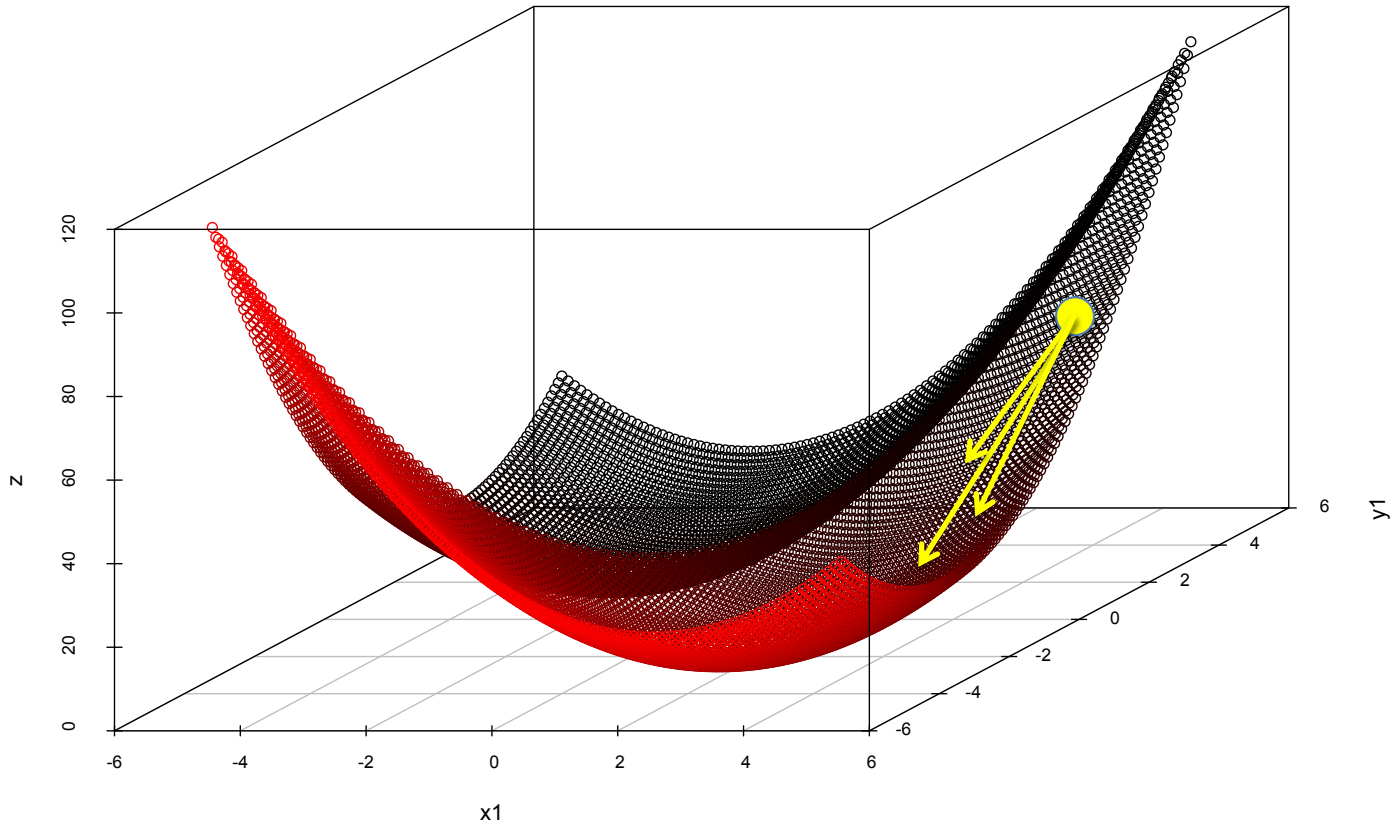
$$z = (x + y)^2$$



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

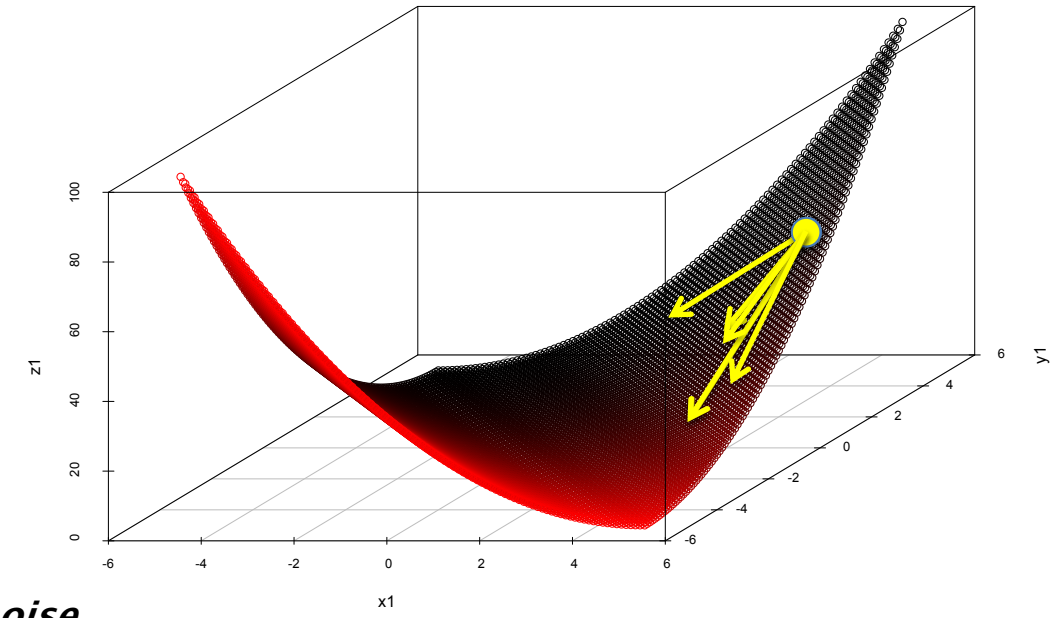
Regularisers without a target:

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

d : interatomic distance

d_{current} : current interatomic distance

σ : restraint standard deviation



Does not change likelihood function.

Does not change derivative.

Does change 2nd derivative – curvature.

Model should be less prone to fitting into noise

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

Typical: σ = 0.01–0.02

Distance threshold: 4.2Å

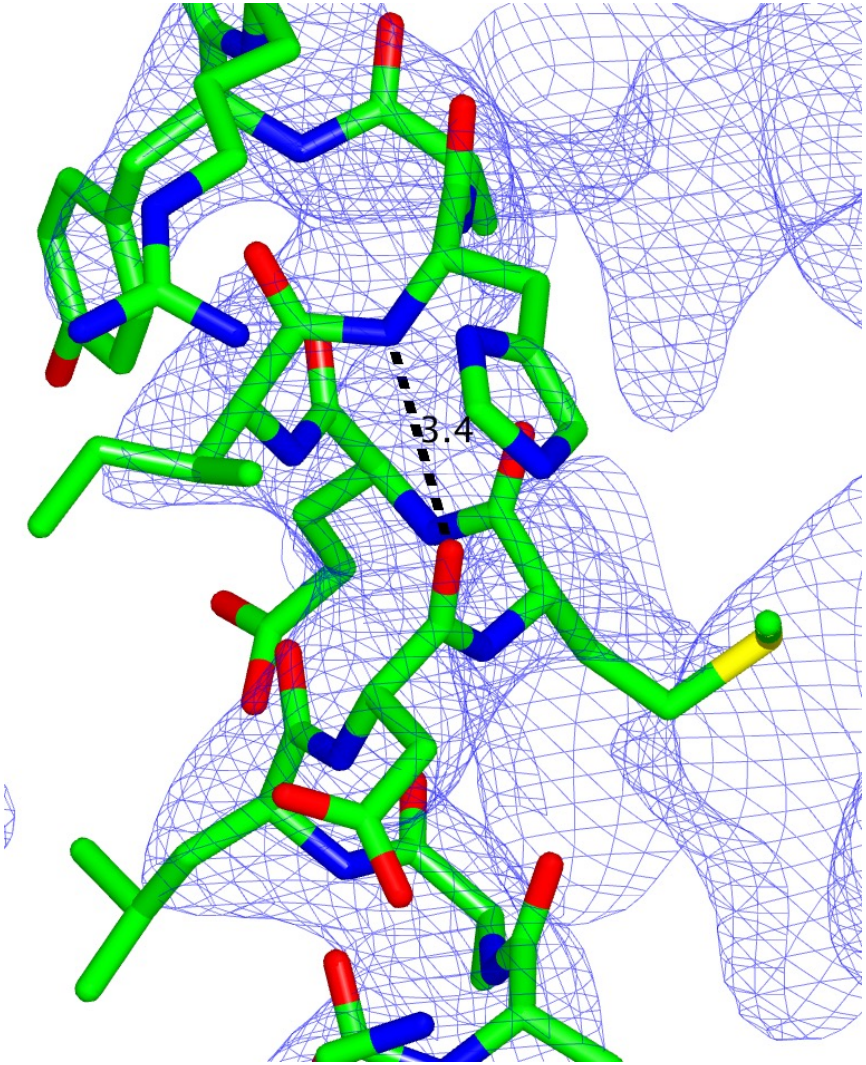
ProSMART

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 – proteins, nucleic acids, ligands, metals, waters
- **Structure analysis**
  - Alignment & comparison - helps analyse differences between models

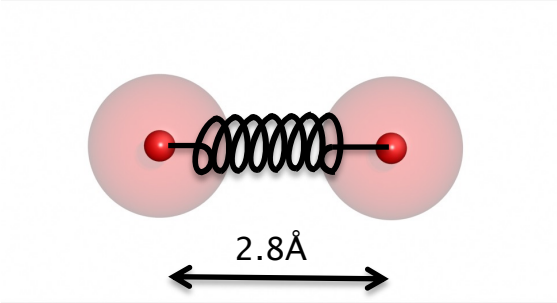
Independent of global conformation

ProSMART External Restraints



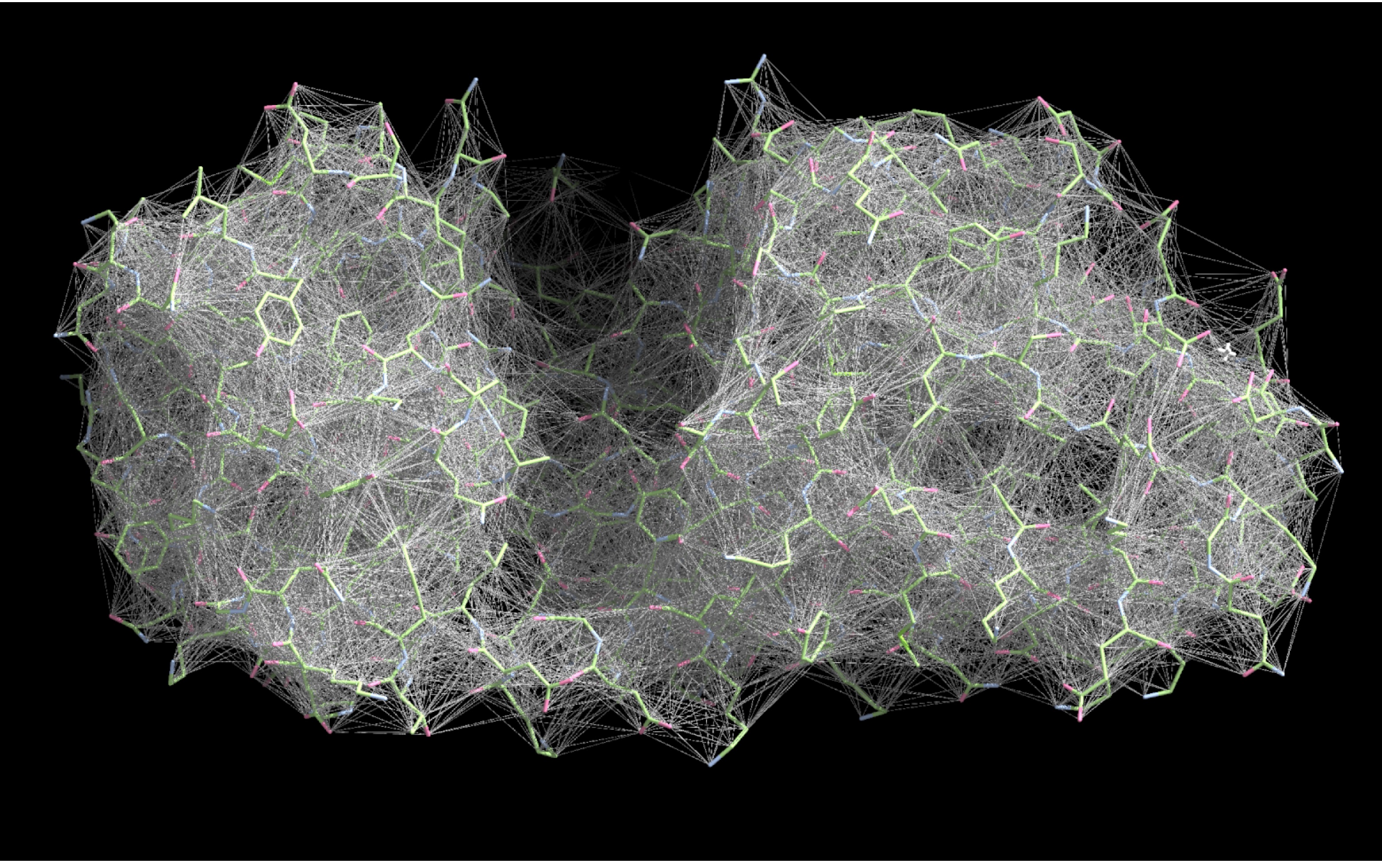
3g4w – 3.7Å

Prior information:

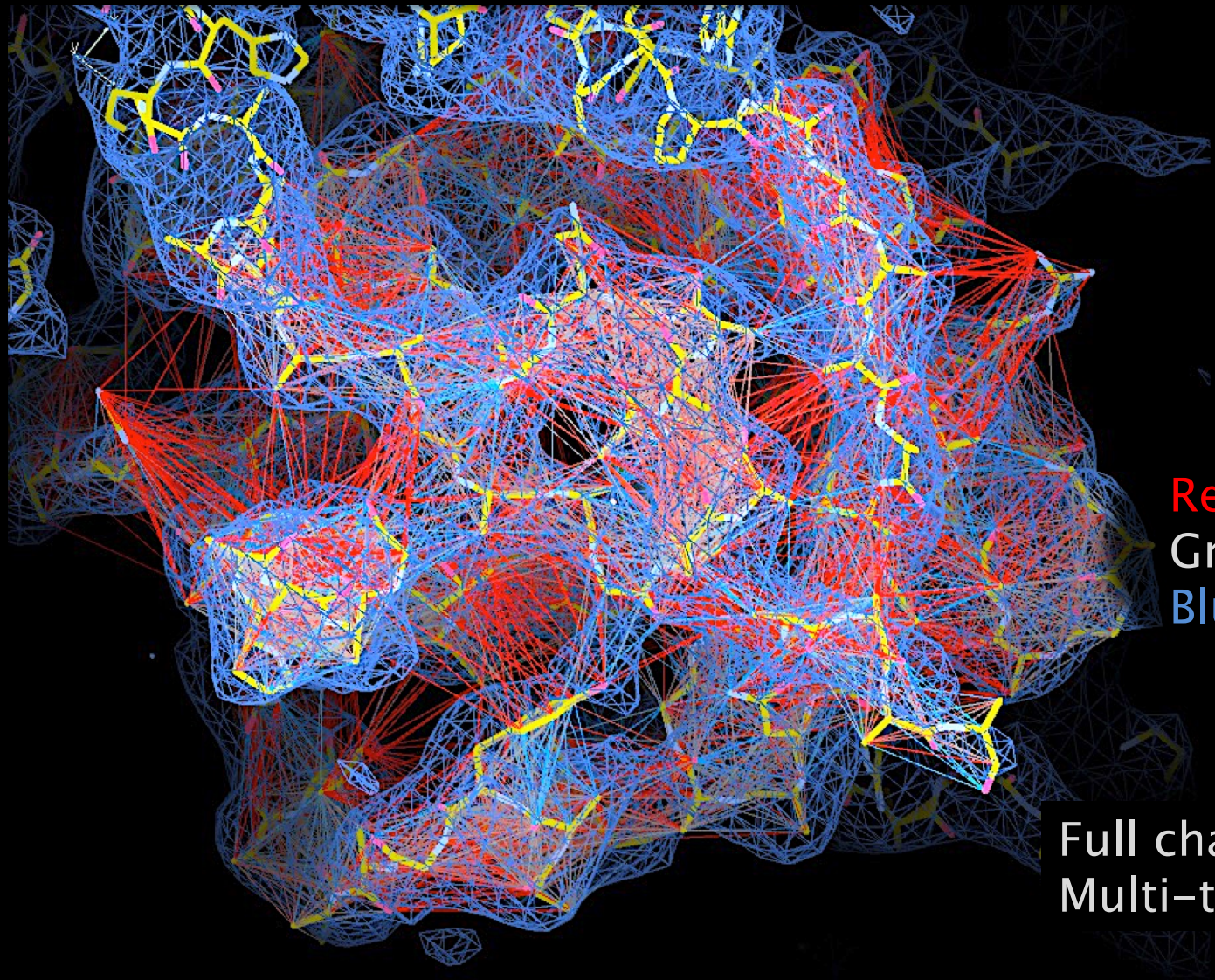


Stabilises structural features

ProSMART External Restraints



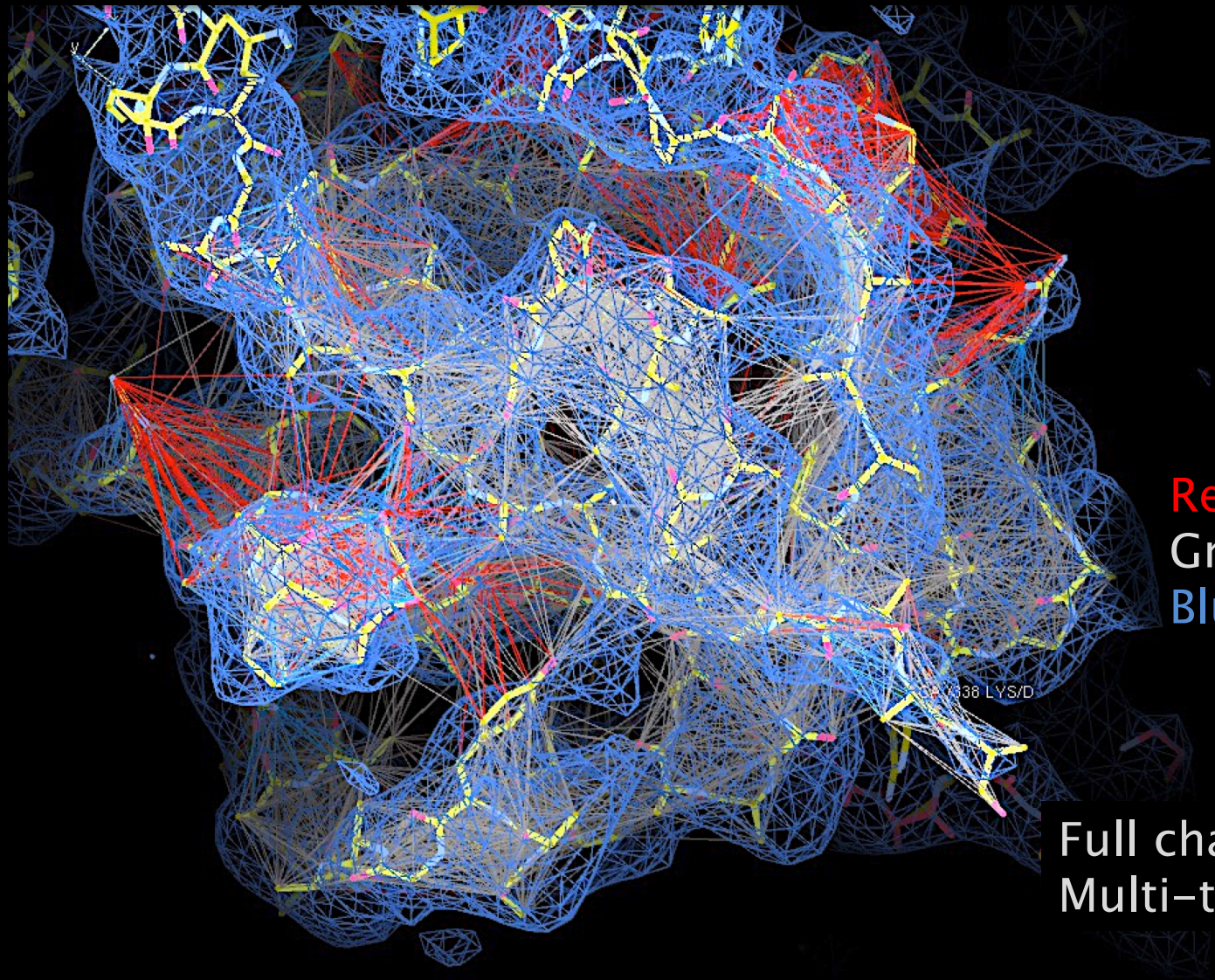
ProSMART Restraints in Coot



Red: long
Grey: similar
Blue: short

Full chain refine
Multi-threaded

ProSMART Restraints in Coot



Red: long
Grey: similar
Blue: short

Full chain refine
Multi-threaded

Example: Ovotransferrin

Ovotransferrin

1ryx – 3.5Å

Low-resolution refinement:

Weak signal
Noisy data

↓

Unstable refinement

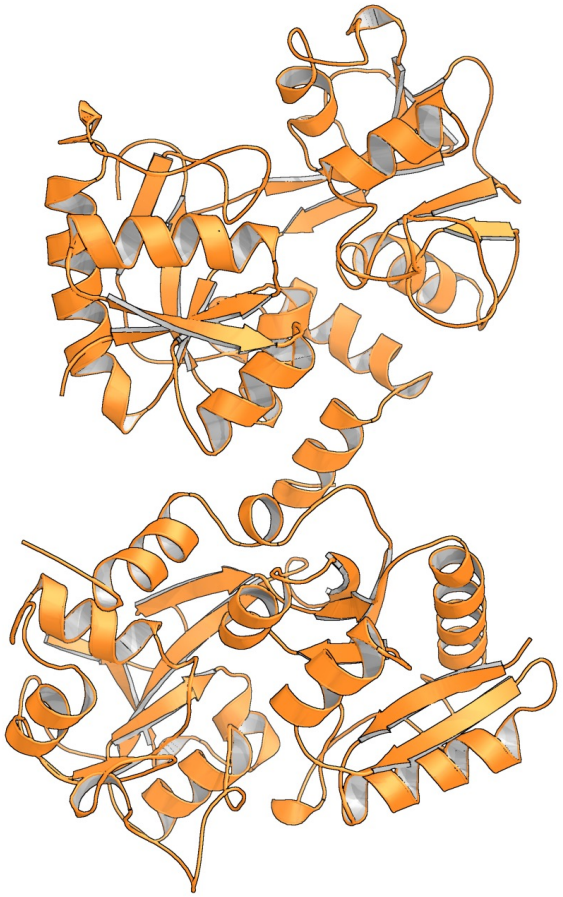
Result:
Poor quality model

Example: Ovotransferrin

Ovotransferrin

1ryx – 3.5Å

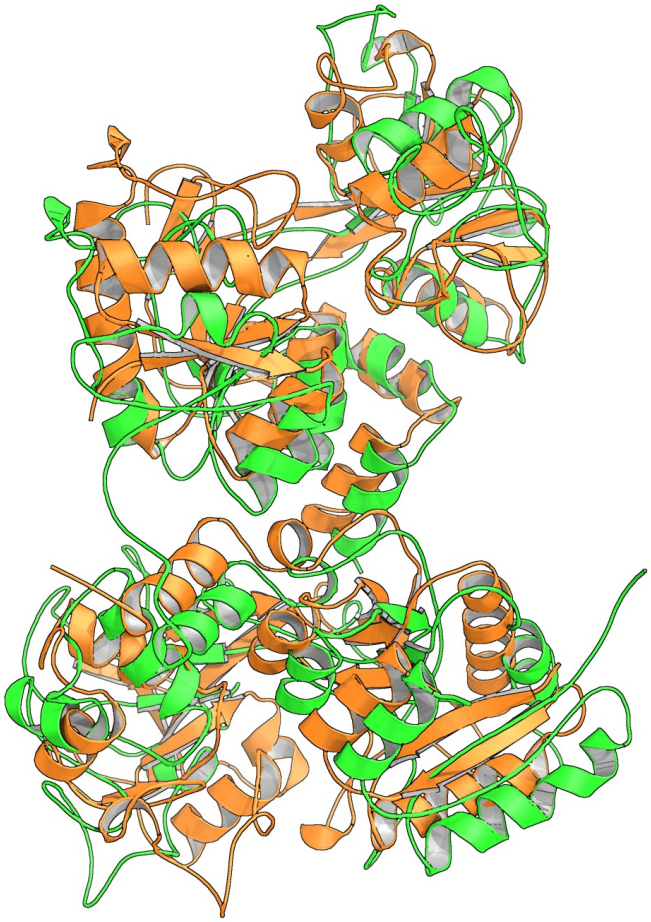
High-resolution homologue



2d3i – 2.15Å

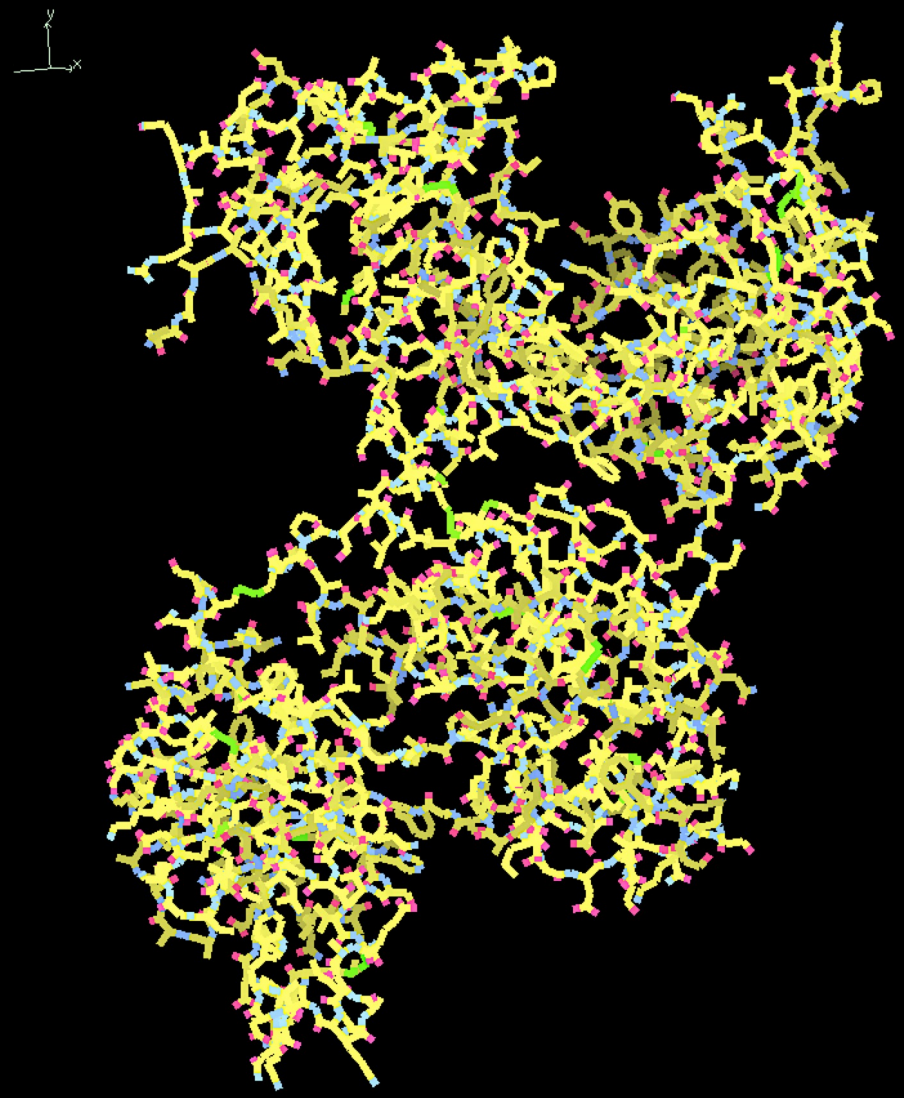
Example: Ovotransferrin

Ovotransferrin



Models don’t superpose well

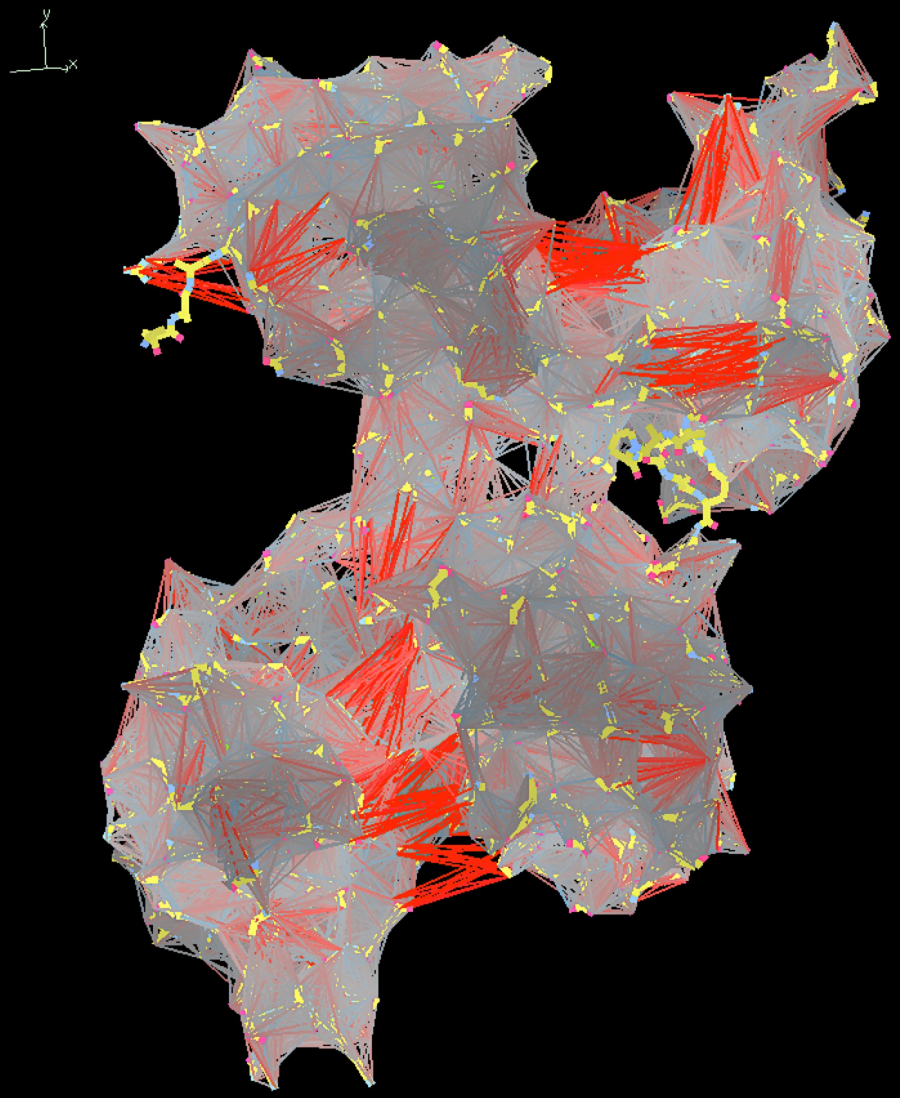
Example: Ovotransferrin



1ryx (3.5Å)

Example: Ovotransferrin

Restraints:
Backbone
Side chains



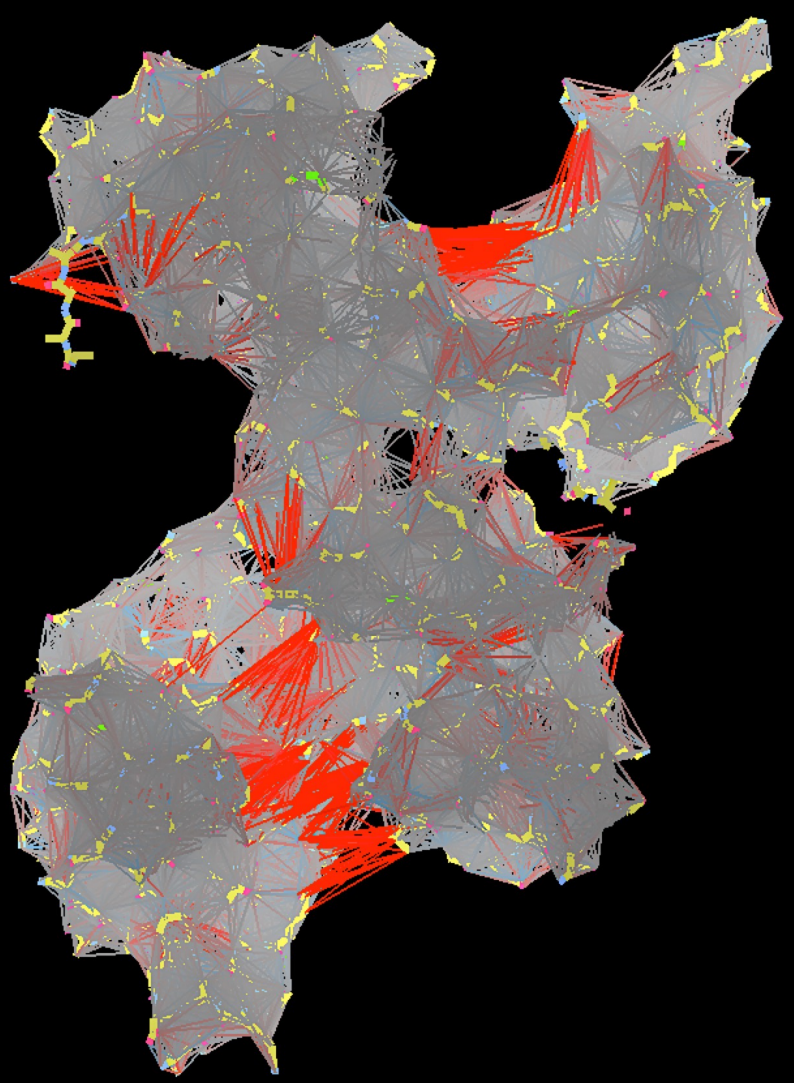
1ryx (3.5Å)
restrained to
2d3i (2.15Å)

