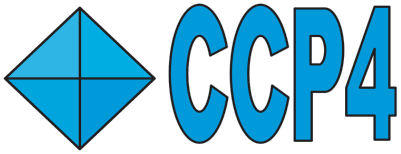
Cryo-EM Macromolecular Refinement

CCP-EM Icknield Workshop – RAL/DLS

6th November 2024

Rob Nicholls

robert.nicholls@stfc.ac.uk



MRC Laboratory of Molecular Biology

Science and Technology Facilities Council

Research Complex at Harwell

REFMAC5
Macromolecular structure refinement

Servalcat
Macromolecular structure refinement

COOT
Visualization and model building

ProSMART
Restraint generation and comparative structural analysis

A few key tools In CCP4/CCP-EM

EMDA
Cryo-EM map calculation and map/model validation

LibG
Nucleic acid restraint generation

AceDRG
Ligand restraint dictionary and conformer generation

MRC-LMB, Cambridge:

Cryo-EM vs MX

Tools designed for low-resolution MX have been repurposed for high-resolution cryo-EM

Some relevant differences:

- "Observations" are electrostatic potential maps
  - *In MX the observations are estimated diffraction spot intensities (usually converted to amplitudes)*
- Able to obtain phase information (although amplitudes and phases are noisy)
  - *map is not updated as model is refined*
- No crystallographic properties (e.g. space groups) or peculiarities (e.g. twinning).
- Point group or helical symmetry (instead of xtal symmetry).
- No fixed unit cell - boundaries are not enforced; artificial boxes are used
- Concept of "resolution"
  - *quoted resolution in MX is the diffraction limit (resolution of the largest Miller indices used)*
  - *In cryo-EM we can consider local resolution*
  - *local map quality varies greatly within and between reconstructions*

One similarity:

- Scattering: High-resolution information loss
  - *most refinement methods developed for MX can be transferred to cryo-EM*

Cryo-EM Model Refinement

Uses maximum a-postiori (MAP) refinement, just like in MX… but

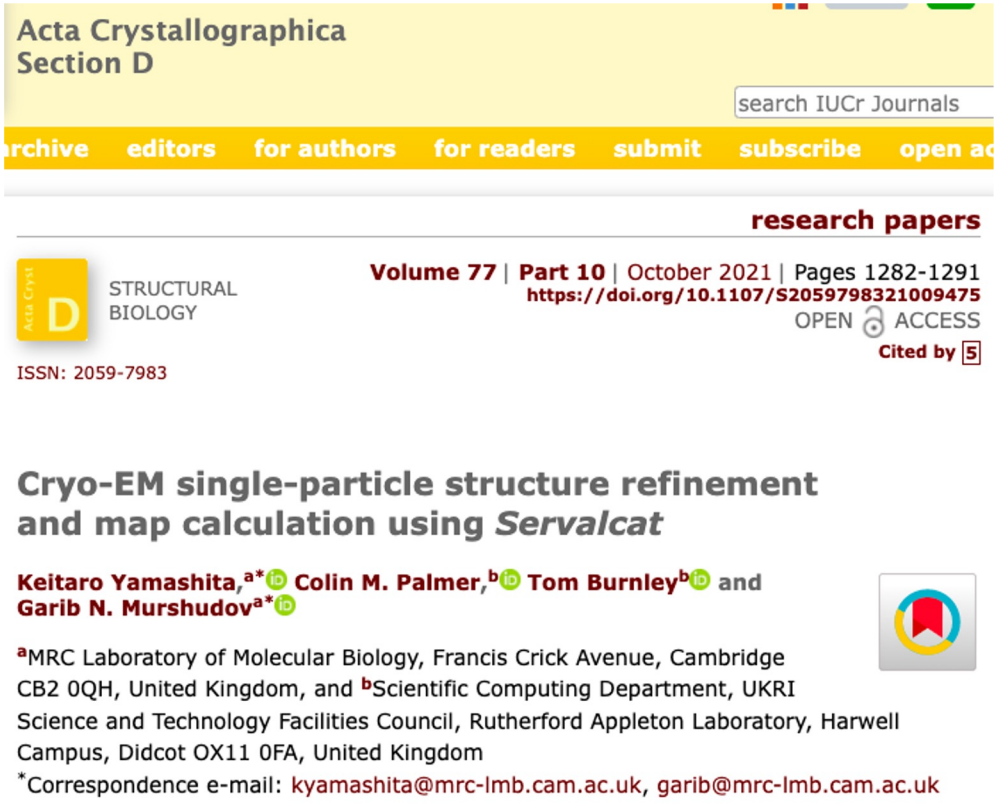
MX Optimise agreement between SF amplitudes – R-factors

Cryo-EM Optimise agreement between SFs – FSCs

Servalcat:

- Symmetry constraints
- Map masking & trimming
- Difference maps
- Bookkeeping via Gemmi
- …

Successor to REFMAC5



Acta Crystallographica
Section D

search IUCr Journals

archive editors for authors for readers submit subscribe open ac

research papers

STRUCTURAL BIOLOGY

Volume 77 | Part 10 | October 2021 | Pages 1282-1291
https://doi.org/10.1107/S2059798321009475
OPEN ACCESS
Cited by 5

ISSN: 2059-7983

Cryo-EM single-particle structure refinement and map calculation using *Servalcat*

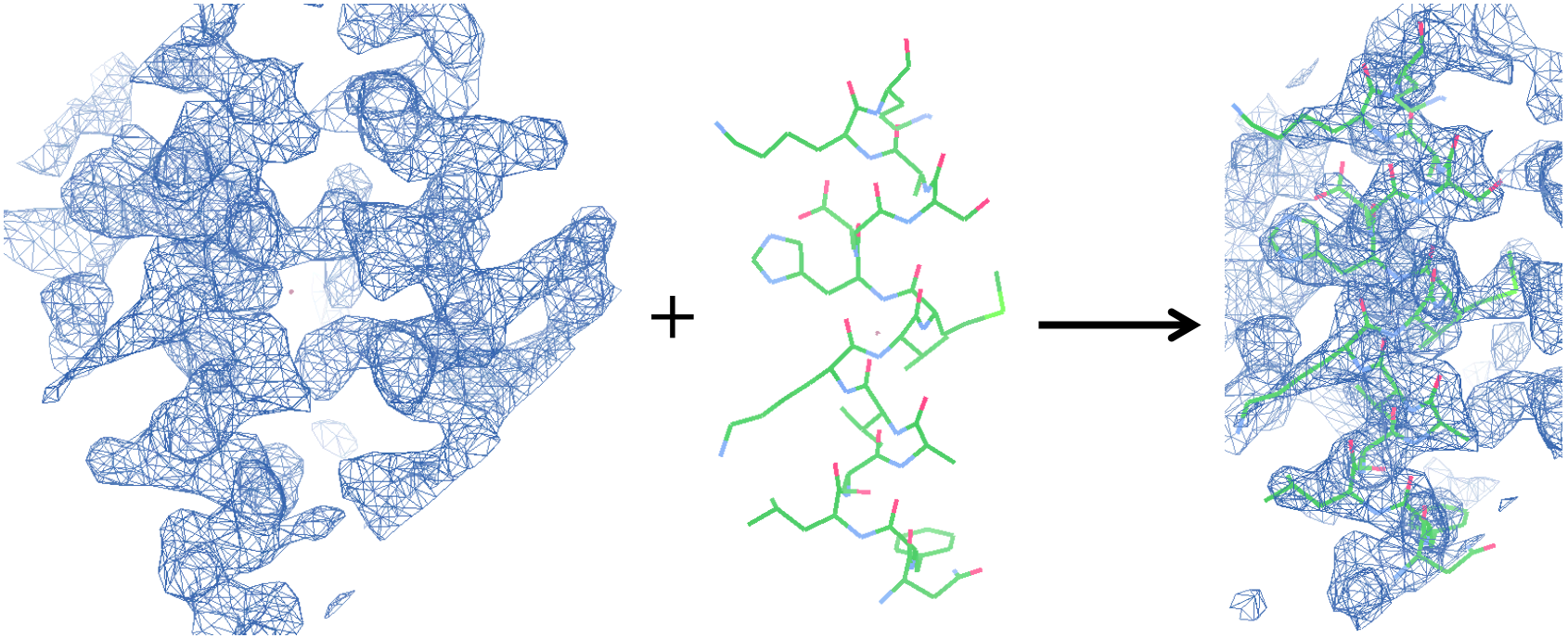
Keitaro Yamashita,[a]* Colin M. Palmer,[b] Tom Burnley[b] and Garib N. Murshudov[a]*

[a]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, United Kingdom, and [b]Scientific Computing Department, UKRI Science and Technology Facilities Council, Rutherford Appleton Laboratory, Harwell Campus, Didcot OX11 0FA, United Kingdom
*Correspondence e-mail: kyamashita@mrc-lmb.cam.ac.uk, garib@mrc-lmb.cam.ac.uk

Purpose of Refinement

To fit an atomic model into observed data

- *Model should agree with the observed data*
- *Model must be chemically and structurally sensible*



Data

Atomic model

Fit and refine

Structure factors (reconstruction/map) are treated as observations...

Model Refinement

Modern refinement programs use maximum a-posteriori estimation

MX/Cryo-EM 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:

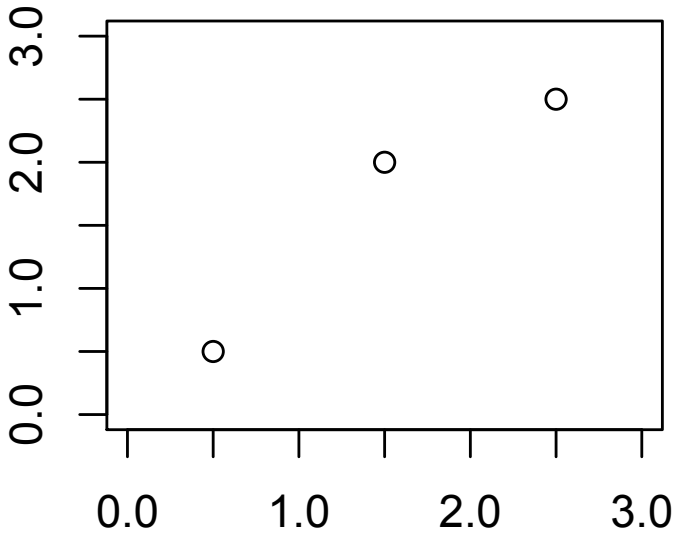
- Bond lengths
- Angles
- Chirals
- Planes
- Some torsion angles
- B-values
- VDW repulsions

These help to ensure that the model is chemically sensible

Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.

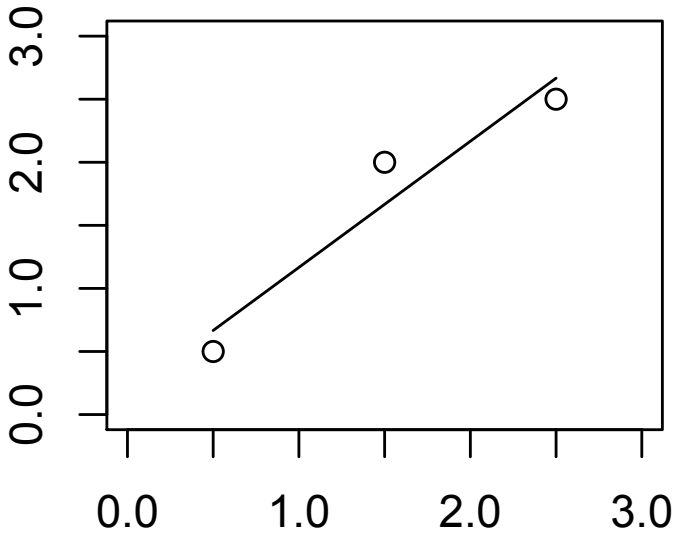


Example: Fitting a line $y = a + bx$

Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.

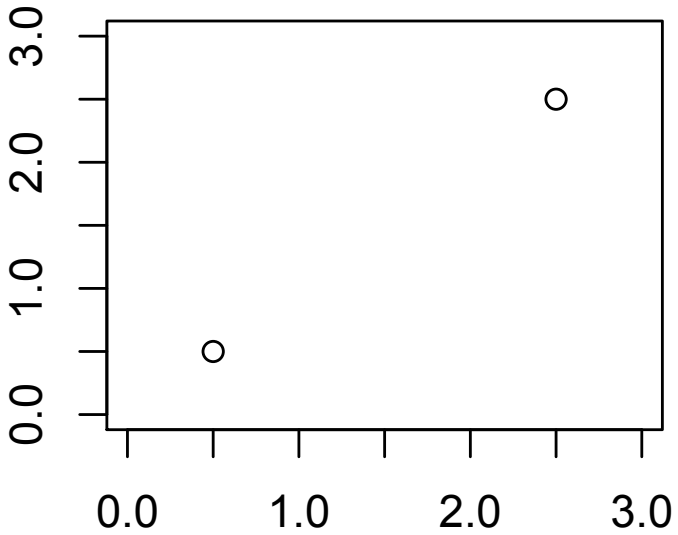


Example: Fitting a line $y = a + bx$

Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.



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

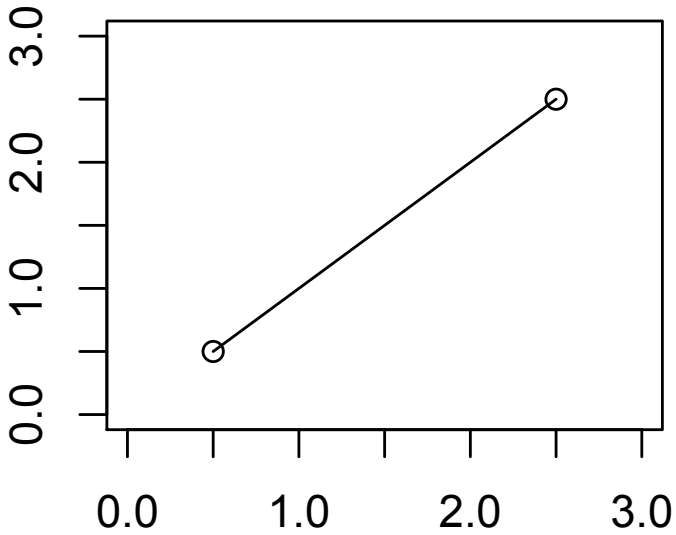
Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.

Can fit a line

Line is unreliable



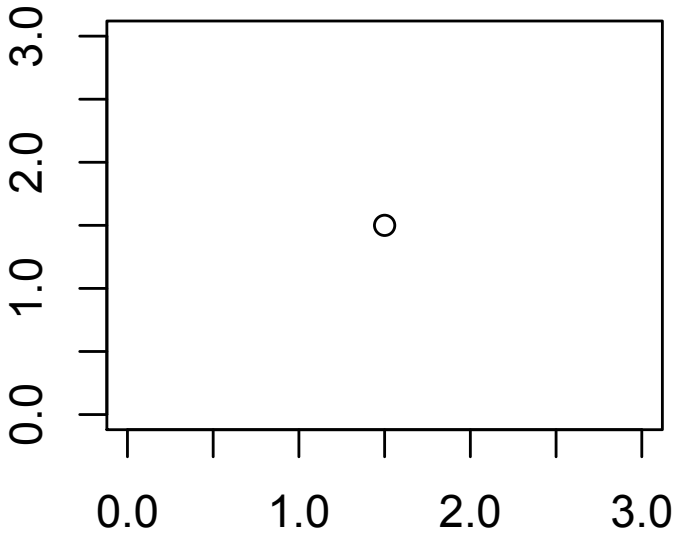
Overfitting
Model Bias

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

Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.



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

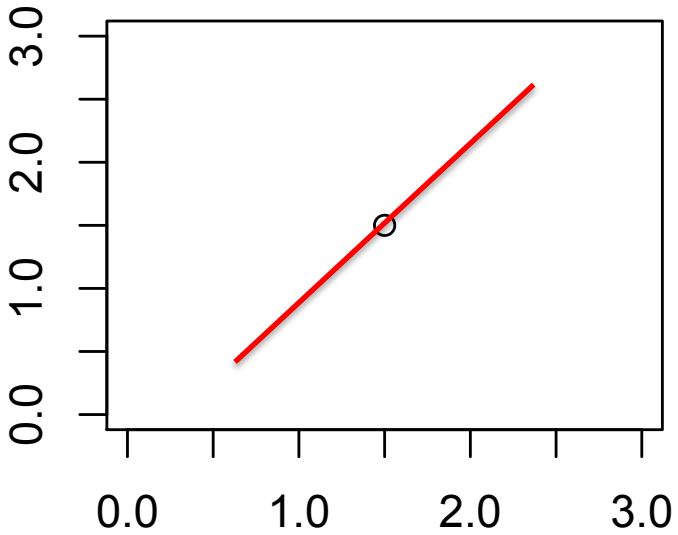
Restraints

Why introduce so many restraints?

Answer: to improve the observation:parameter ratio.

Insufficient observations!

Unstable refinement



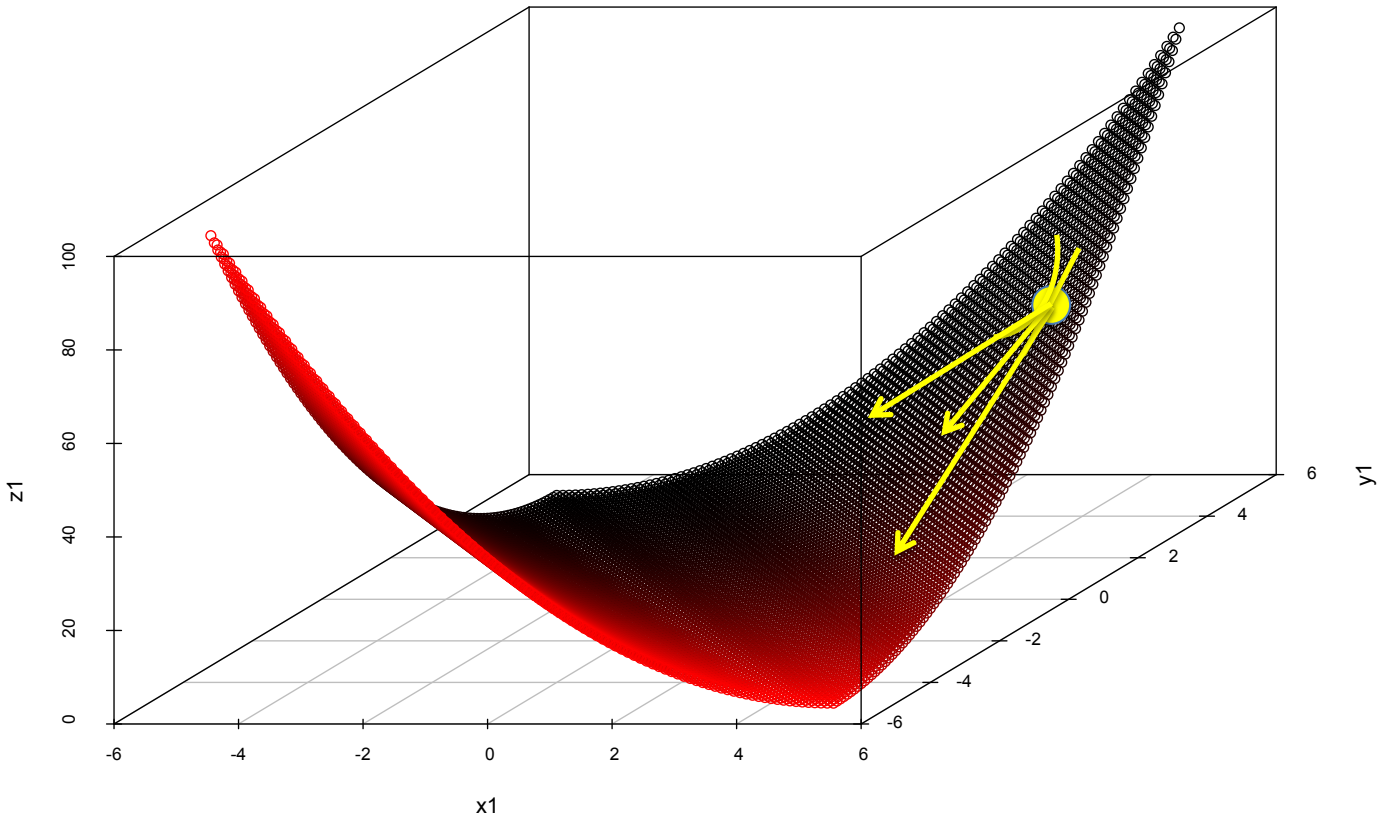
Ill-posed problem

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

Regularisation

Example:

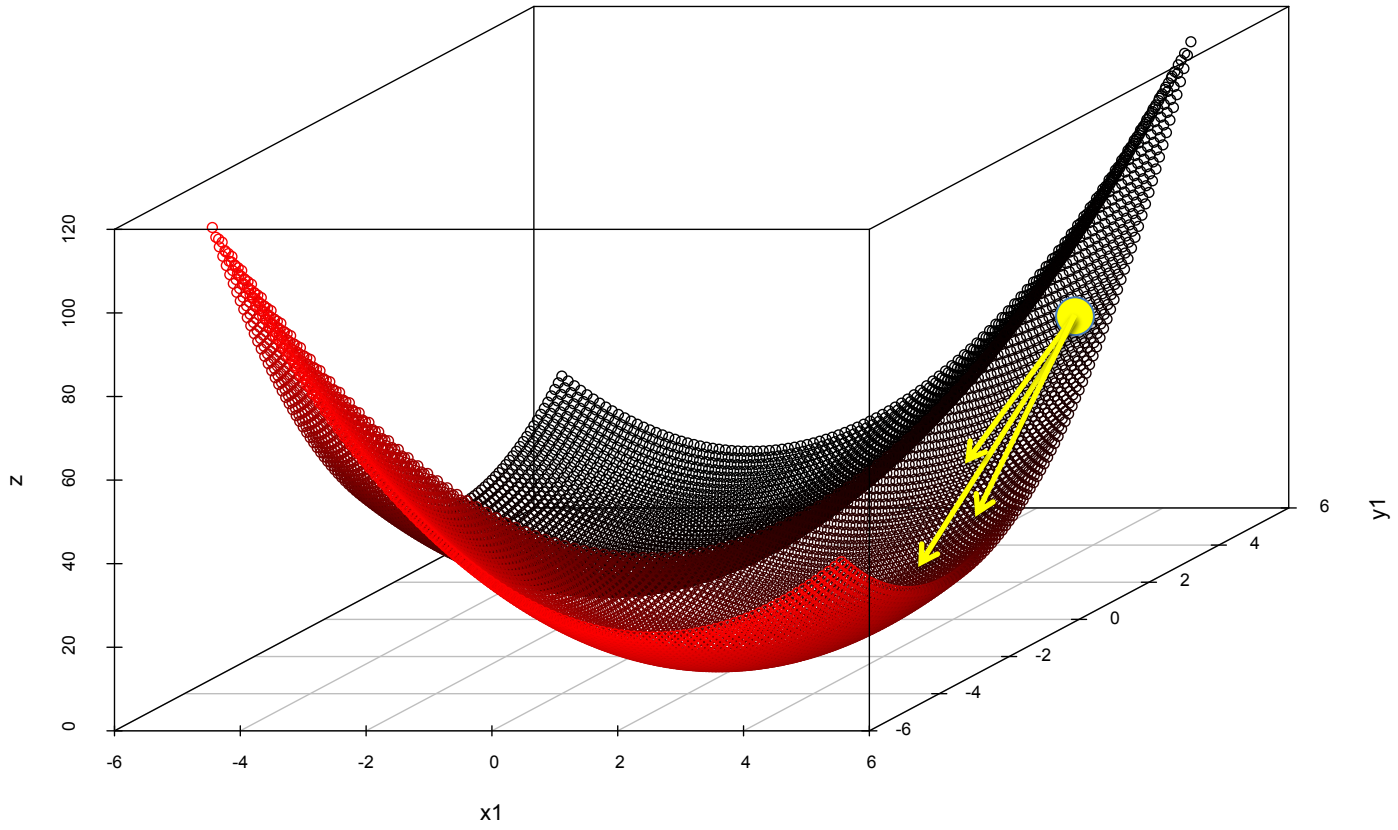
$$z = (x + y)^2$$



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

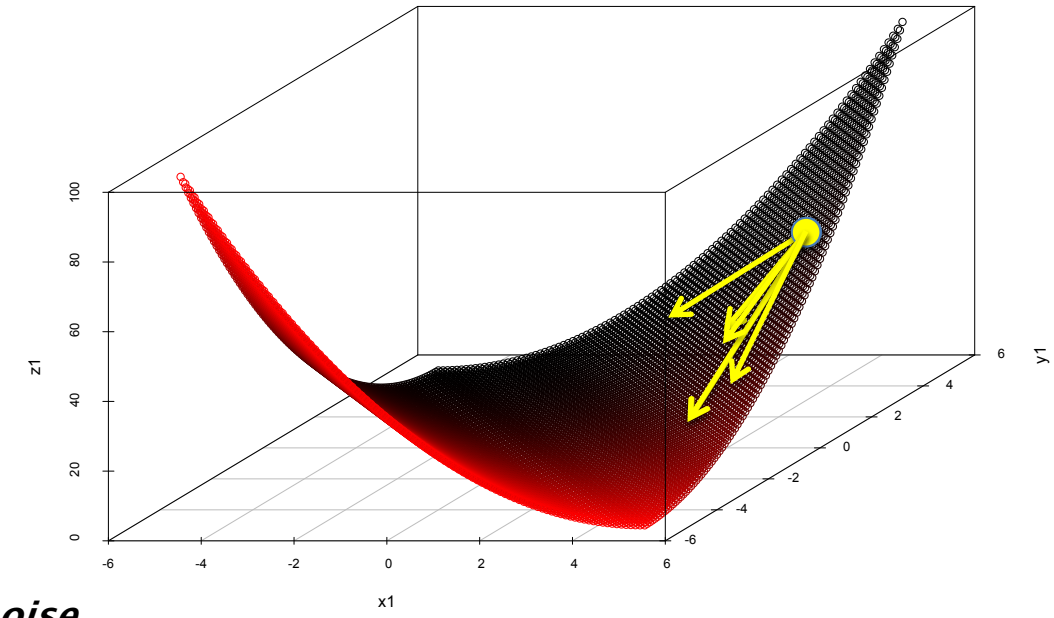
Regularisers without a target:

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

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

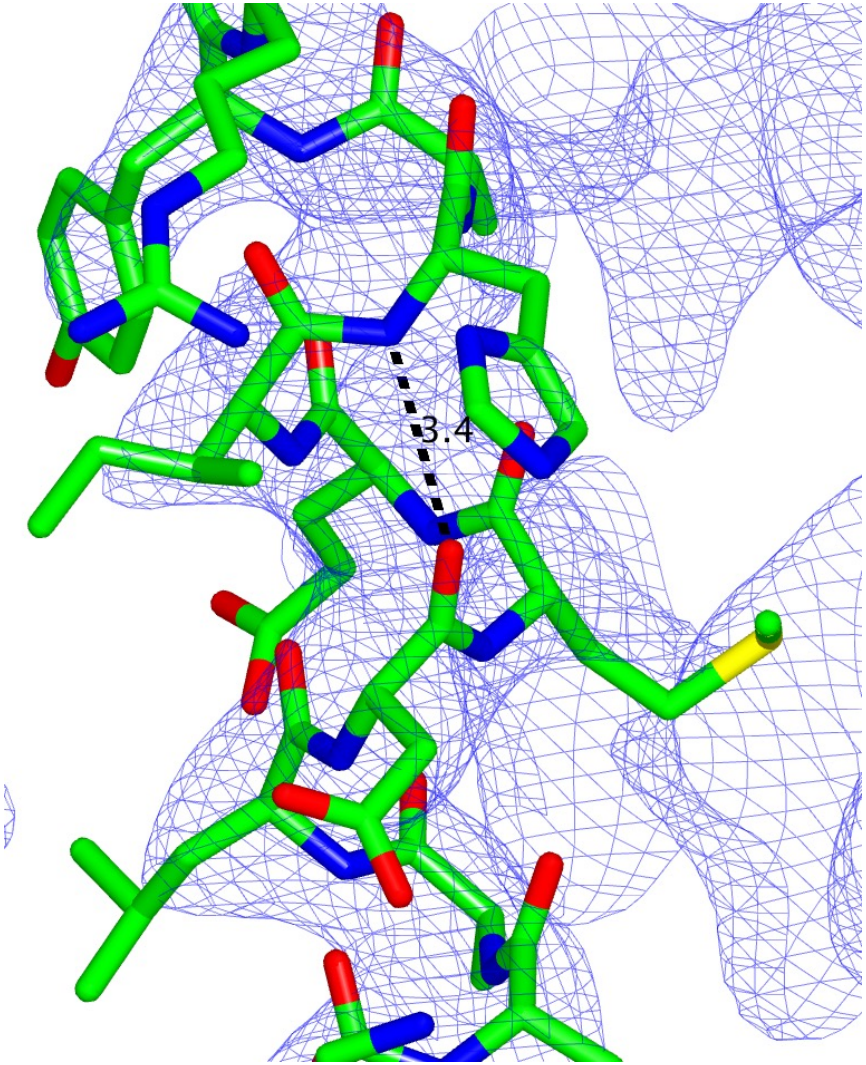
ProSMART

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

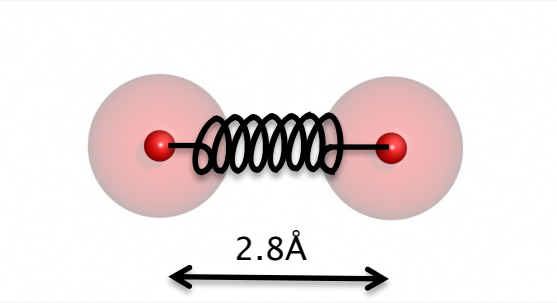
Independent of global conformation

ProSMART External Restraints



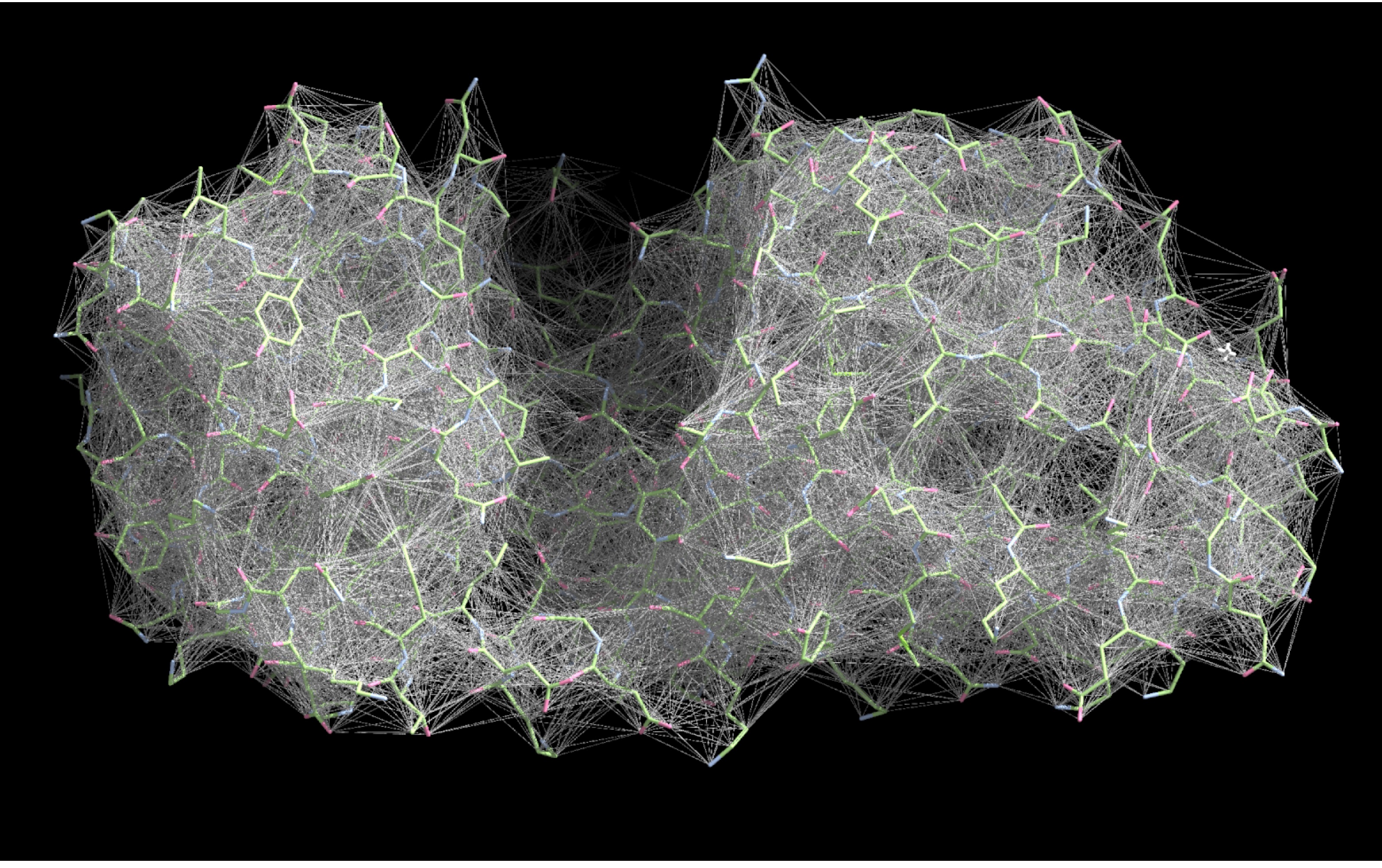
3g4w – 3.7Å

Prior information:

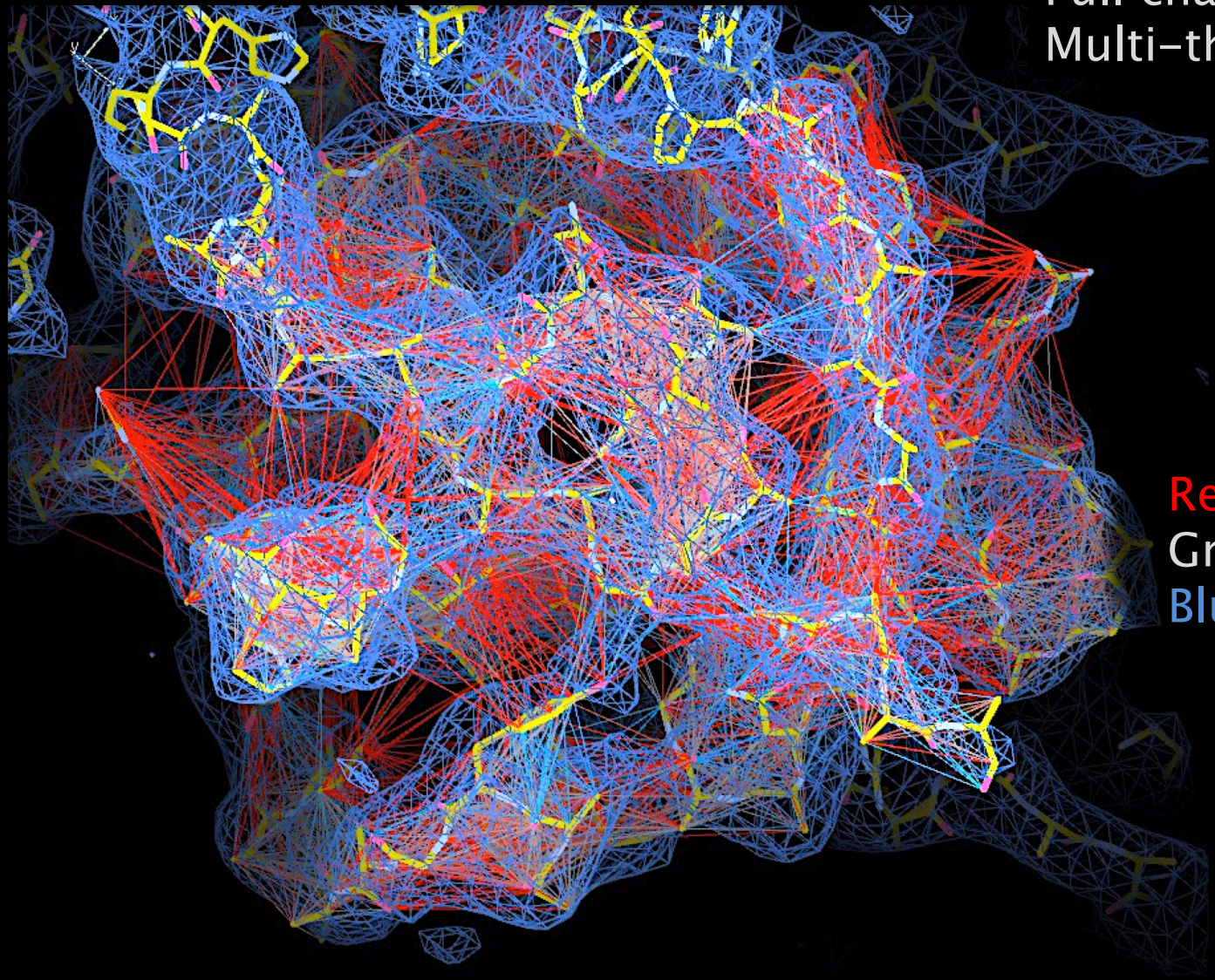


Stabilises structural features

ProSMART External Restraints



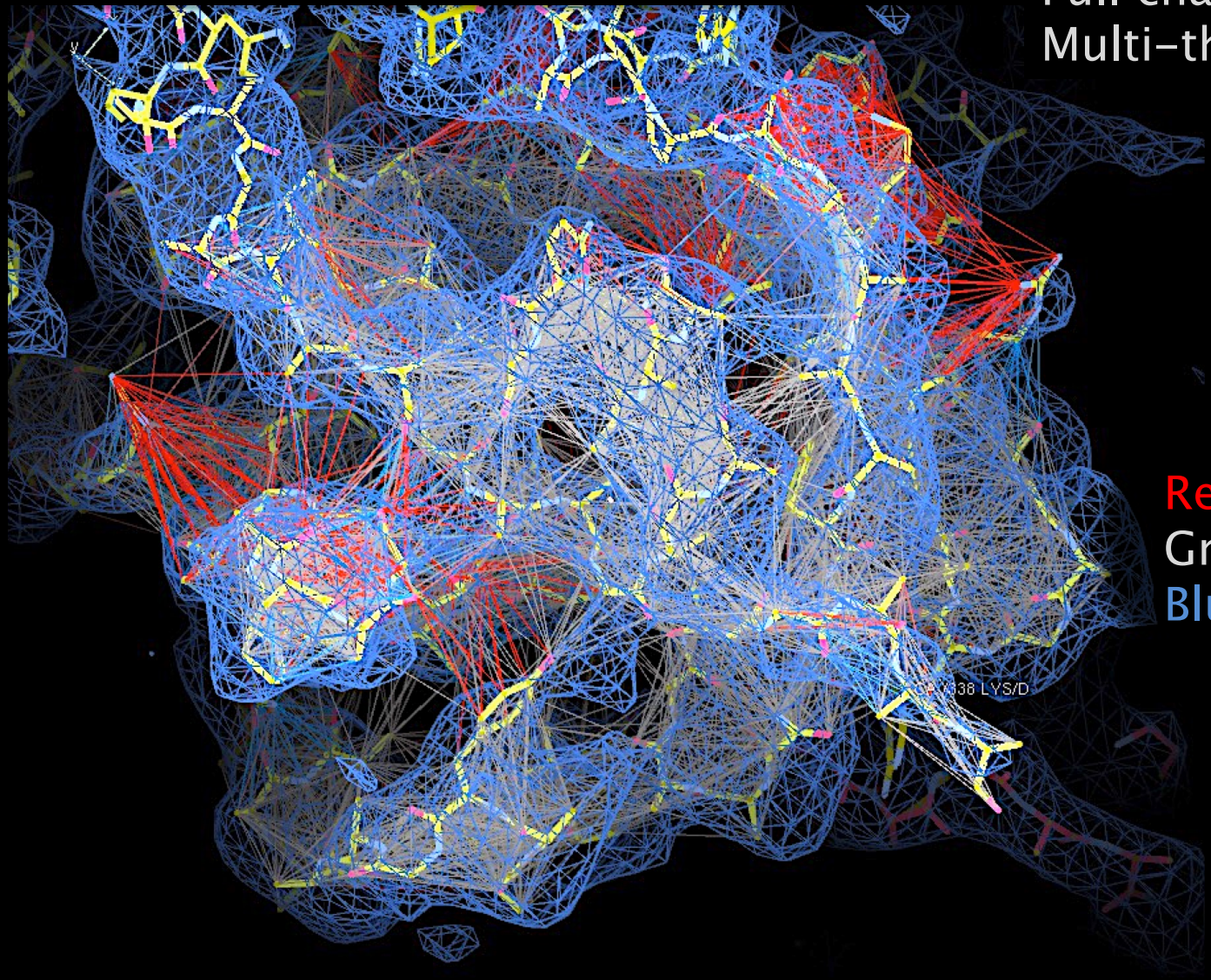
ProSMART Restraints in Coot



Full chain refine
Multi-threaded

Red: long
Grey: similar
Blue: short

ProSMART Restraints in Coot



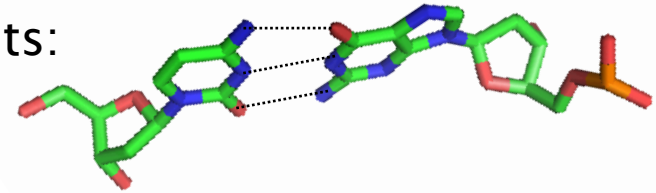
What if there are no high-resolution homologues?

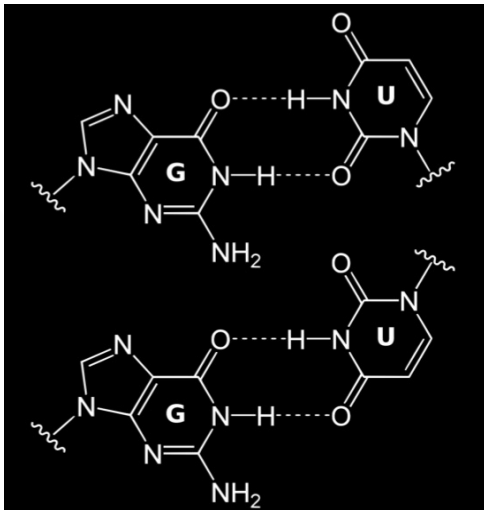
But we still need to stabilise refinement…

- Jelly-body restraints
- ProSMART – Alphafold restraints
- Generic external restraints:
  - ProSMART – protein
    (secondary structure h-bonds)
  - LibG – DNA/RNA
    (base-pair, base-stacking)

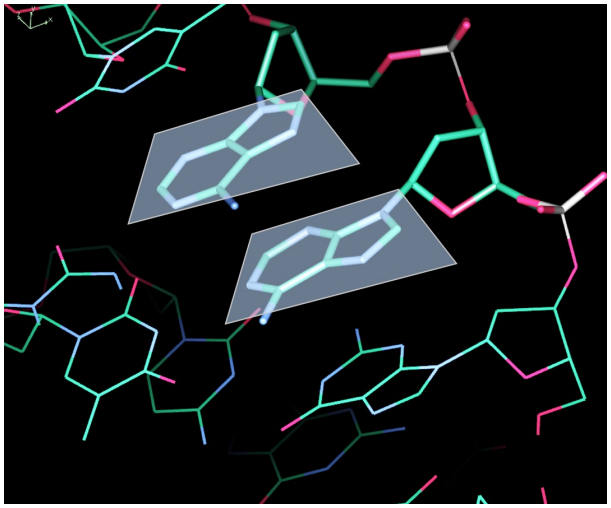
LibG Nucleic Acid Restraints

1) Base-pair restraints:

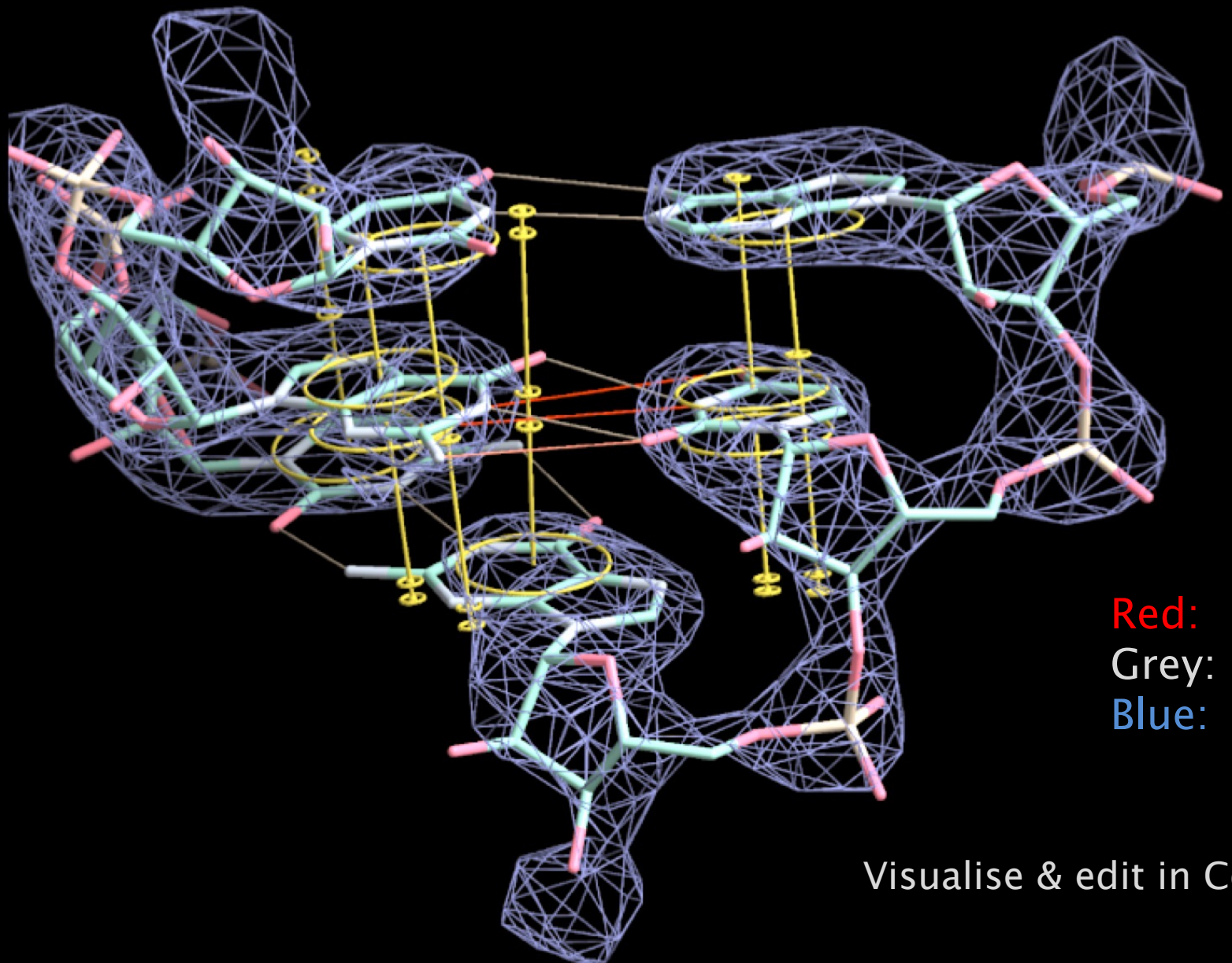




2) Parallel plane restraints:



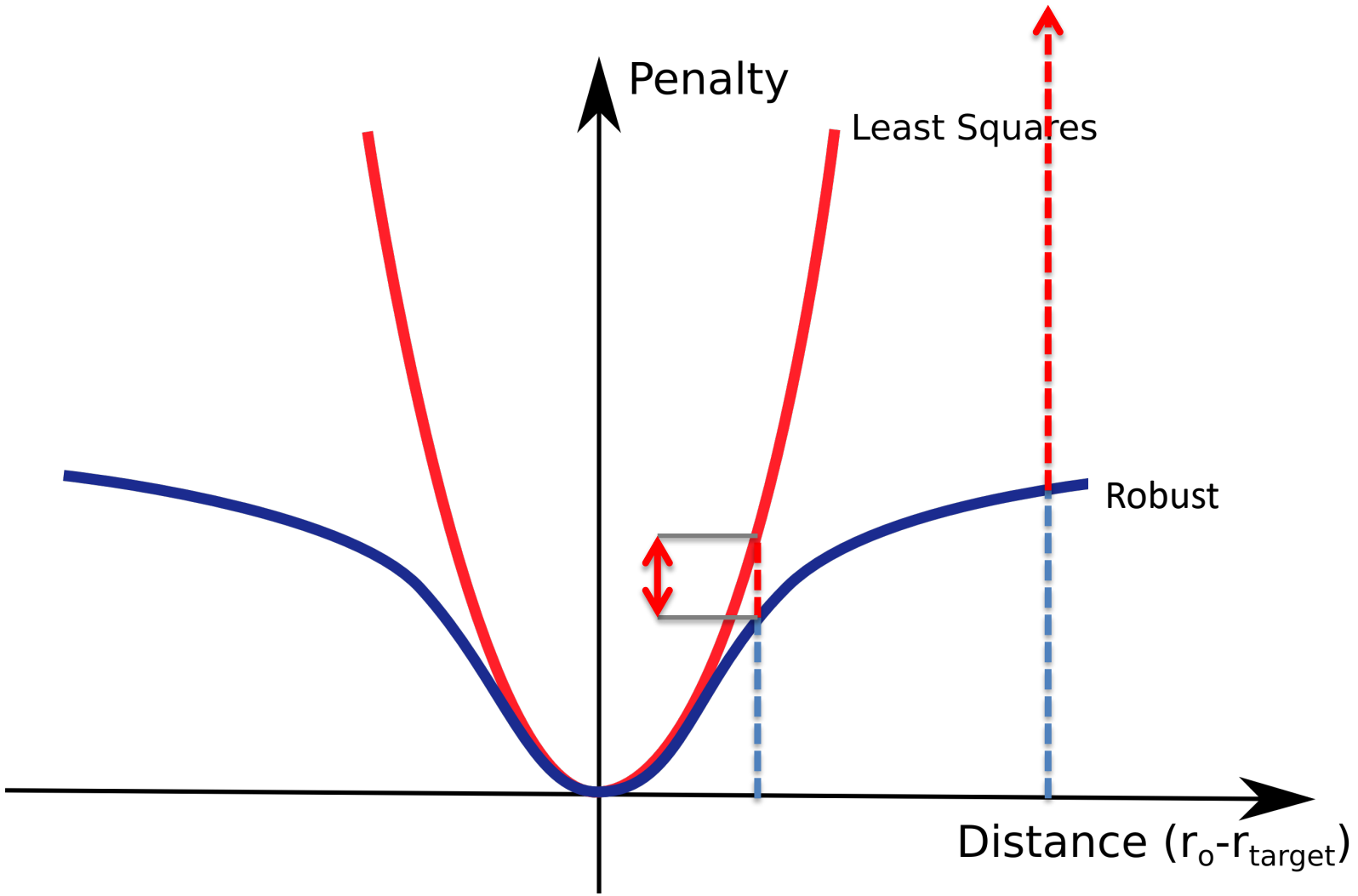
LibG Nucleic Acid Restraints



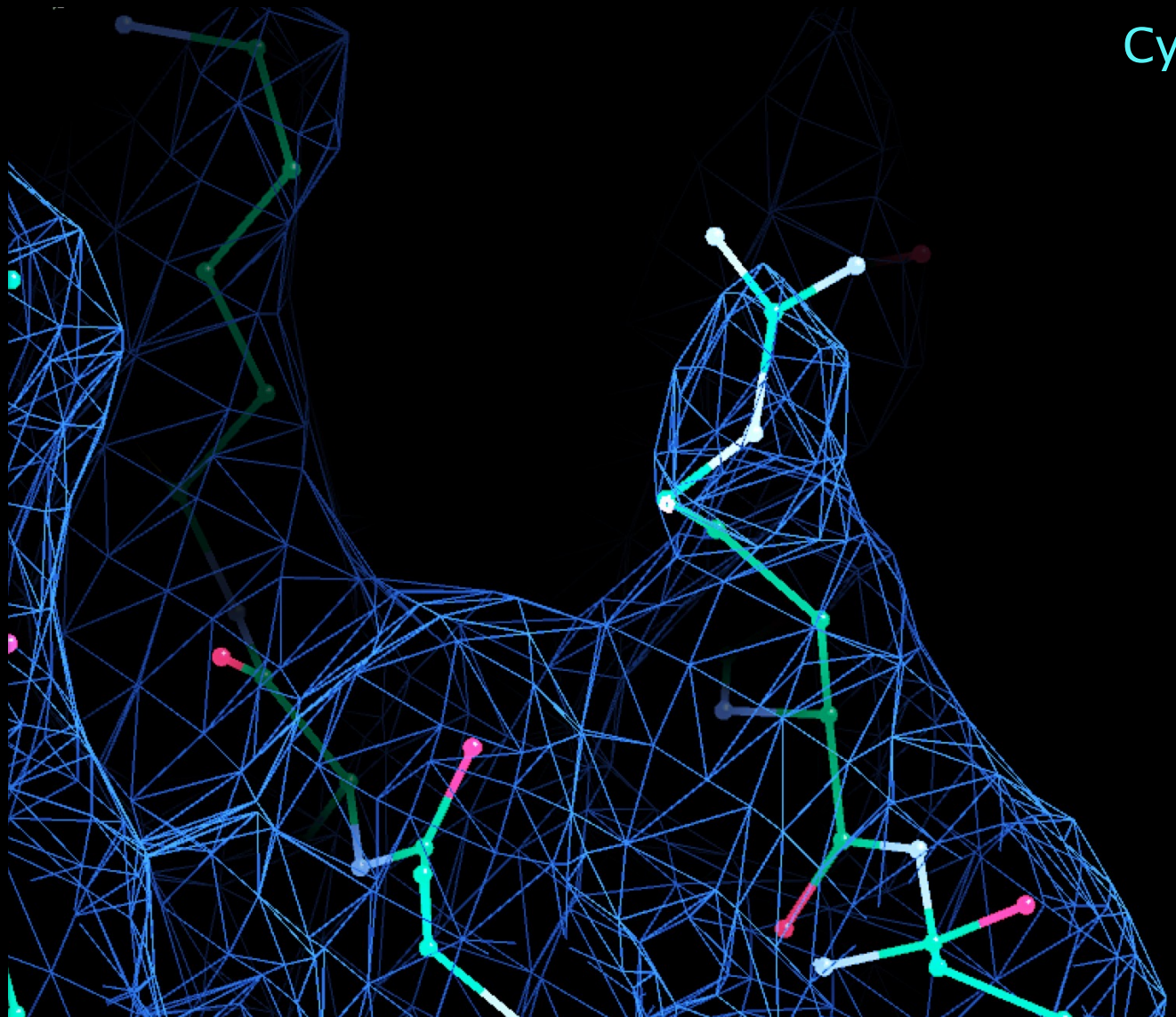
Red: long
Grey: similar
Blue: short

Visualise & edit in Coot

Robust Estimation

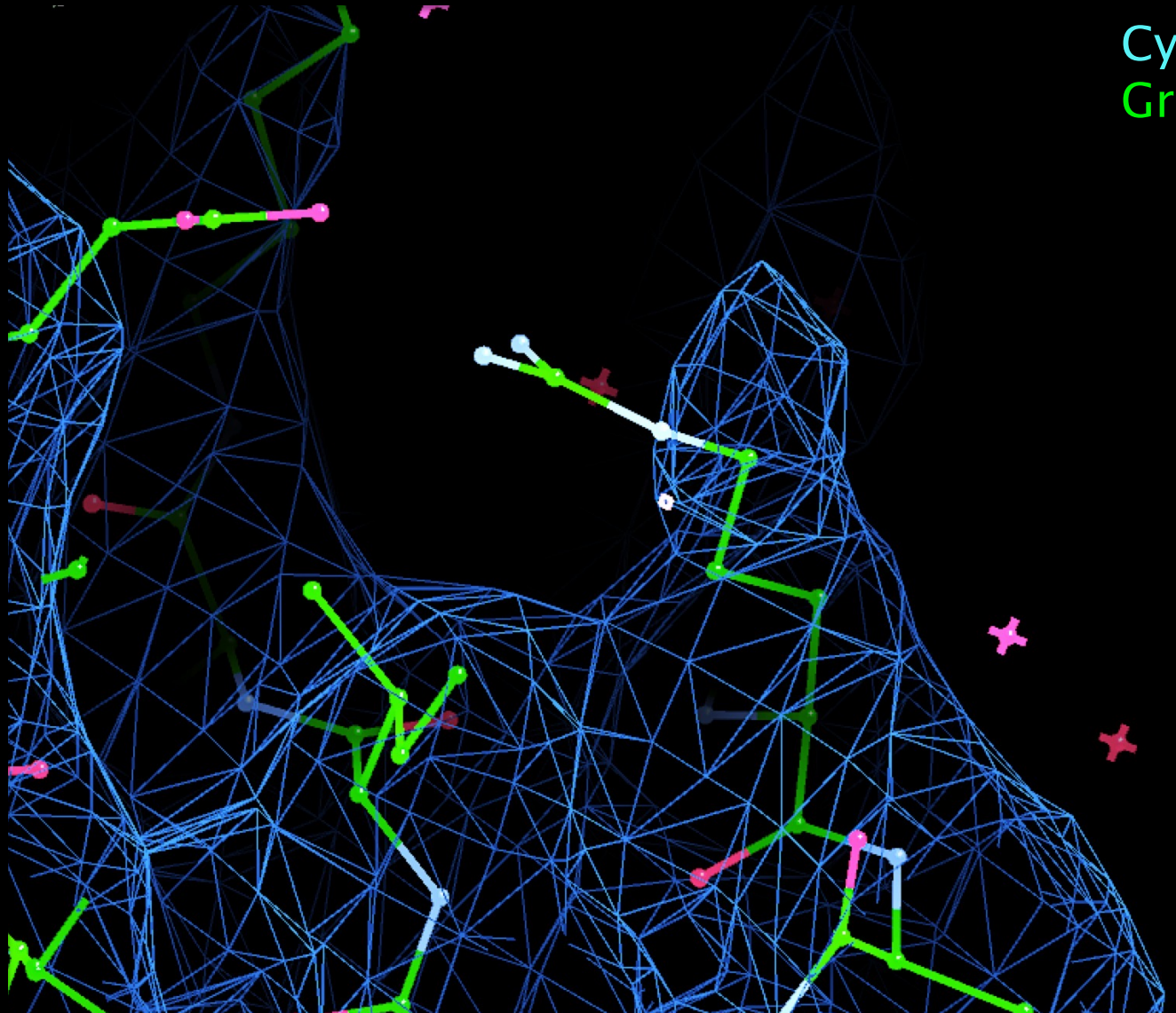


Robust Estimation



Cyan: original

Robust Estimation



Cyan: original
Green: homolog

Robust Estimation

Cyan: original
Green: homolog
Yellow: refined

Red: long
Grey: similar
Blue: short

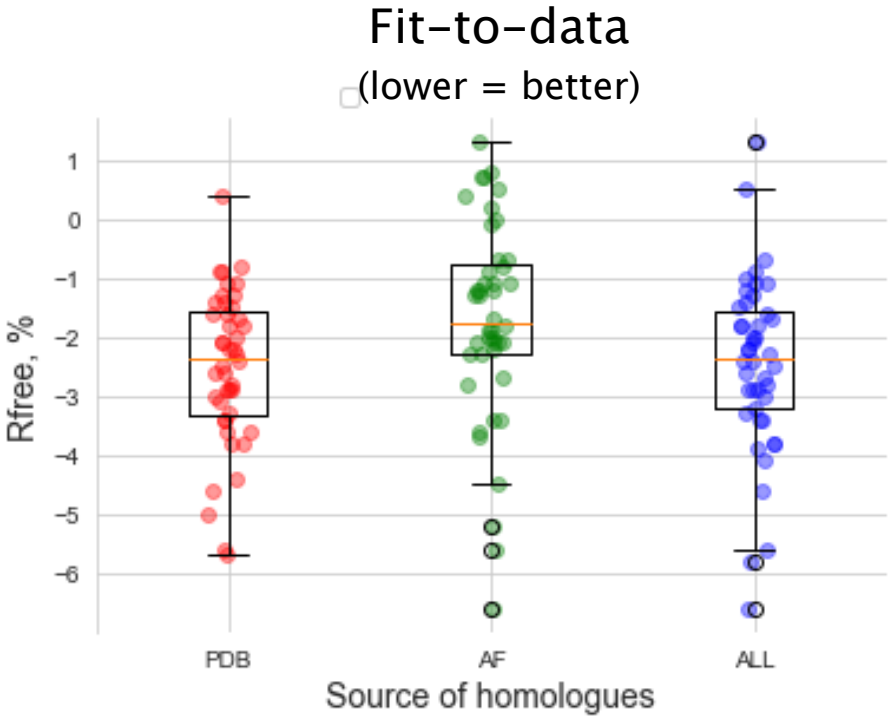
Robust Estimation

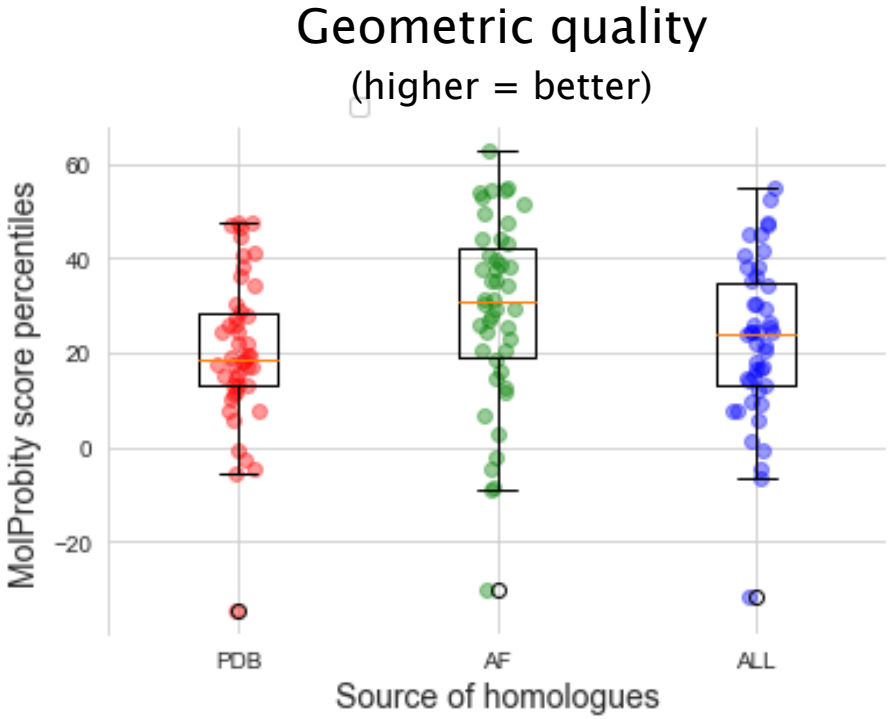
Cyan: original
Green: homolog
Yellow: refined

Red: long
Grey: similar