Red: long
Grey: similar
Blue: short

Example: Ovotransferrin

Restraints:
- Backbone
- Side chains

After re-refinement

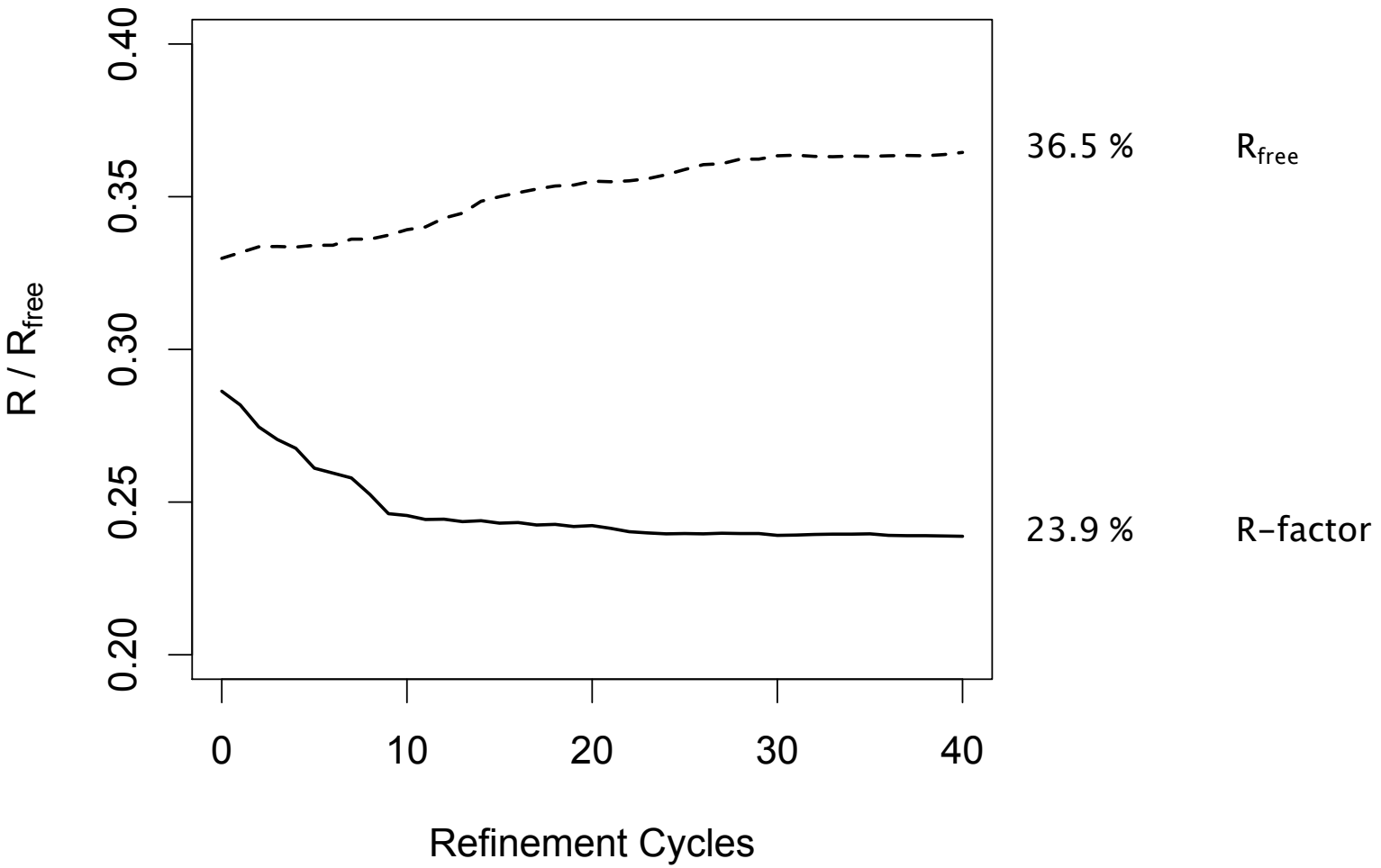


1ryx (3.5Å)
restrained to
2d3i (2.15Å)

Red: long
Grey: similar
Blue: short

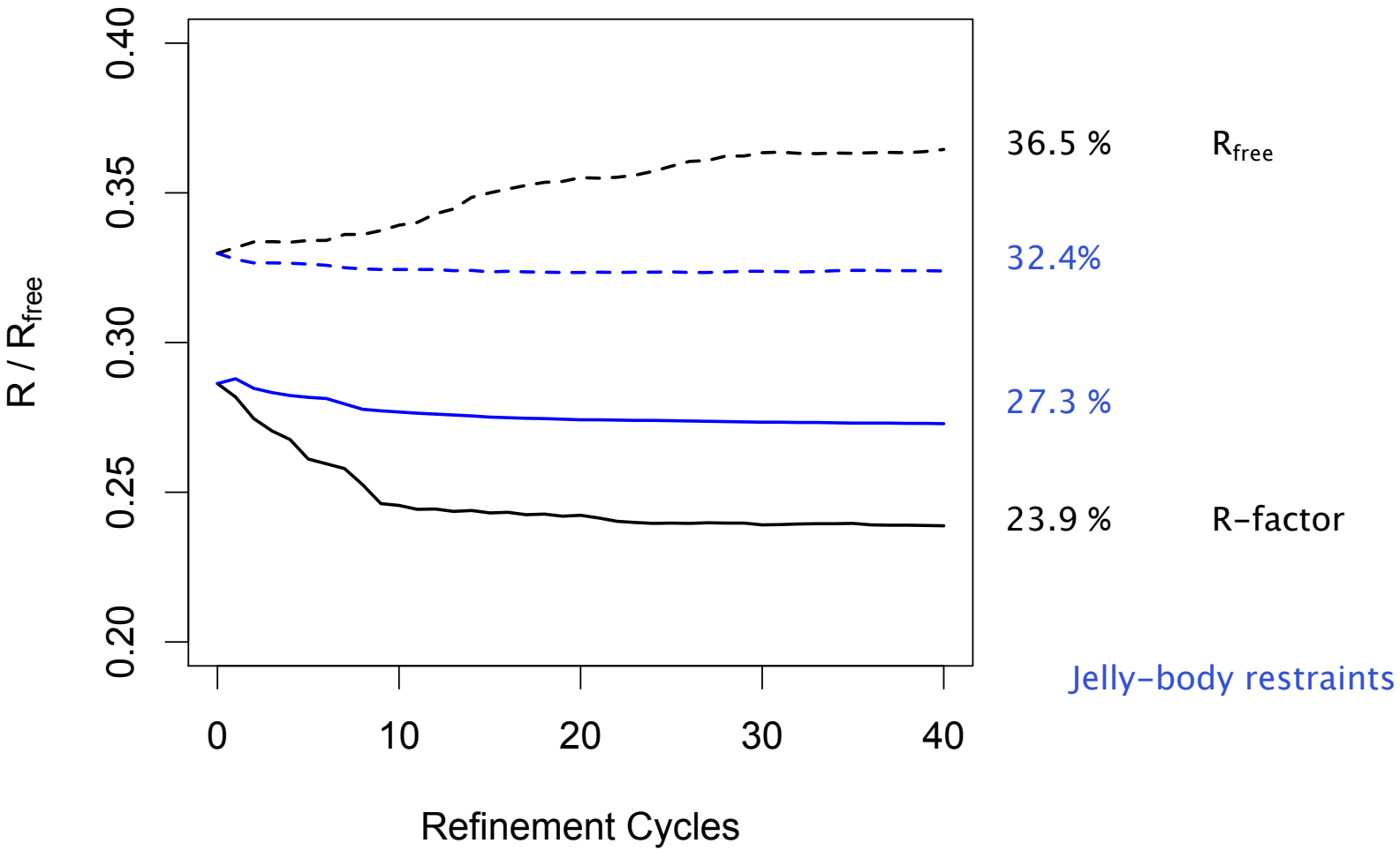
External Restraints

Ovotransferrin



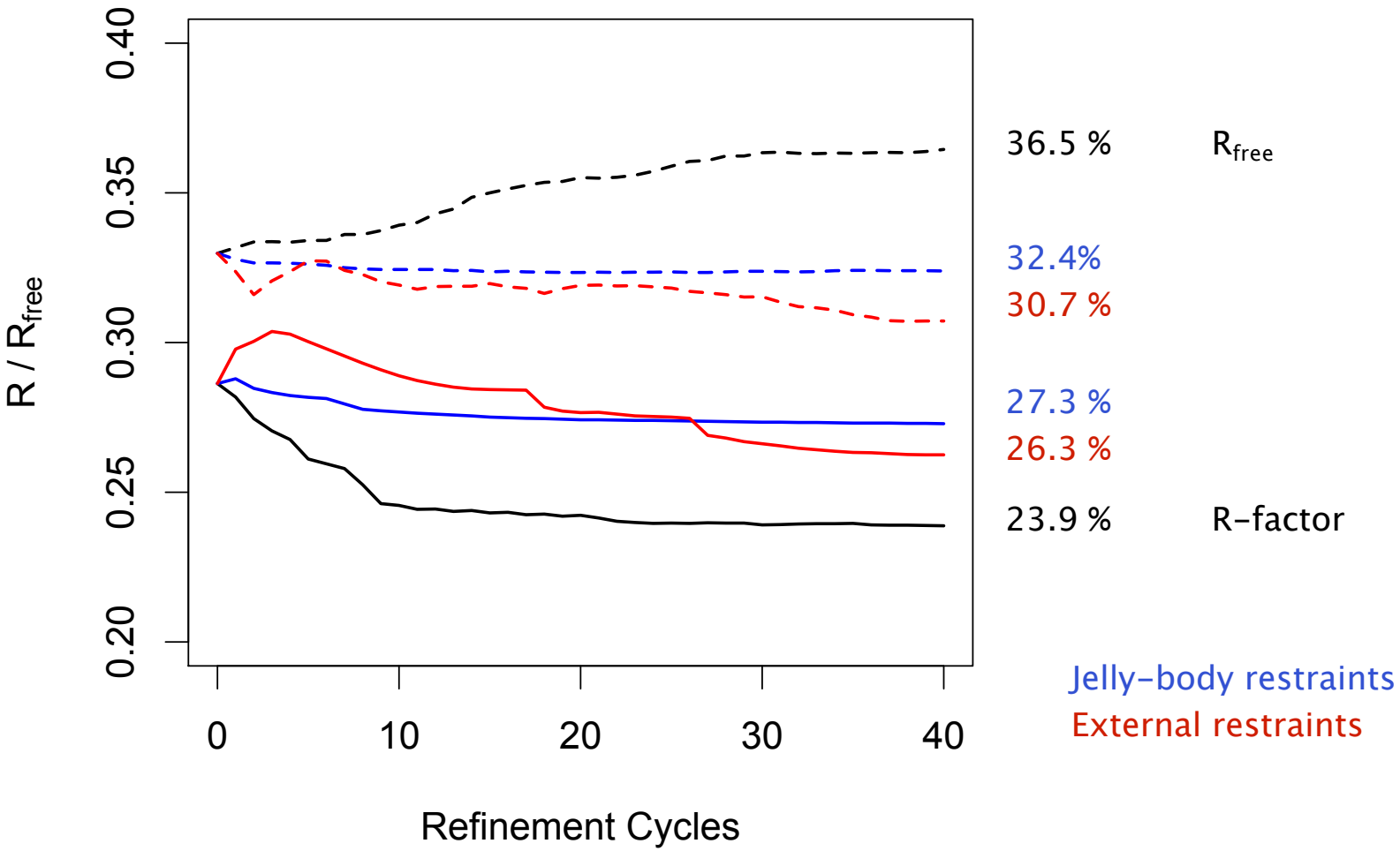
External Restraints

Ovotransferrin



External Restraints

Ovotransferrin

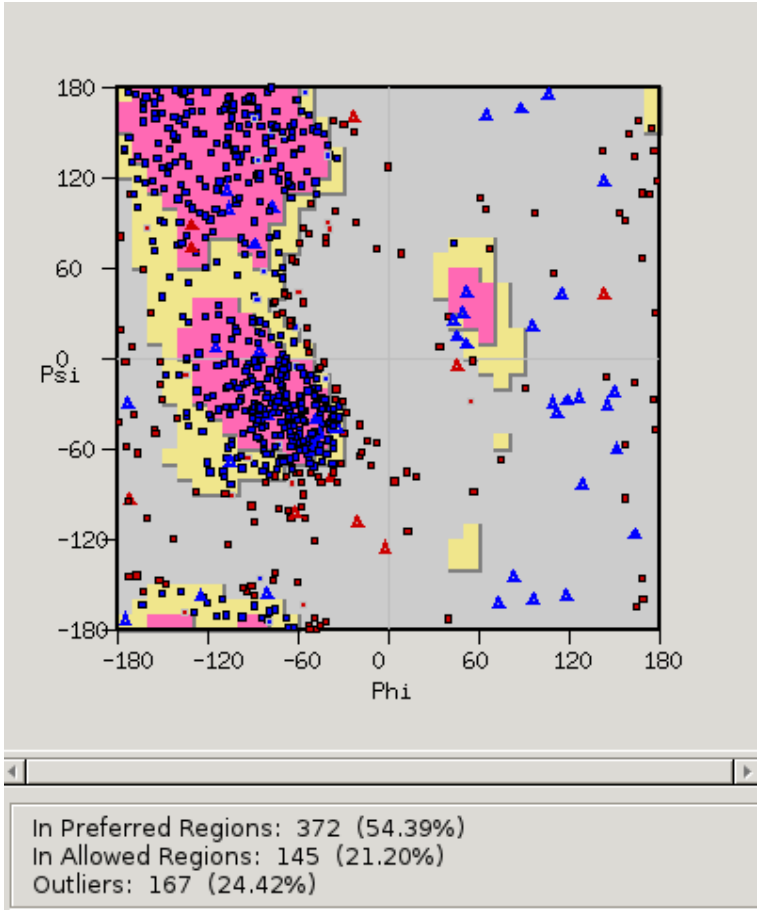


External Restraints

Ovotransferrin

Original Structure

R/R_{free} : 0.286/0.330

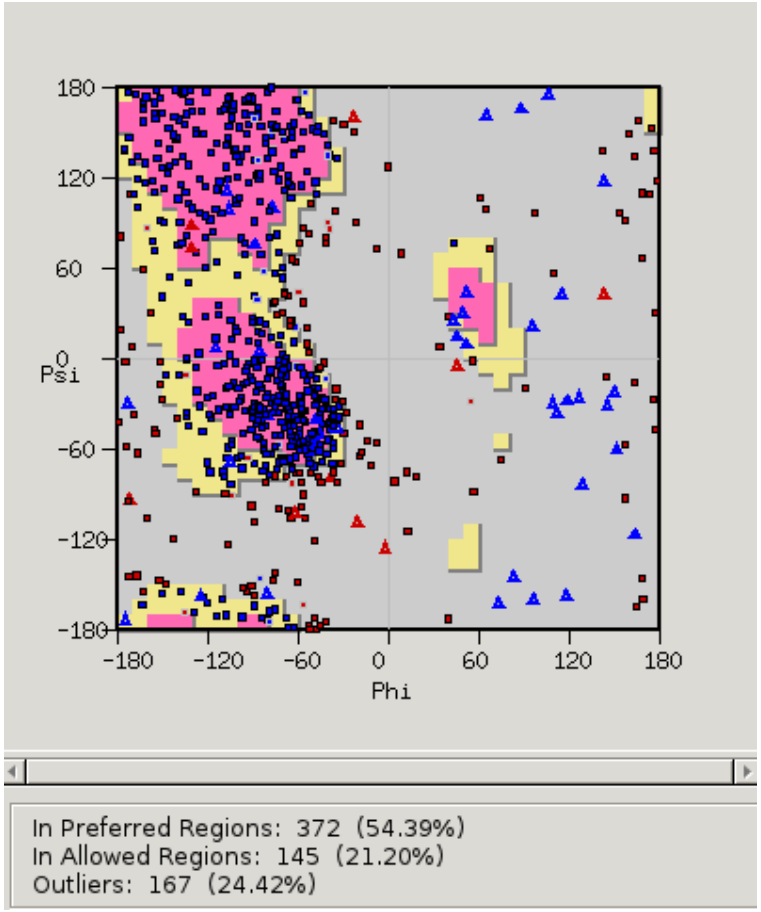


External Restraints

Ovotransferrin

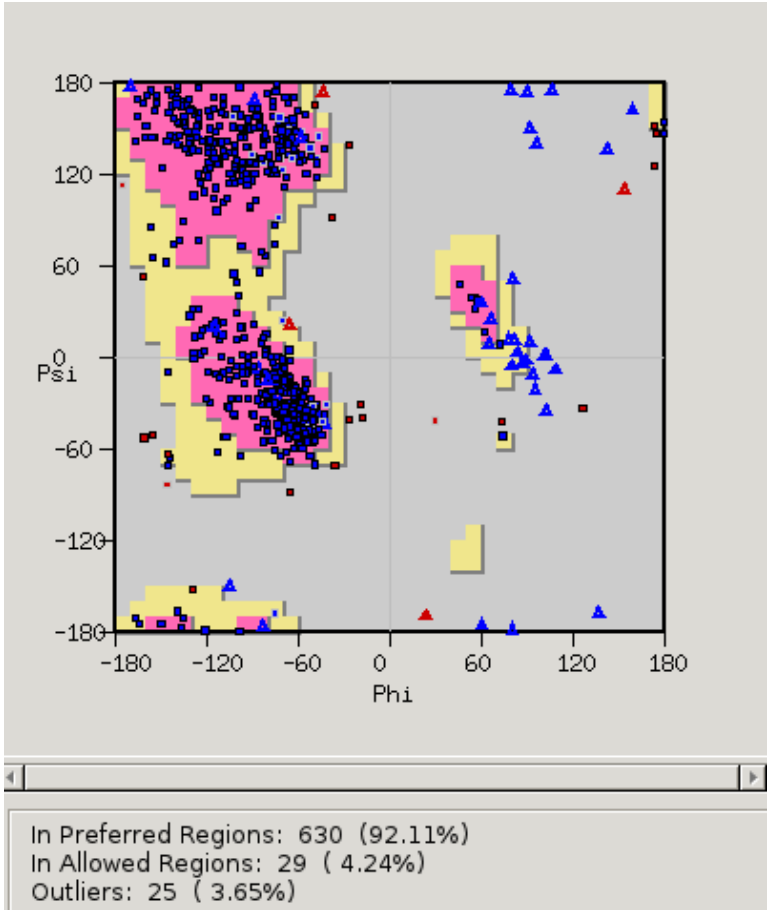
Original Structure

R/R_{free} : 0.286/0.330

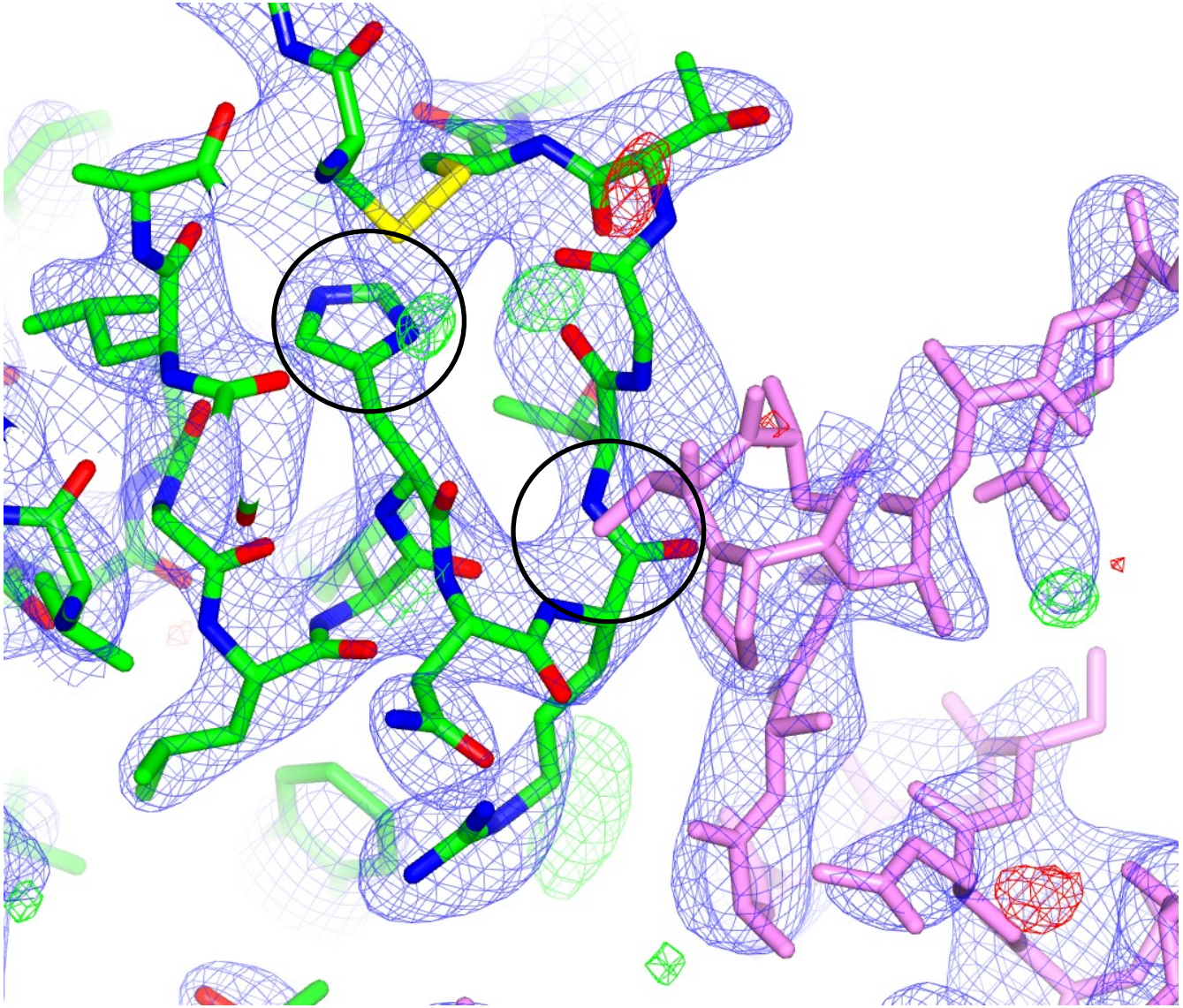


Re-refined with External Restraints

R/R_{free} : 0.263/0.307



External Restraints



1.3σ

Original Structure

R/R_{free} : 0.286/0.330

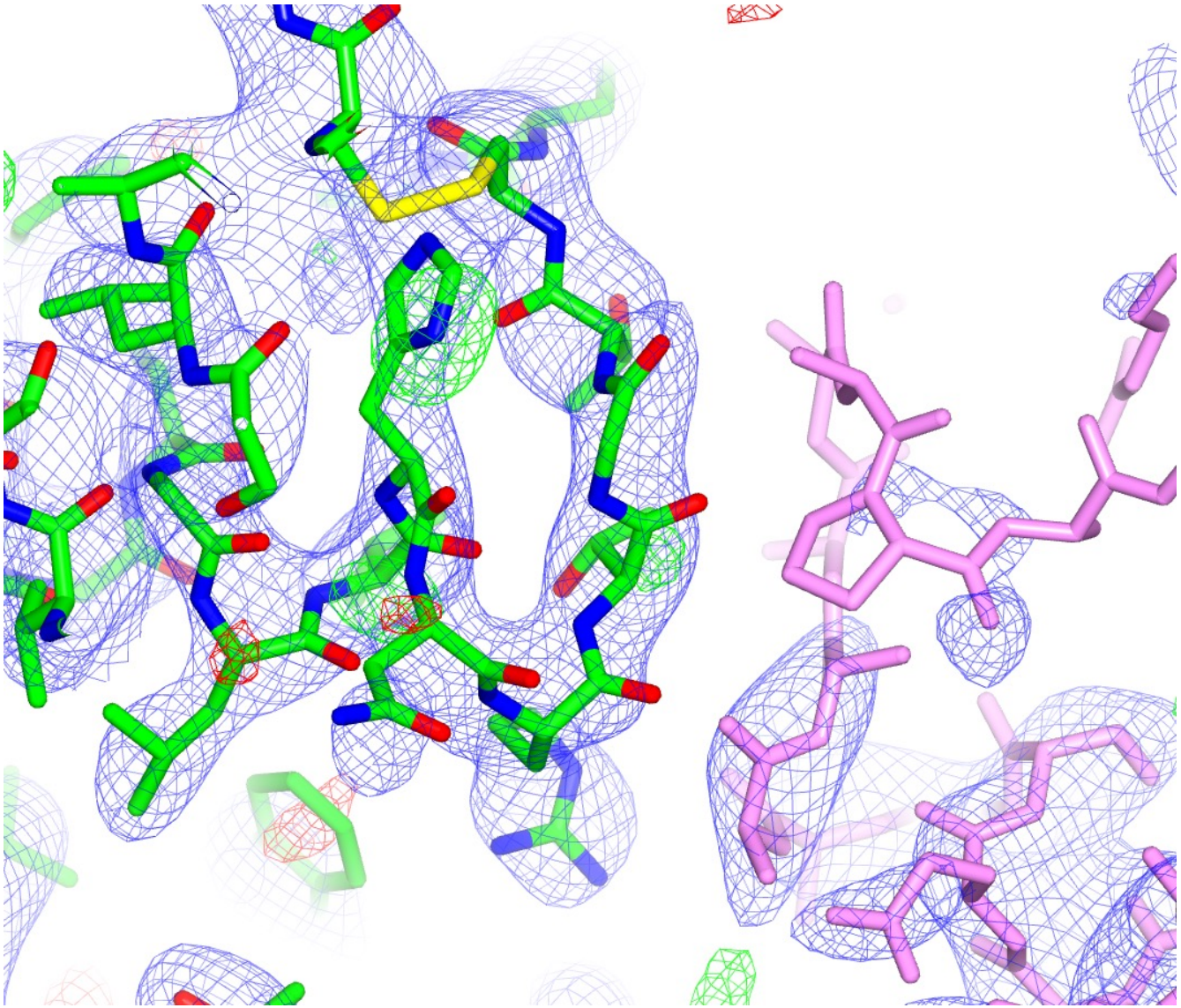
↓

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

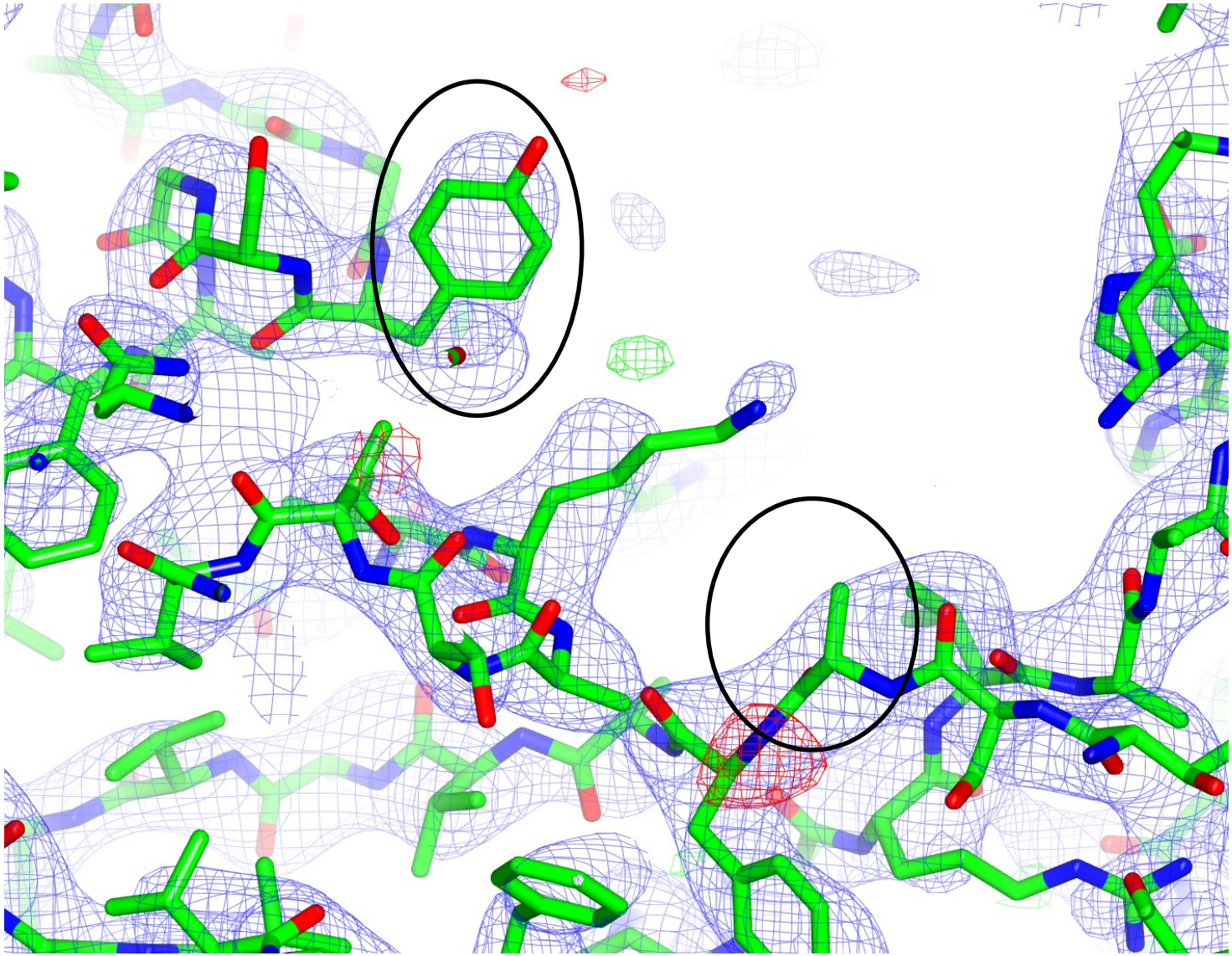
↓

Jelly body

(40 cycles)

R/R_{free} : 0.253/0.304

External Restraints



1.3σ

Original Structure

R/R_{free} : 0.286/0.330

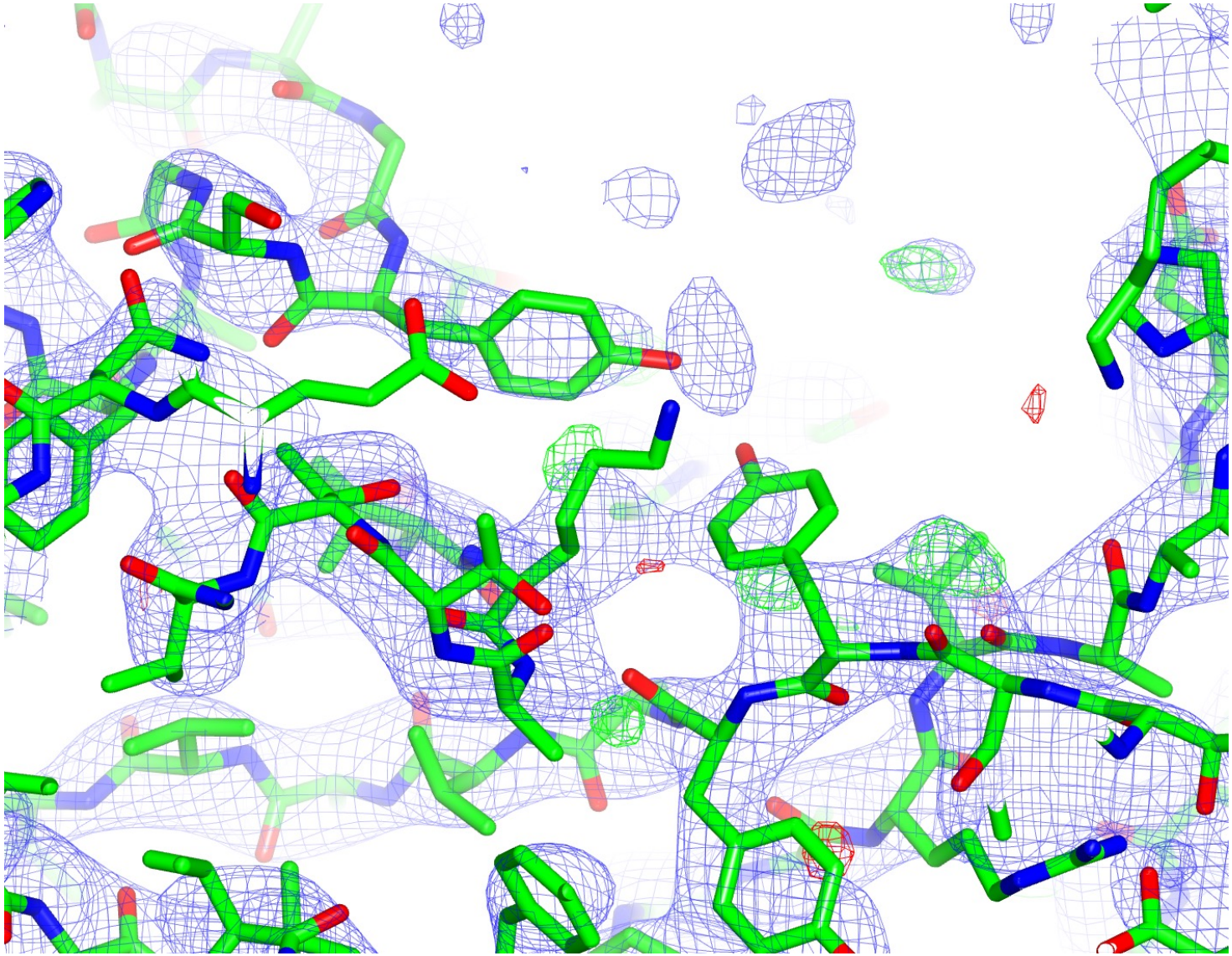
↓

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

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

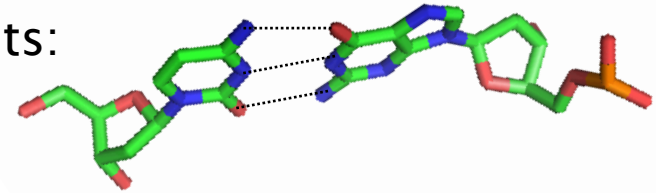
Quality of prior information is important **– consider manual re-refinement**