Blue: short

Utilising Predicted Structures

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





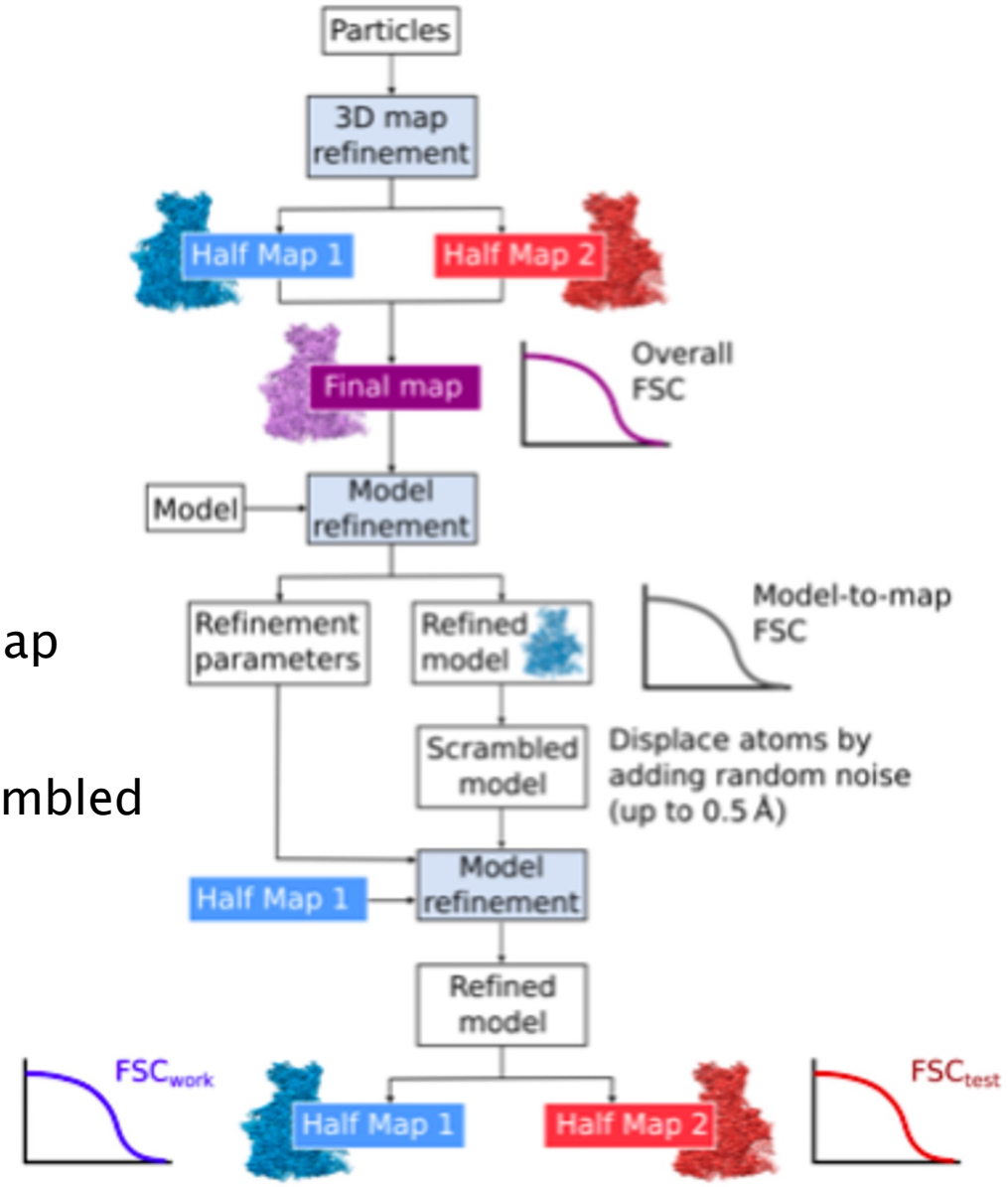
- AlphaFold2 models can be used for external restraint generation.
- Close homologues usually perform better, if available. Try both.

(unpublished; thanks to Oleg Kovalevskiy)

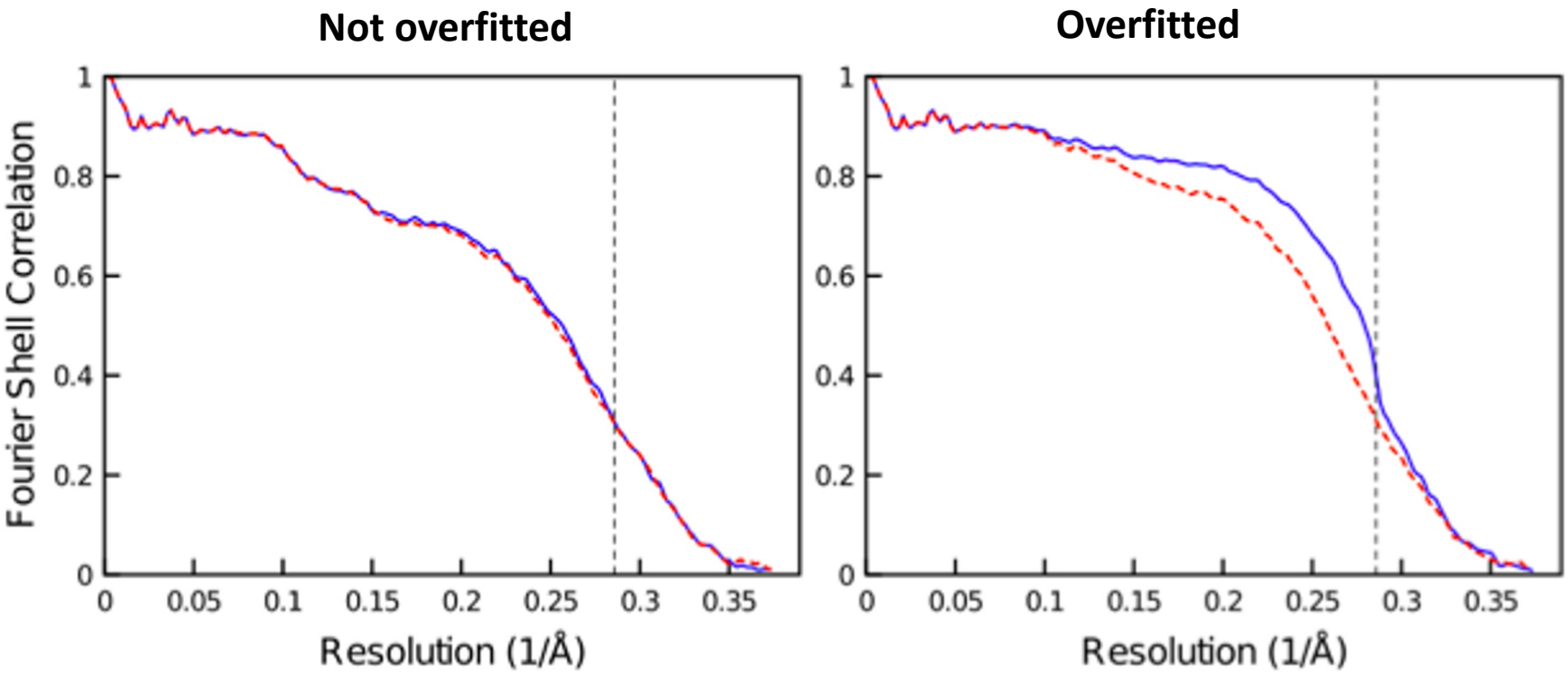
Cryo-EM Model Refinement

Half-map cross-validation:

1. Initial refinement using full map
2. Second refinement using scrambled model and half map 1
3. Compare model vs map for:
   a. Half-map 1 – work
   b. Half-map 2 – test