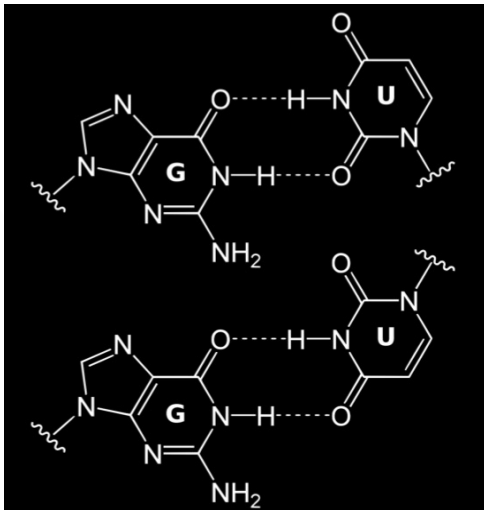
– PDB_REDO is useful


DRINKS

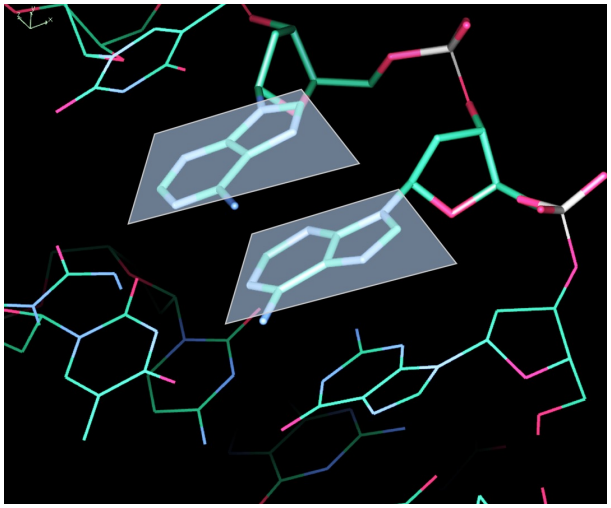
LibG Nucleic Acid Restraints

1) Base-pair restraints:

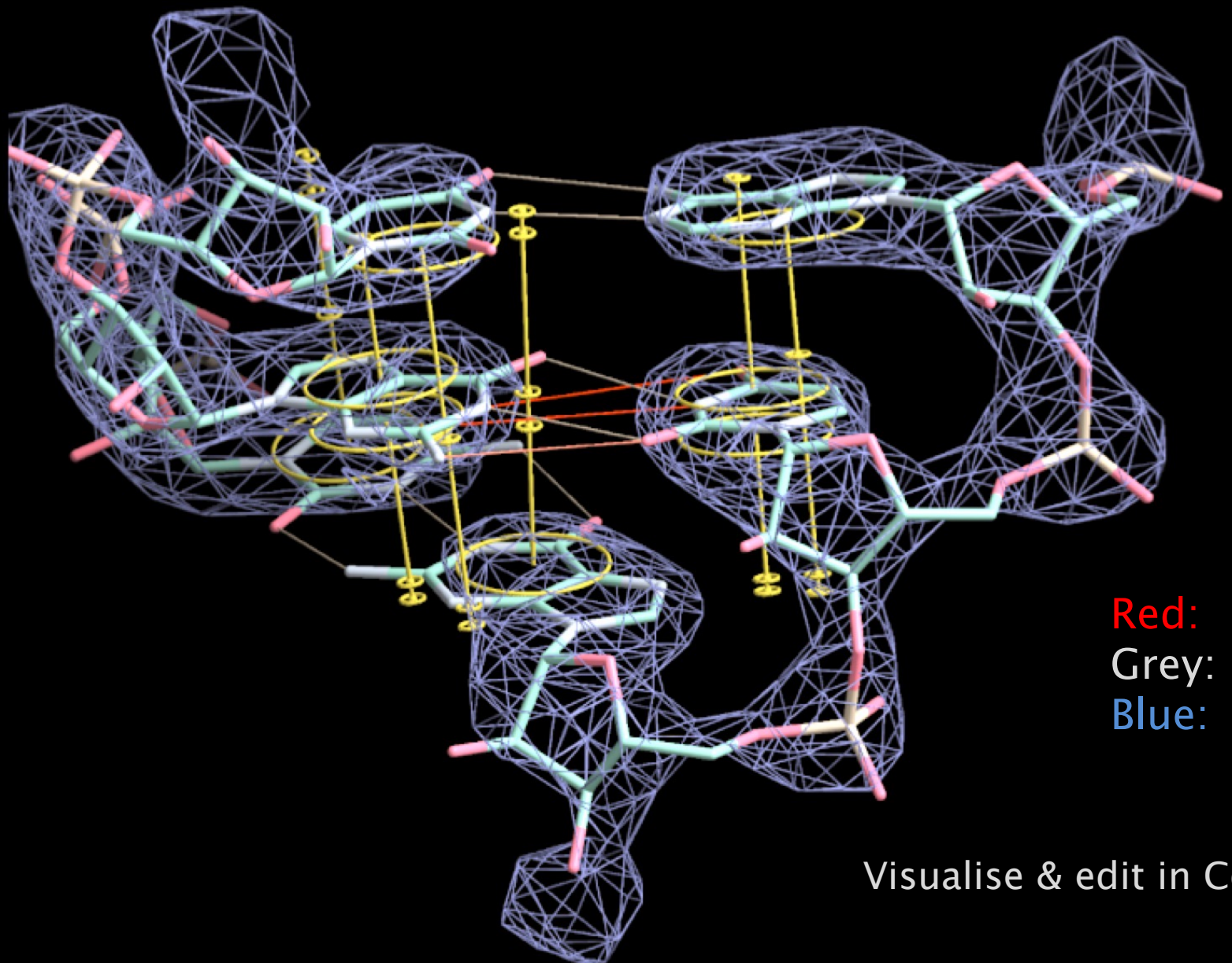




2) Parallel plane restraints:



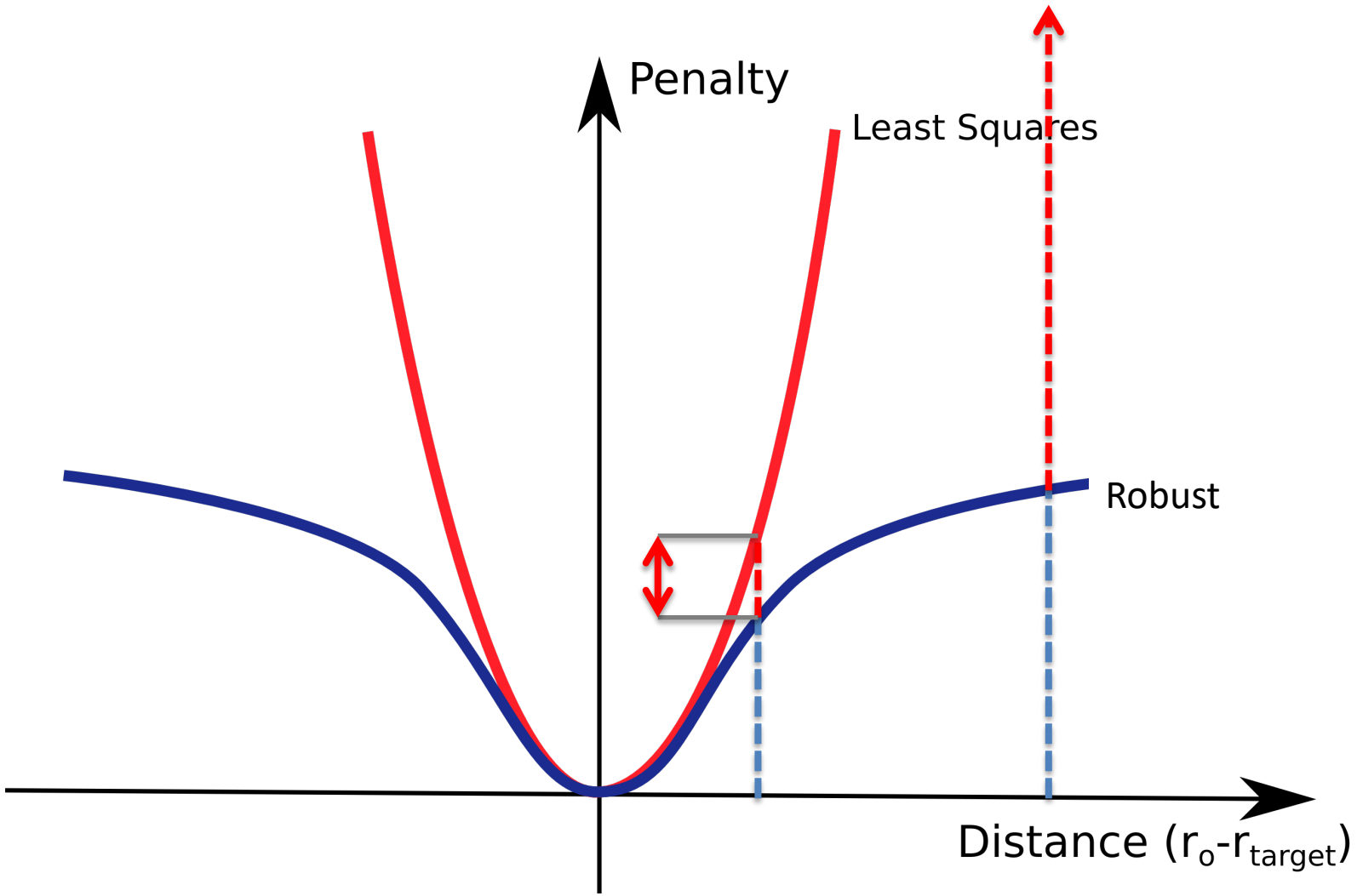
LibG Nucleic Acid Restraints



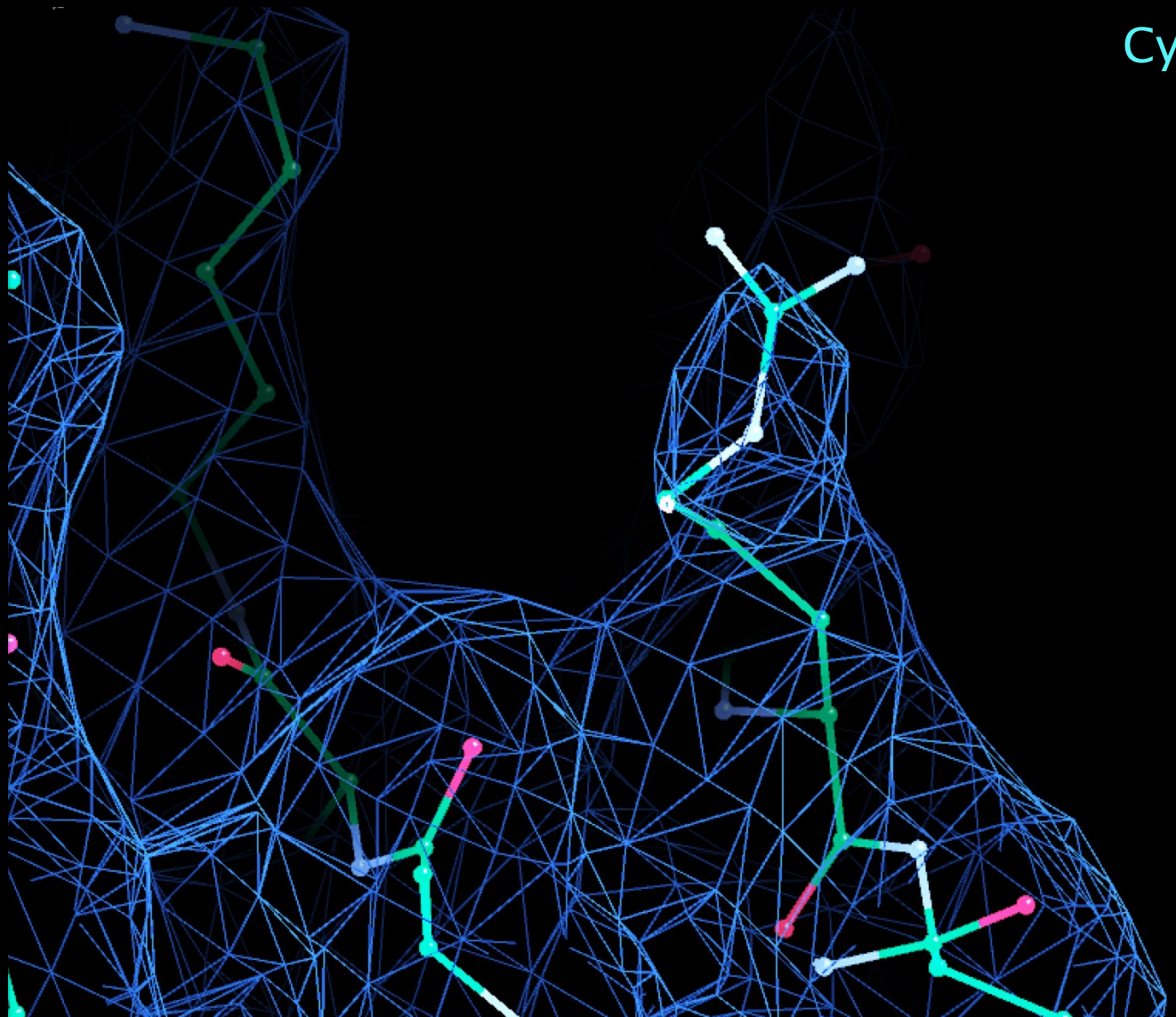
Red: long
Grey: similar
Blue: short

Visualise & edit in Coot

Robust Estimation

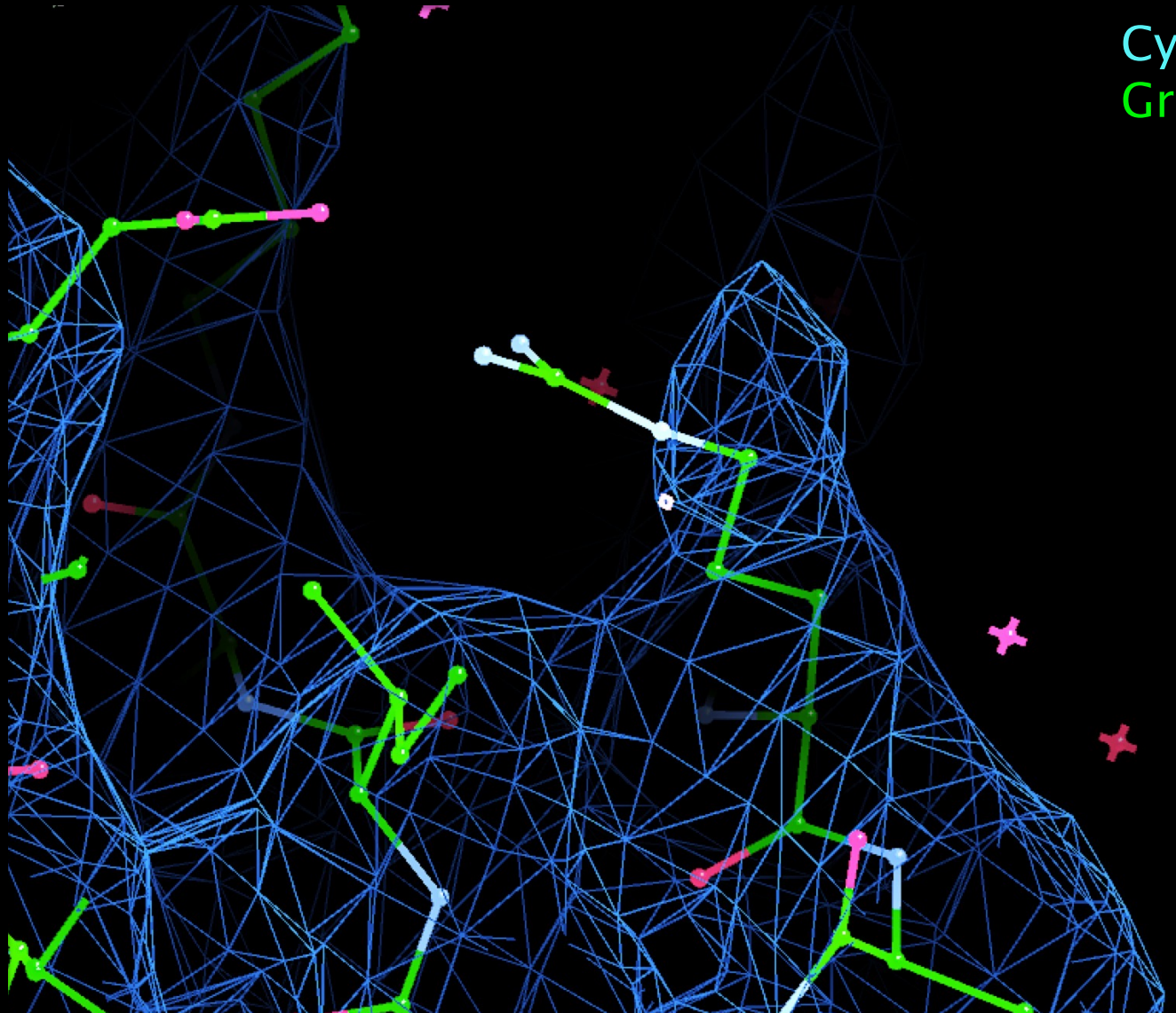


Robust Estimation



Cyan: original

Robust Estimation



Cyan: original
Green: homolog

Robust Estimation

Cyan: original
Green: homolog
Yellow: refined

Red: long
Grey: similar
Blue: short

Robust Estimation

Cyan: original
Green: homolog
Yellow: refined

Red: long
Grey: similar
Blue: short

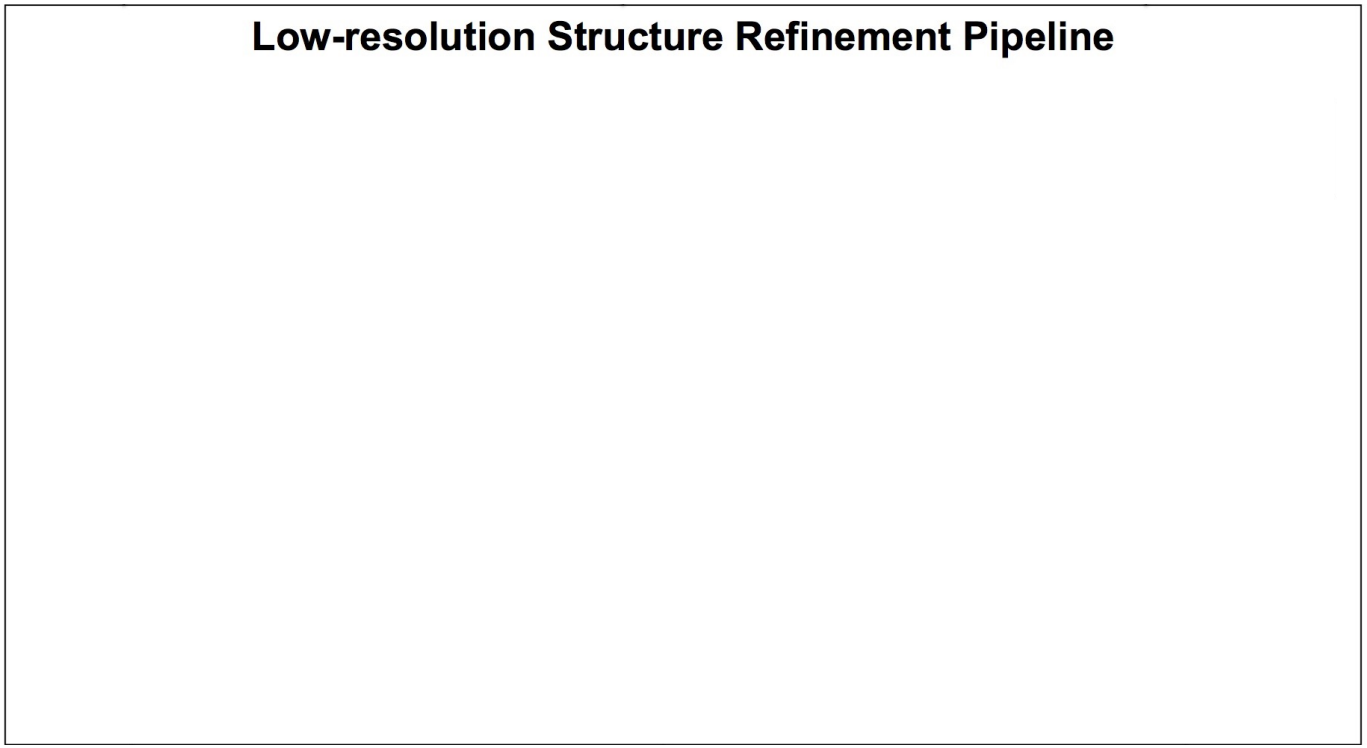
Automated pipeline - LORESTR

- 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

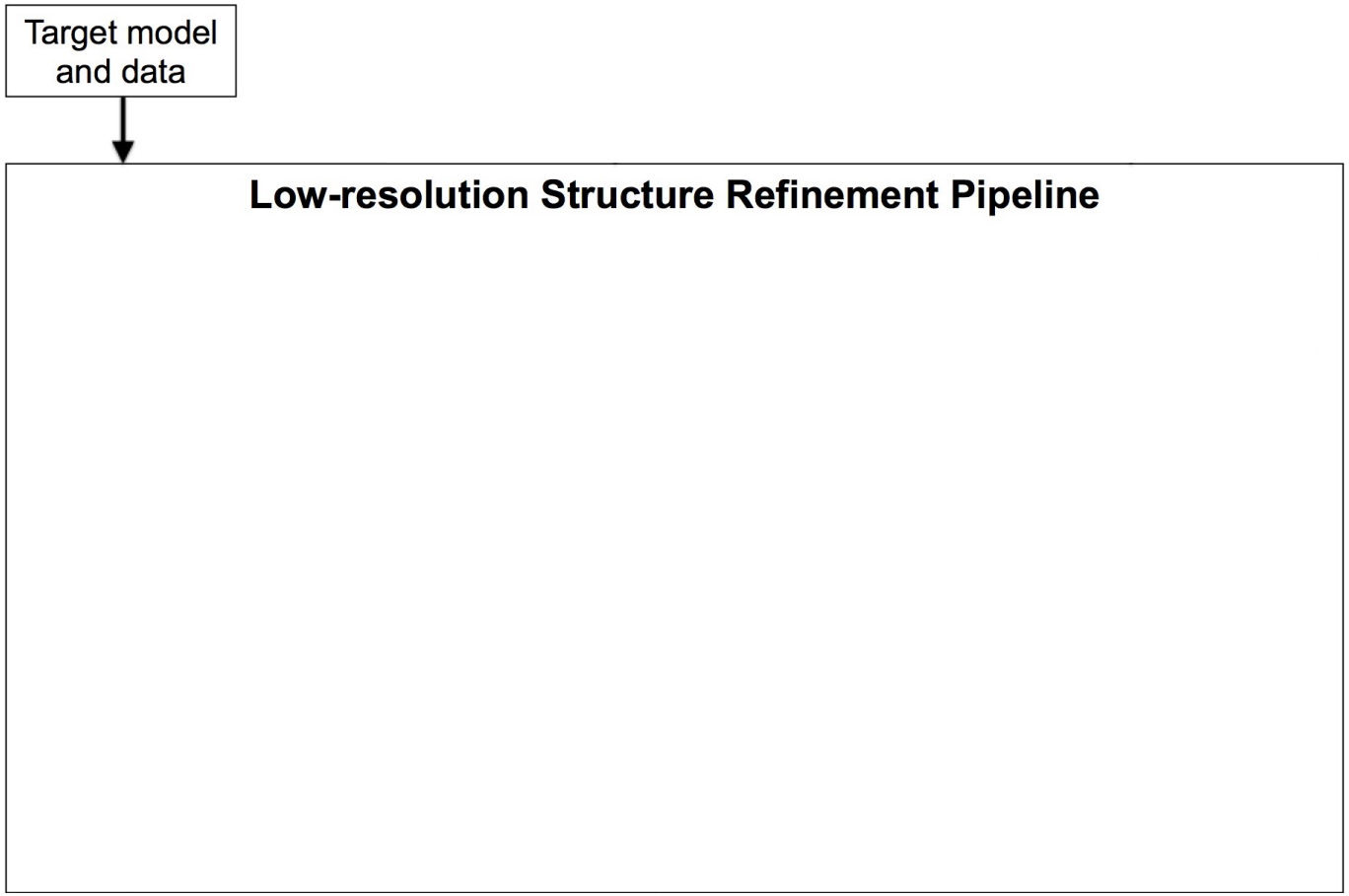
Solution:

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

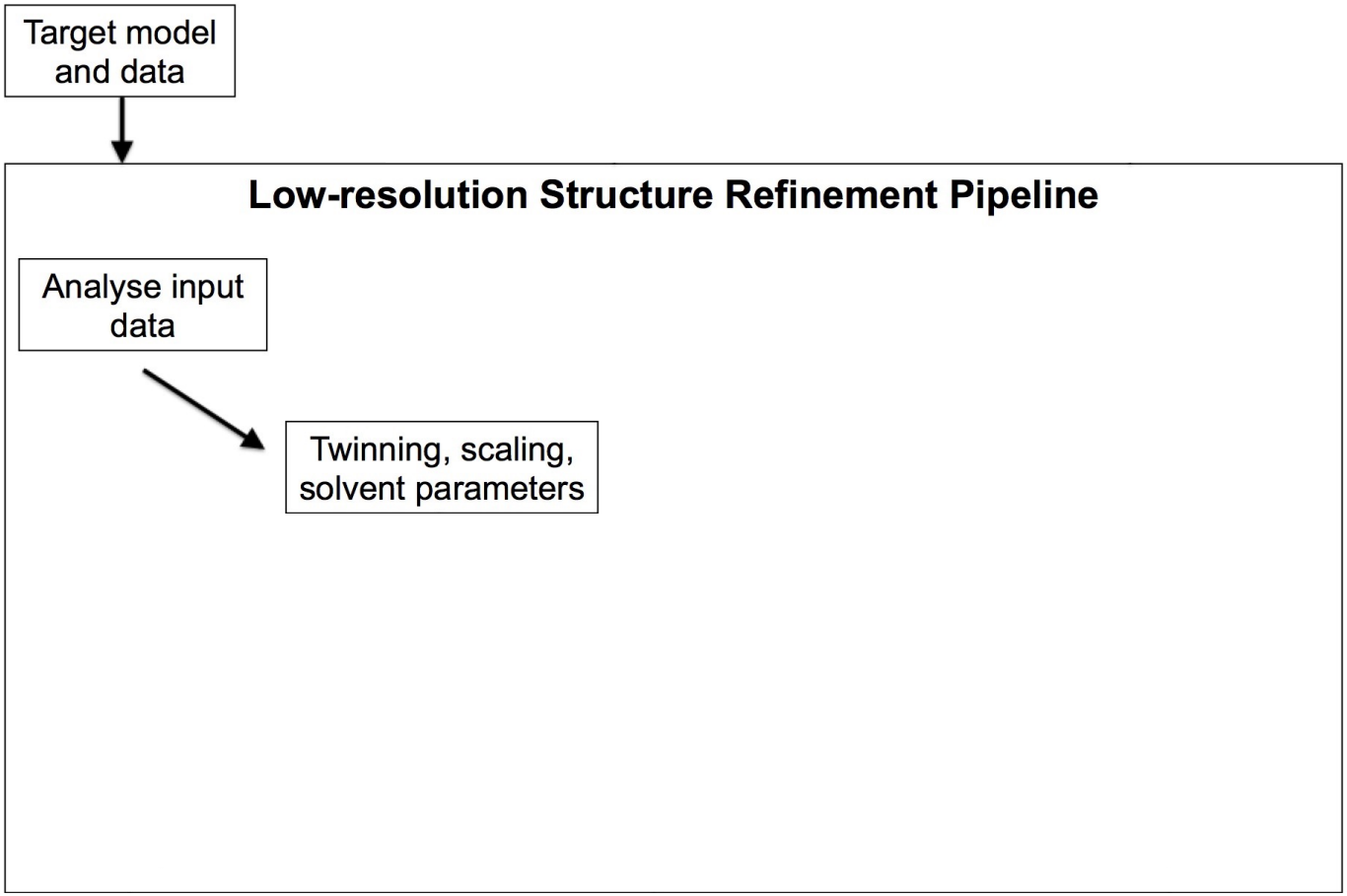
Automated pipeline - LORESTR



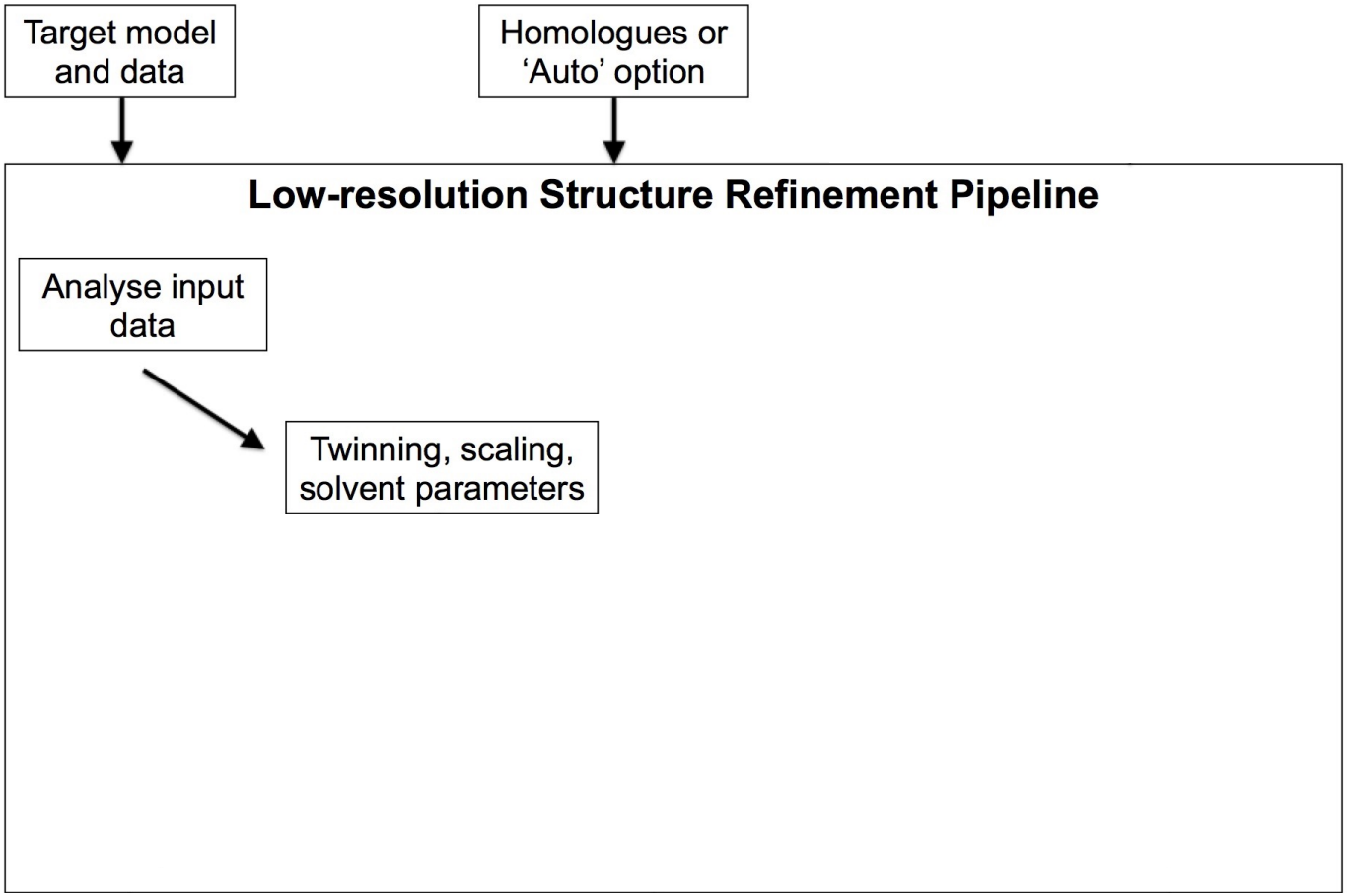
Automated pipeline - LORESTR



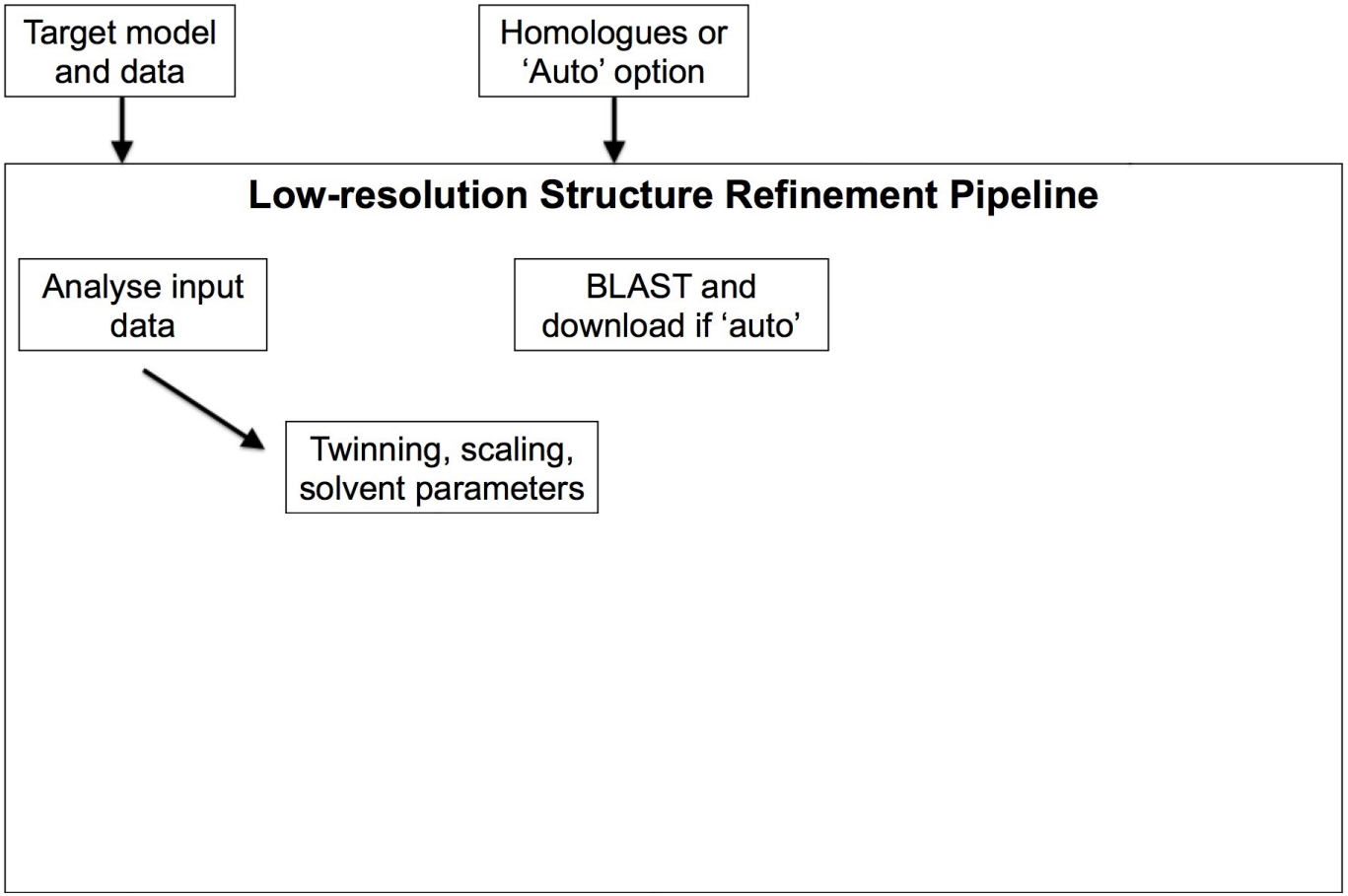
Automated pipeline - LORESTR



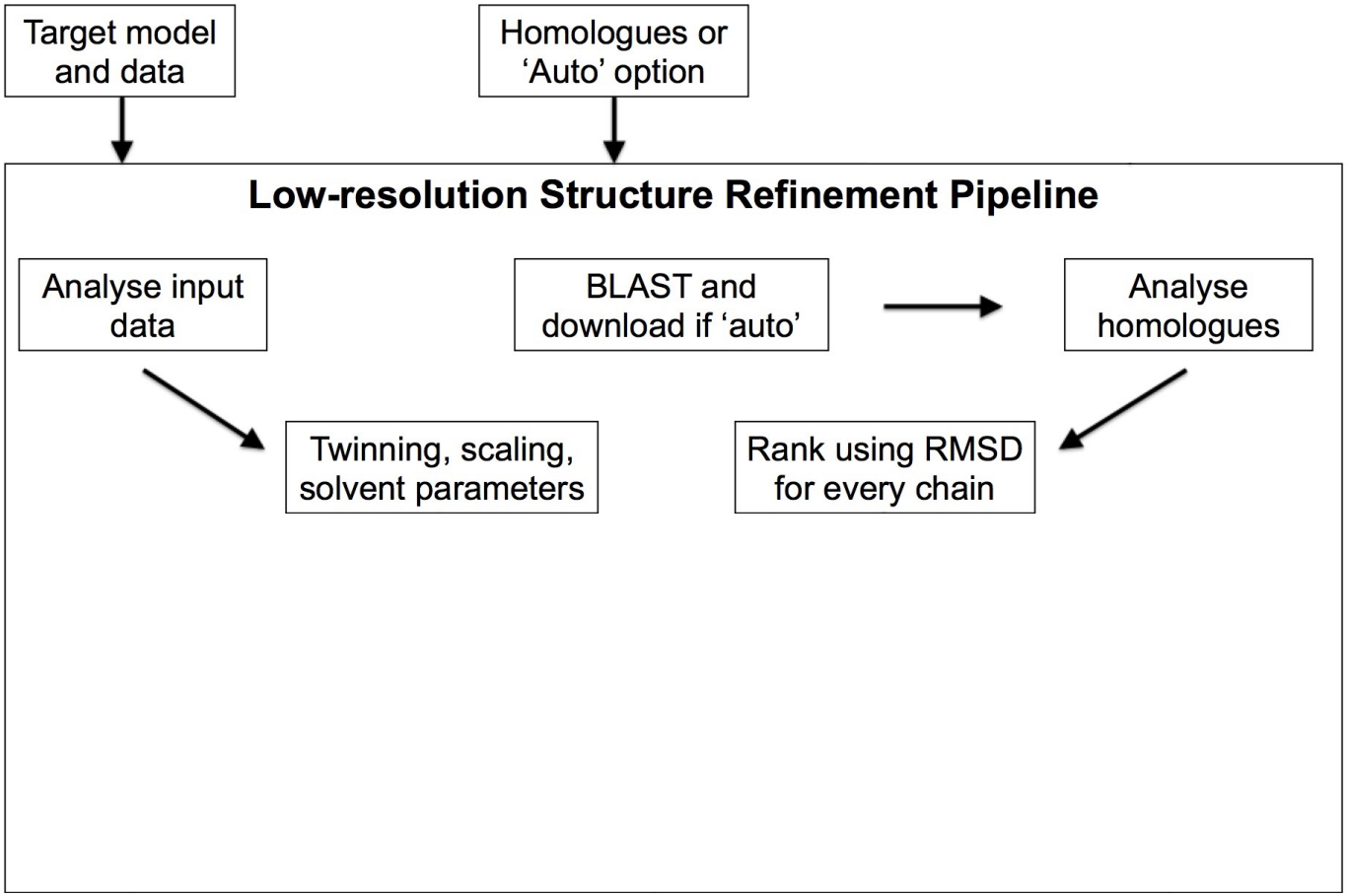
Automated pipeline - LORESTR



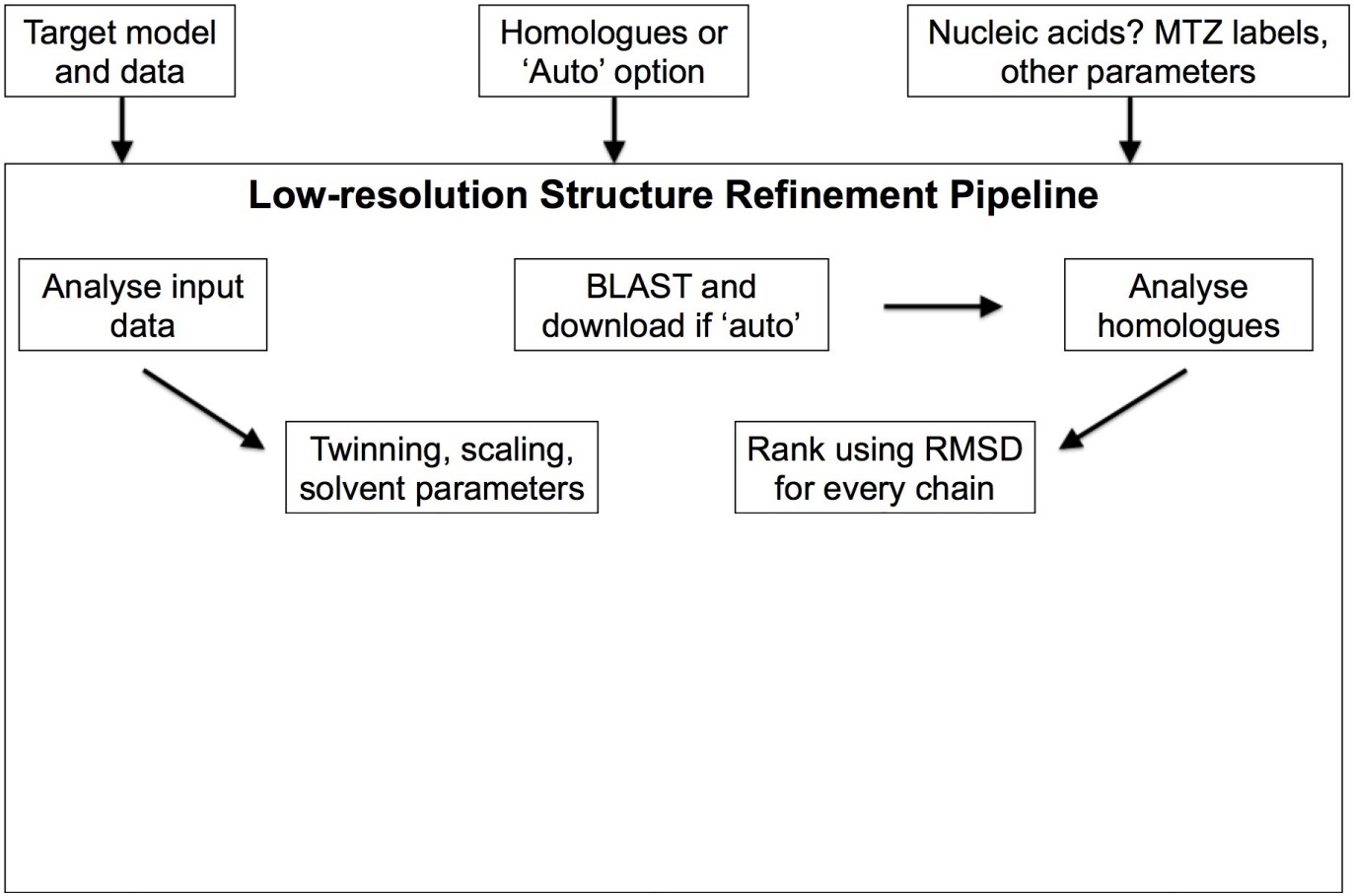
Automated pipeline - LORESTR



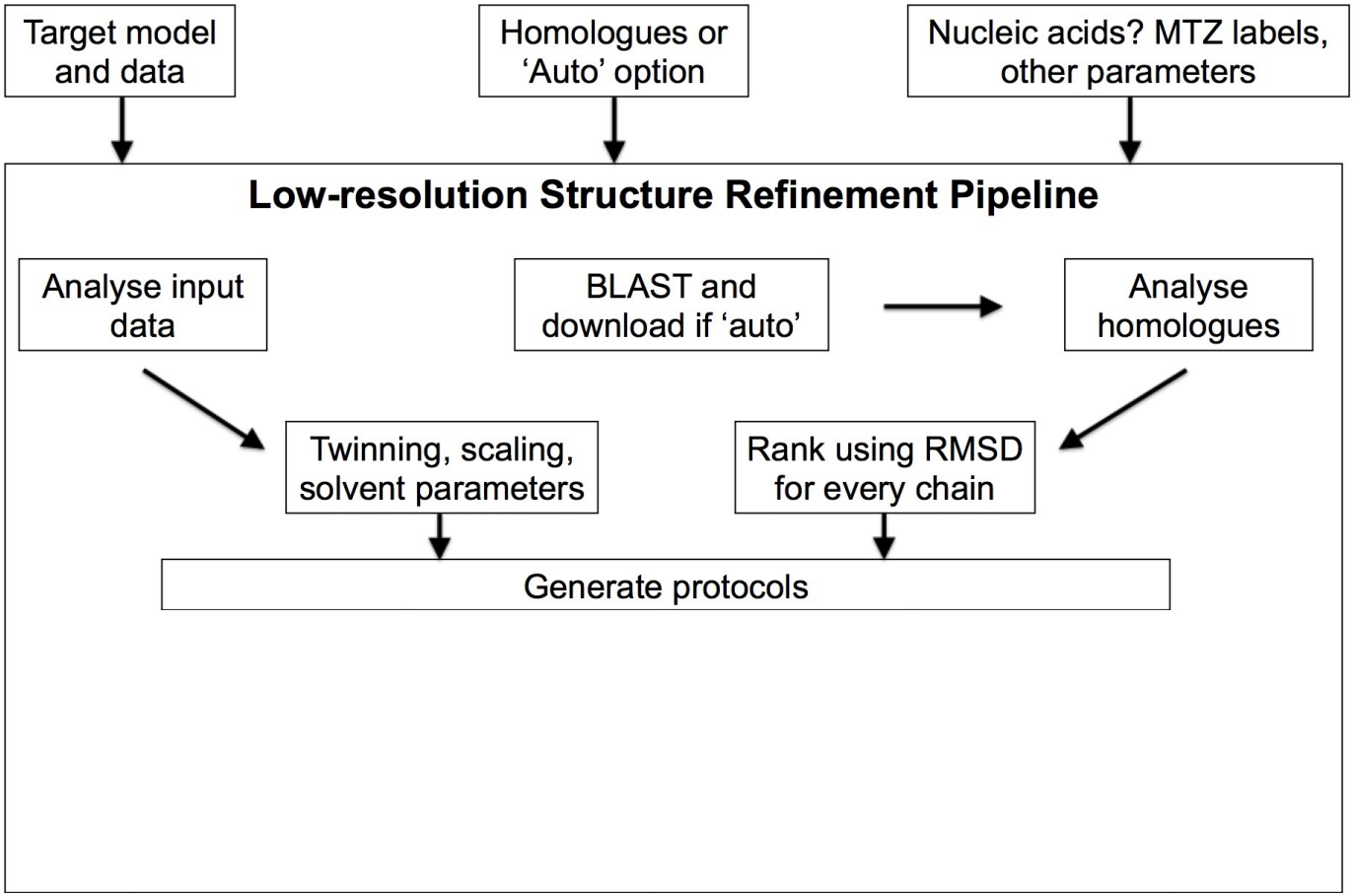
Automated pipeline - LORESTR



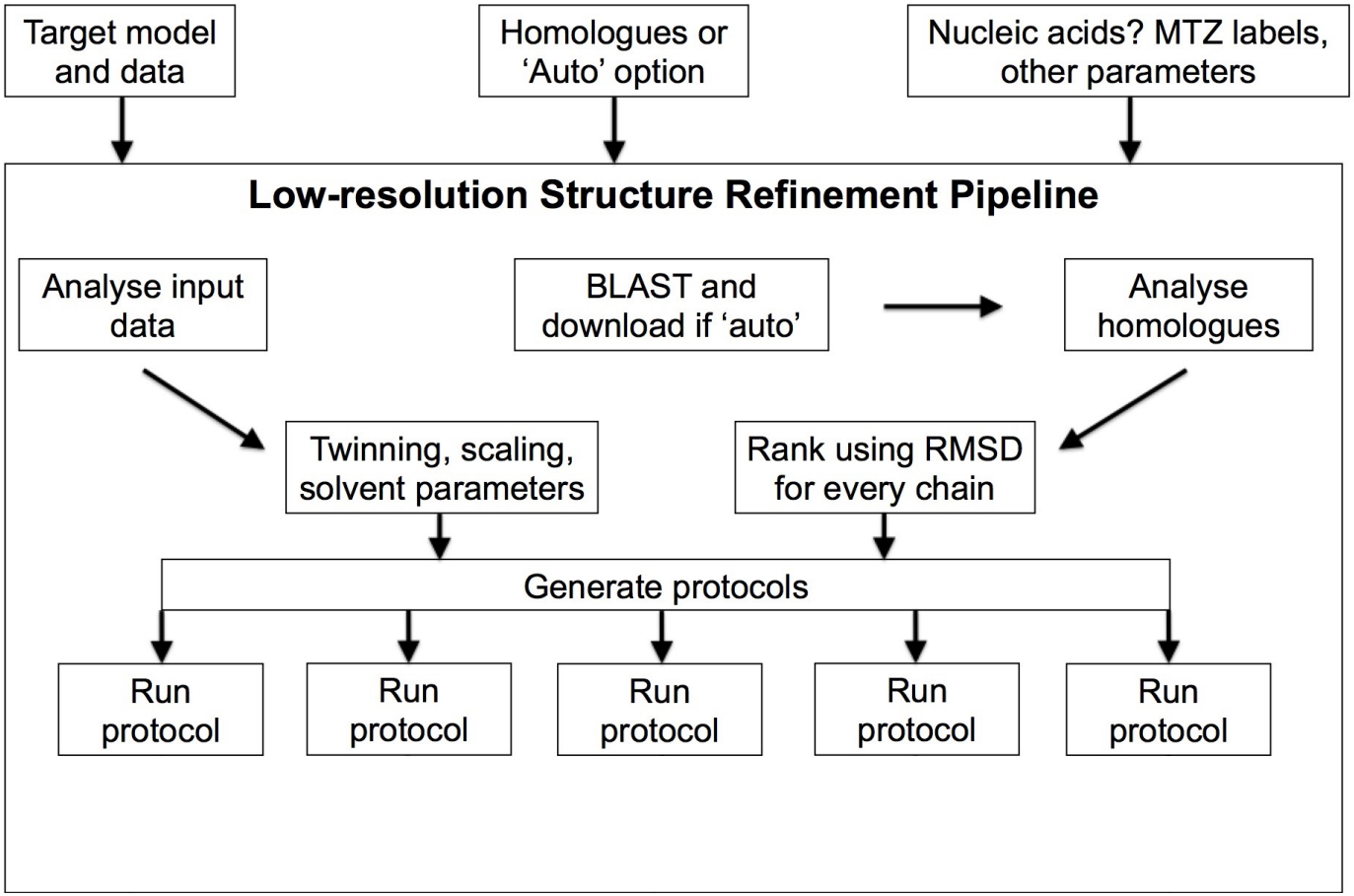
Automated pipeline - LORESTR



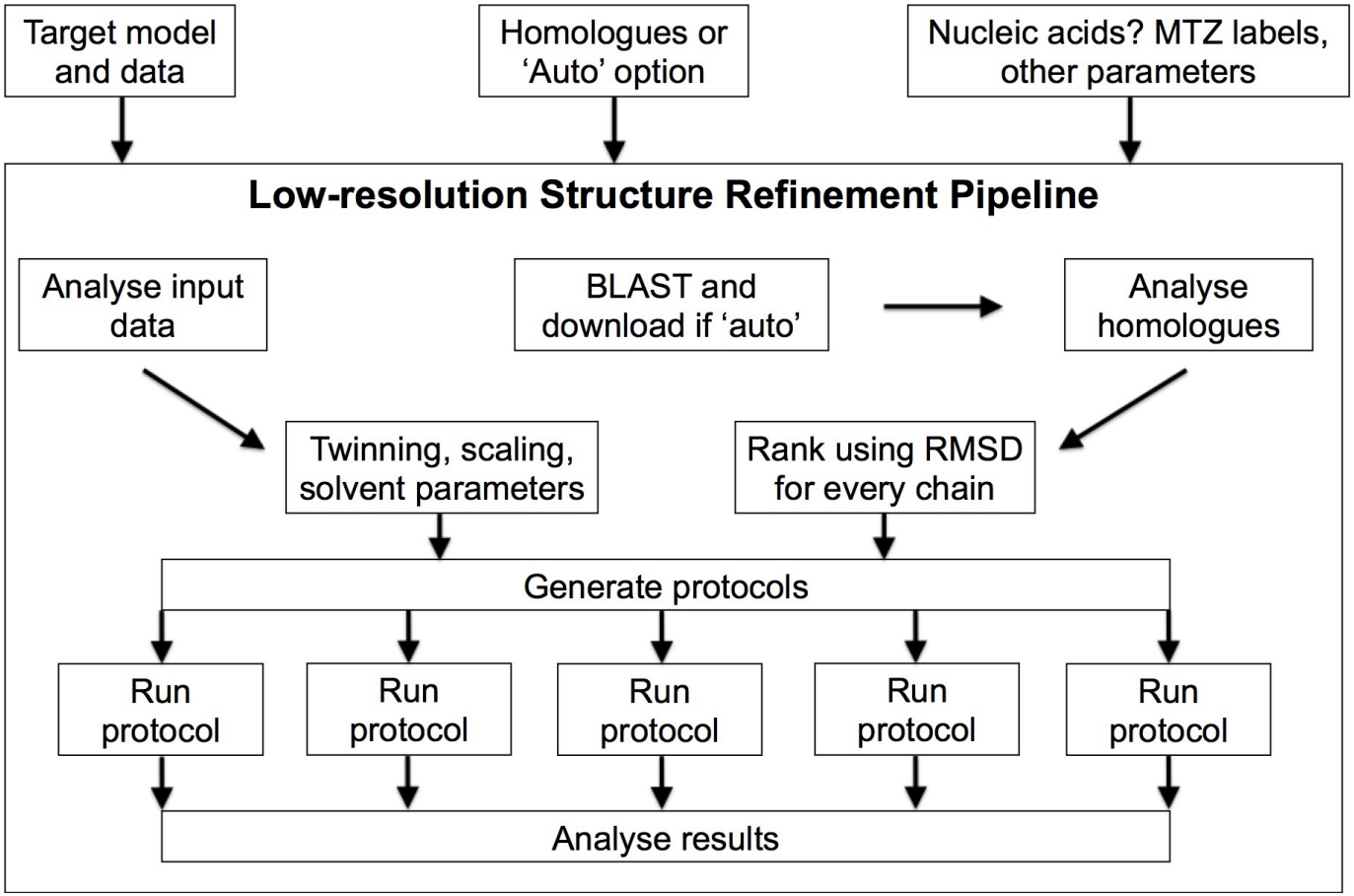
Automated pipeline - LORESTR



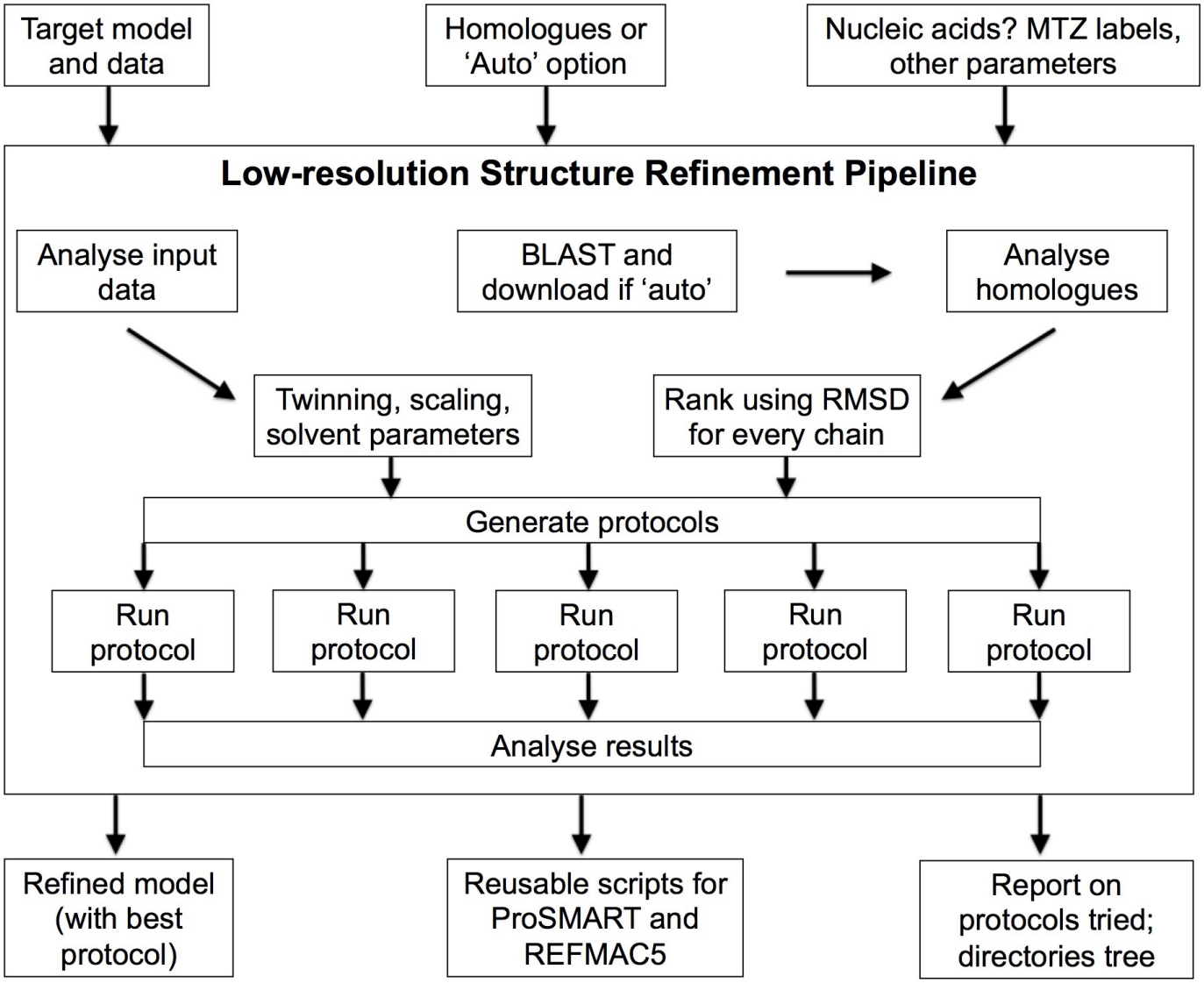
Automated pipeline - LORESTR



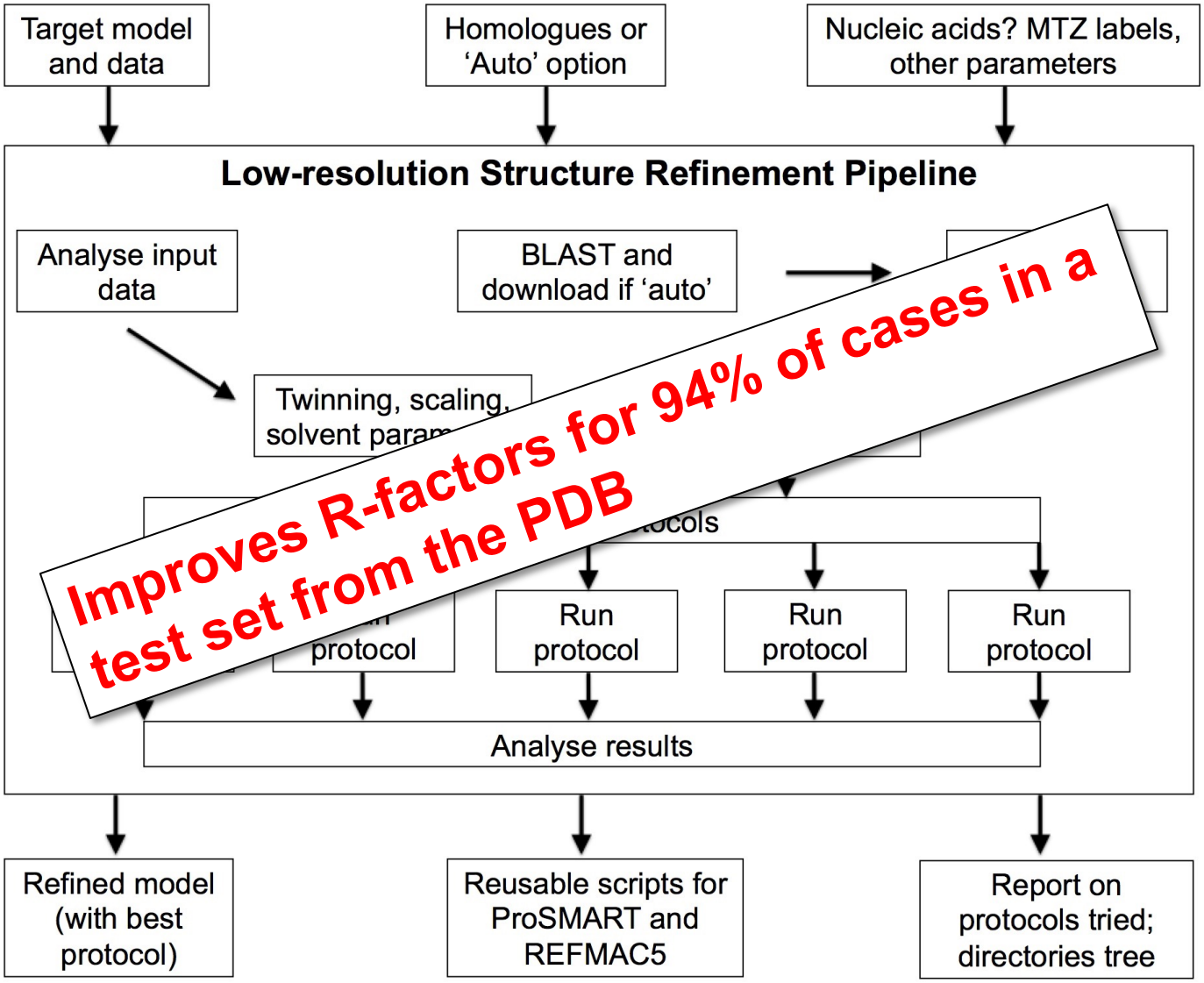
Automated pipeline - LORESTR



Automated pipeline - LORESTR

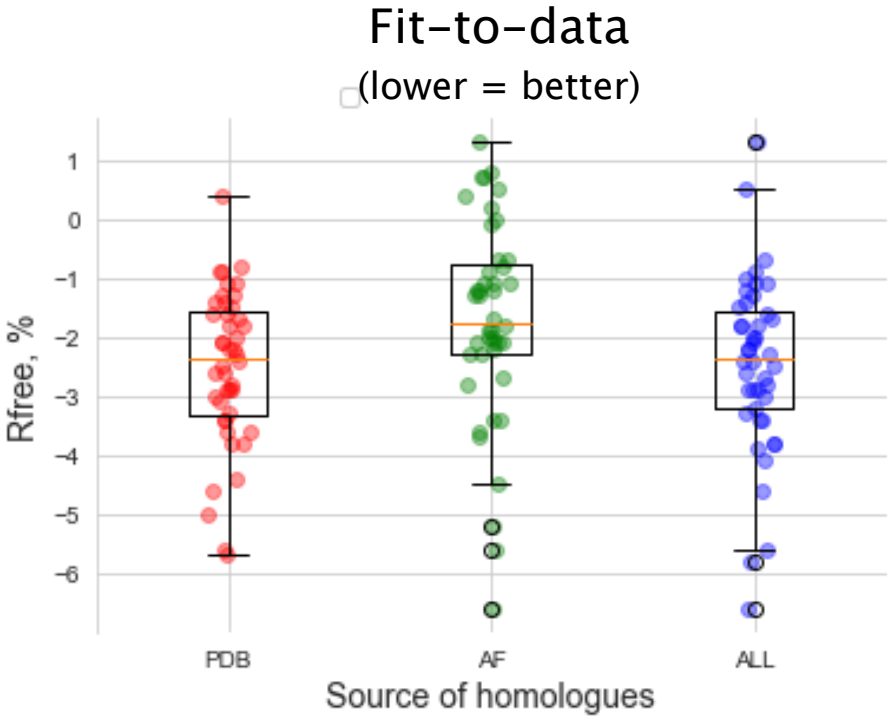


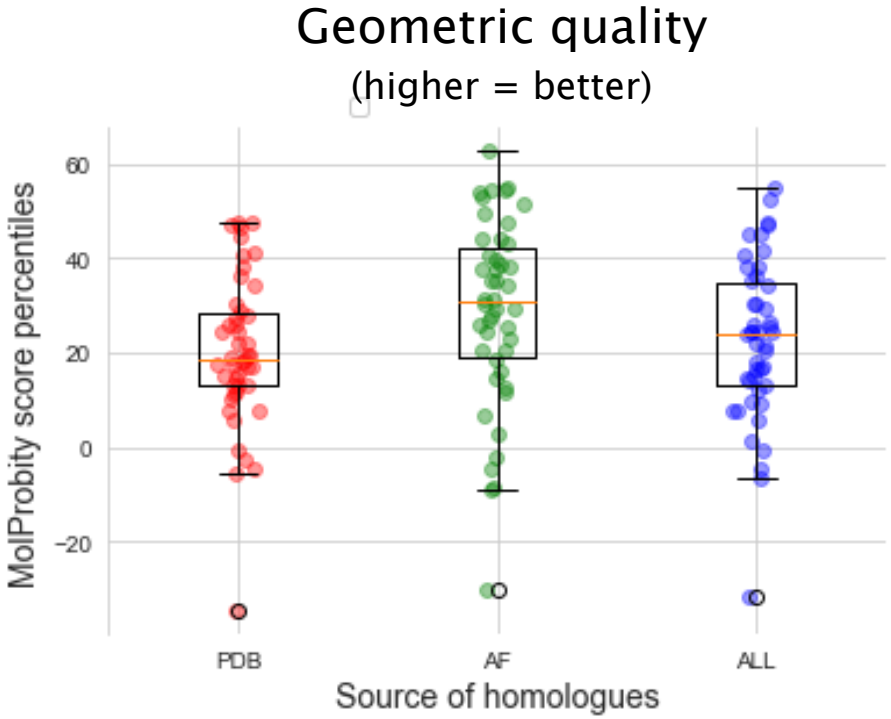
Automated pipeline - LORESTR



Utilising Predicted Structures

LORESTR performance vs original model – restraints from PDB/AF2 models





- AlphaFold2 models can be used for external restraint generation.
- Close homologues usually perform better, if available; try both.
- LORESTR automatically gets and prepares AF2 models ready for ProSMART.

(unpublished; thanks to Oleg Kovalevskiy)

Summary

CCP4 Tools for model building and refinement:

REFMAC5: Refinement, jelly body restraints, map sharpening/blurring
Refmacat: Currently uses REFMAC5 for refinement

AceDRG: Ligand dictionary and conformer generation

ProSMART: External restraints, comparative analysis

LibG: Nucleic acid restraints

LORESTR: Automated low-resolution pipeline

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)

- Jelly body – sometimes up to 200 cycles
- Rigid body refinement – but nowadays jelly body is preferred
- Shift field refinement – very quick, post-MR, pre-refinement

Medium stages – during model building

- Auto local NCS – wherever possible
- External restraints (try 40 cycles?) – when homologue available
- Otherwise, jelly body… but not together!