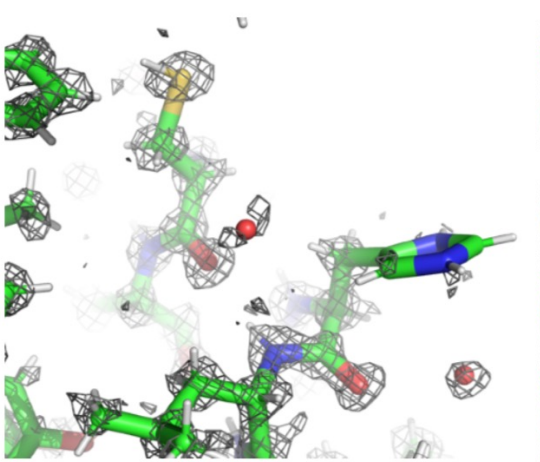
Half-Map Cross-Validation

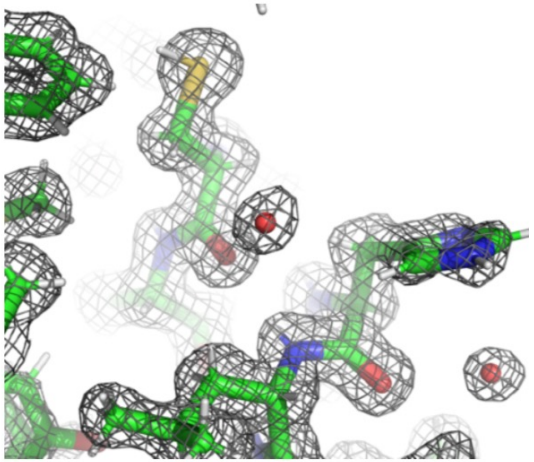


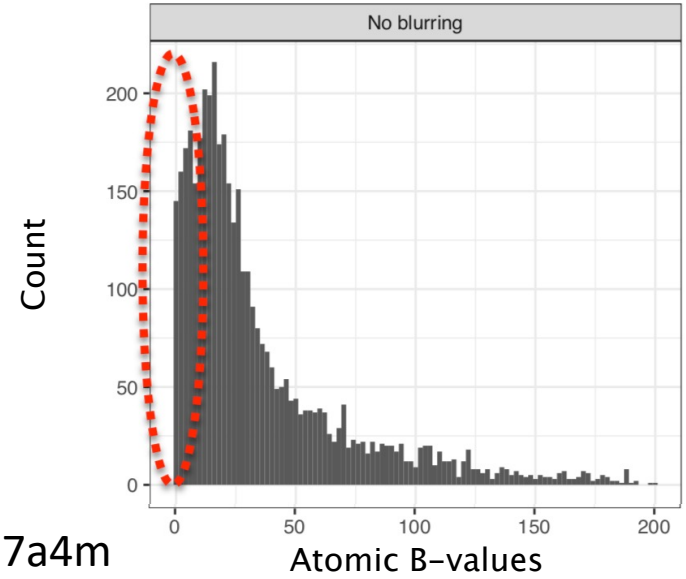
FSC_{work} : model refined against half-map 1; compared to half-map 1
FSC_{free} : model refined against half-map 1; compared to half-map 2

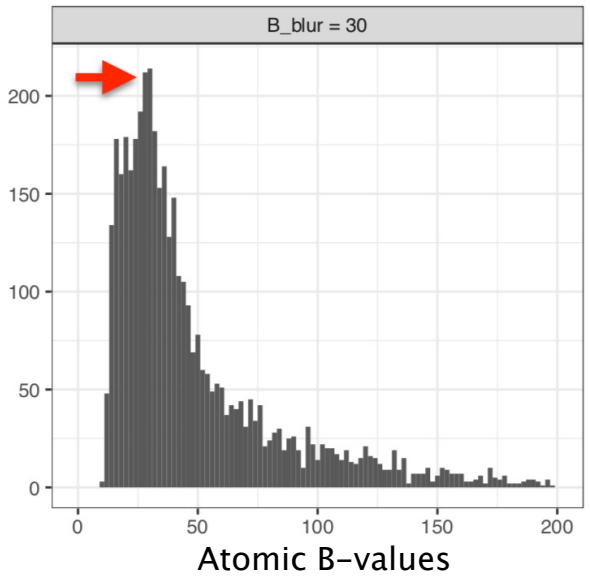
Unsharpened Unweighted Half-Maps

Maps often oversharpened – negatively affects refinement; causes incorrect B-values
So always refine against unsharpened, unweighted half-maps