- H-bond and DNA/RNA restraints – if no homologue available
- Secondary structure conformation restraints – model building tool
- Add hydrogens (?) – try both. Generally: without earlier, with later.

Medium-final stages

- TLS – at medium resolutions, but beware domain boundary issues.
- Anisotropic B-factors – only at high resolution.
- Twin refinement – only if you are sure. Better to leave off until later.

Final stages of refinement

- Jelly body – around 20 cycles

Relevant Publications

Primary Citations:

REFMAC: Murshudov *et al.* (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Cryst.* D53, 240-55.

Murshudov *et al.* (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Cryst.* D67, 355-67.

ProSMART: Nicholls *et al.* (2014) Conformation-Independent Structural Comparison of macromolecules with ProSMART. *Acta Cryst.* D70, 2487-99.

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

LORESTR: Kovalevskiy *et al.* (2016) Automated refinement of macromolecular structures at low resolution using prior information. *Acta Cryst.* D72, 1149-61.

AceDRG: Long *et al.* (2017) AceDRG: a stereochemical description generator for ligands. *Acta Cryst*. D66, 486-501.

Coot: Emsley & Cowtan (2004) Coot: model-building tools for molecular graphics. *Acta Cryst*. D60, 2126-32.

Emsley *et al.* (2010) Features and development of Coot. *Acta Cryst*. D66, 486-501.

Relevant Publications

Low-resolution refinement with REFMAC, 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

CCP4 Core

Eugene Krissinel
Andrey Lebedev
Charles Ballard
Ronan Keegan
Ville Uski
Maria Fando

MRC-LMB Computational Structural Biology Group

Garib Murshudov - *REFMAC5*
Keitaro Yamashita - *REFMACAT*
Fei Long - *AceDRG, LibG*
Paul Emsley - *Coot*
Lucrezia Catapano - *Coot*

CCP-EM

Tom Burnley
Colin Palmer
Rangana Warshamanage
Agnel Joseph
Martyn Wynn

Other Collaborators

Marcin Wojdyr (Global Phasing Ltd)
Oleg Kovalevskiy - *LORESTR*
Robbie Joosten
Jon Agirre
Marcus Fischer
Alan Brown
Ben Bax
Martin Noble
Stuart McNicholas