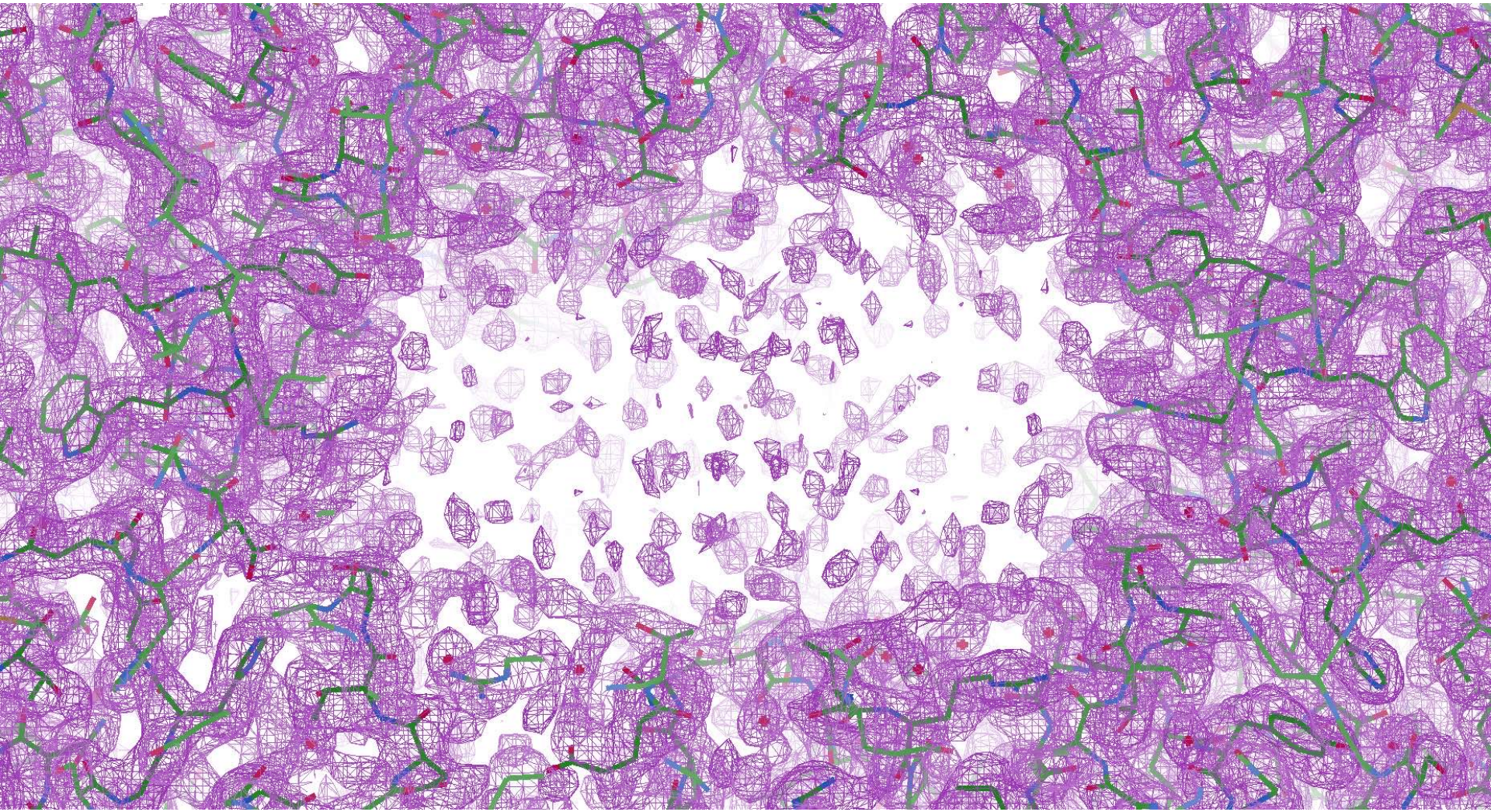




EMD-11638, 7a4m

Map Sharpening/Blurring

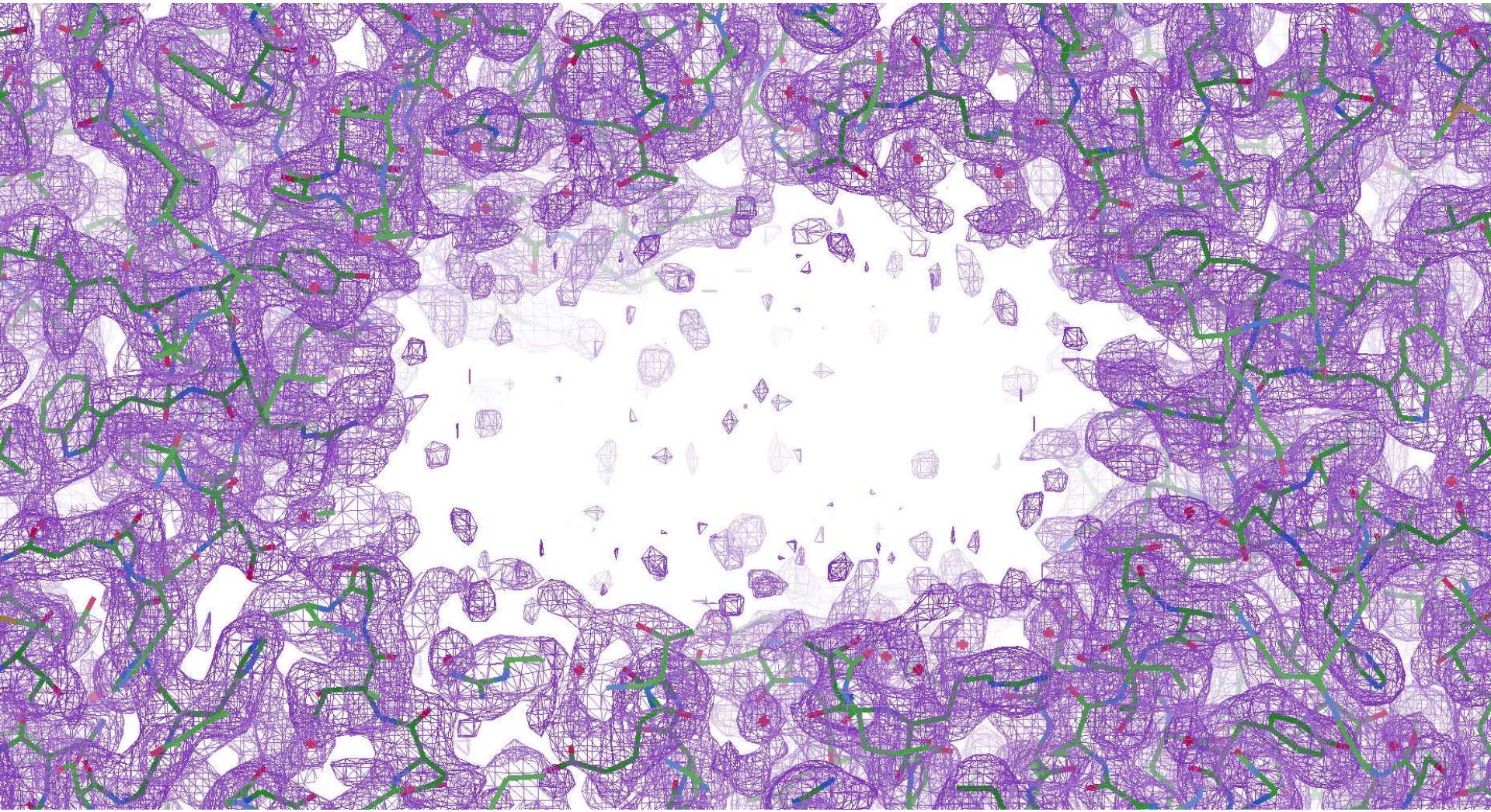
Default map **5a1a (2.2Å)**



Map Sharpening/Blurring

Blur 20 Å^2

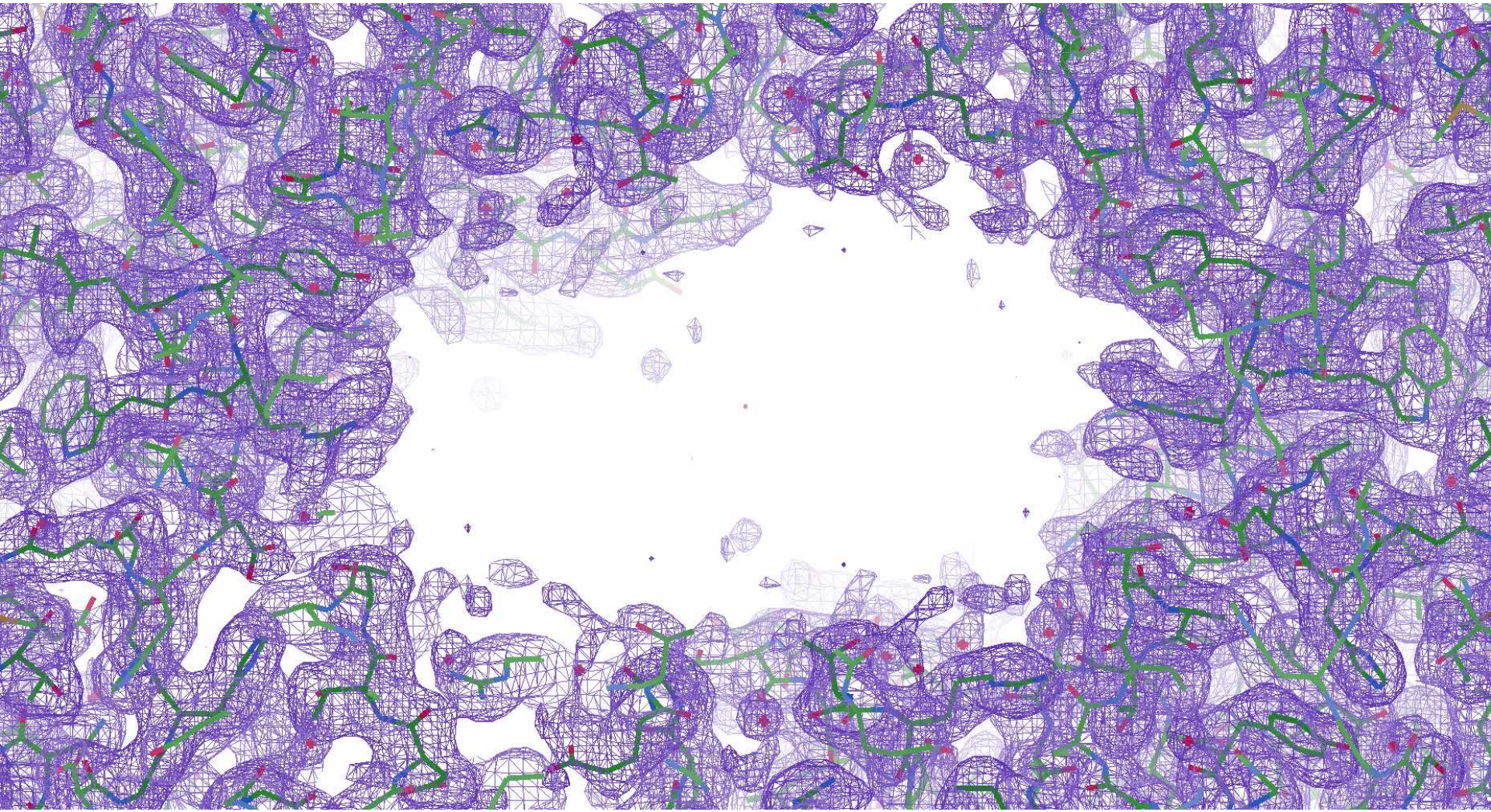
5a1a (2.2Å)



Map Sharpening/Blurring

Blur 40 Å^2

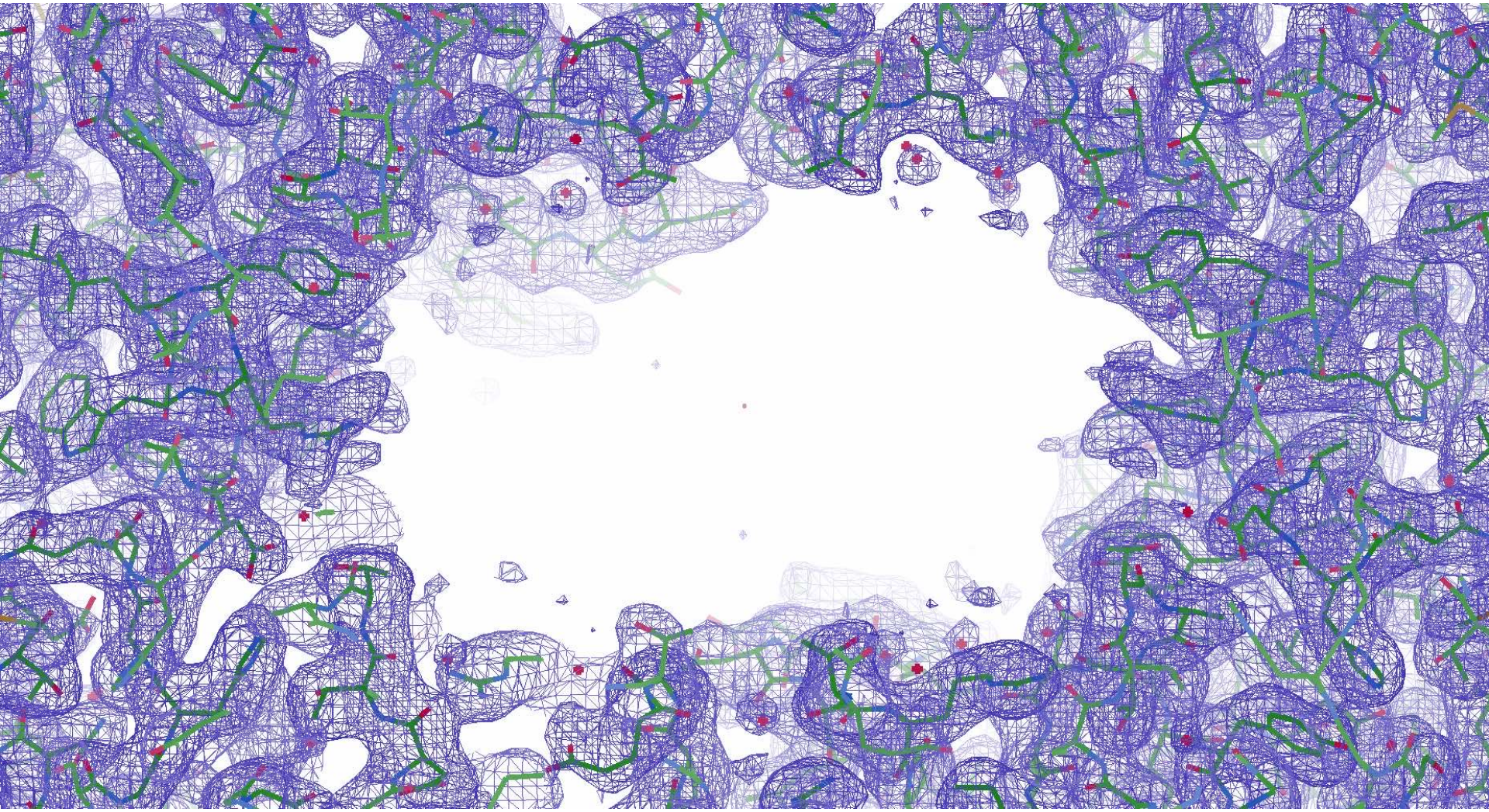
5a1a (2.2Å)



Map Sharpening/Blurring

Blur 60 Å^2

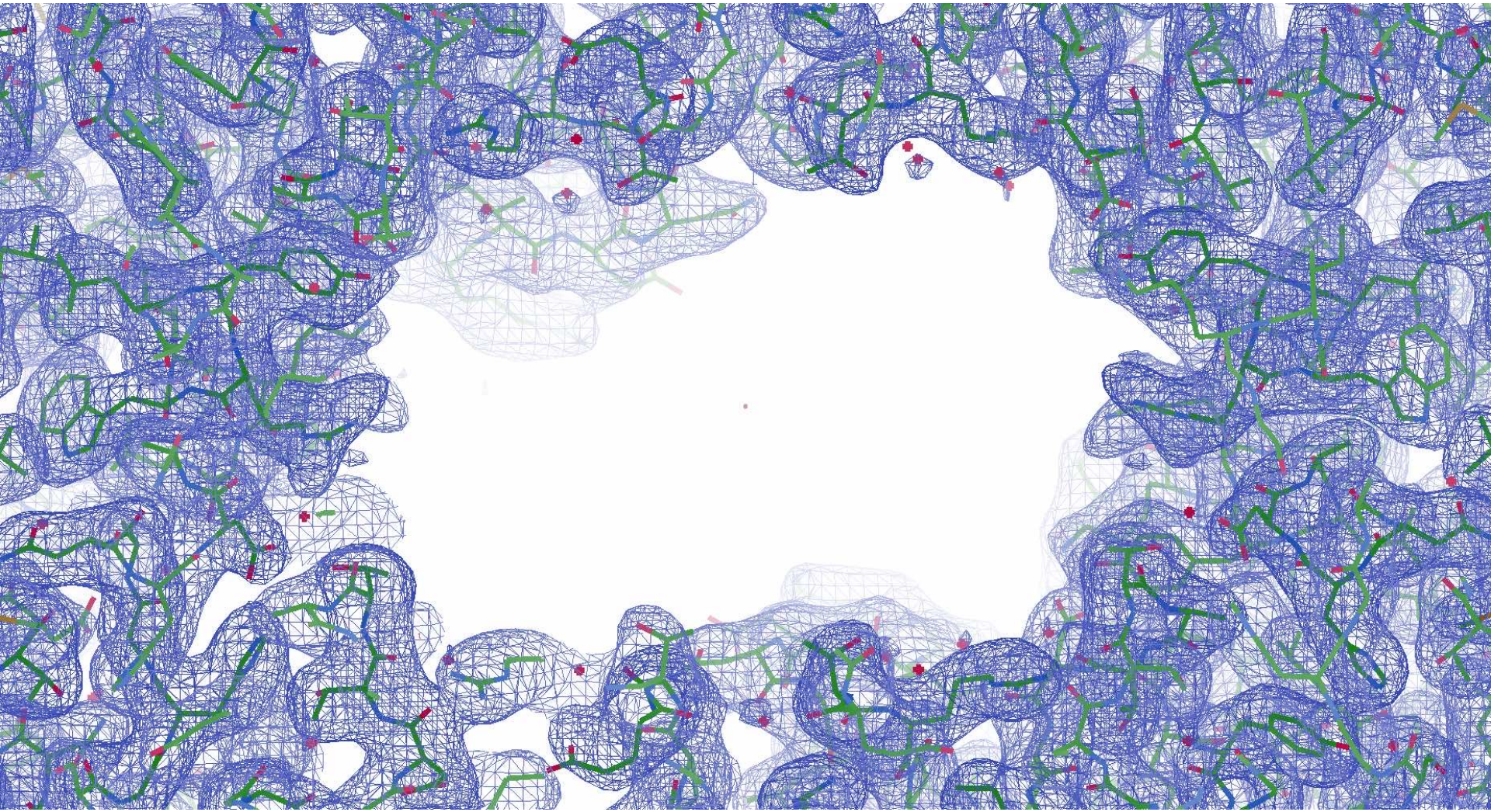
5a1a (2.2Å)



Map Sharpening/Blurring

Blur 80 Å^2

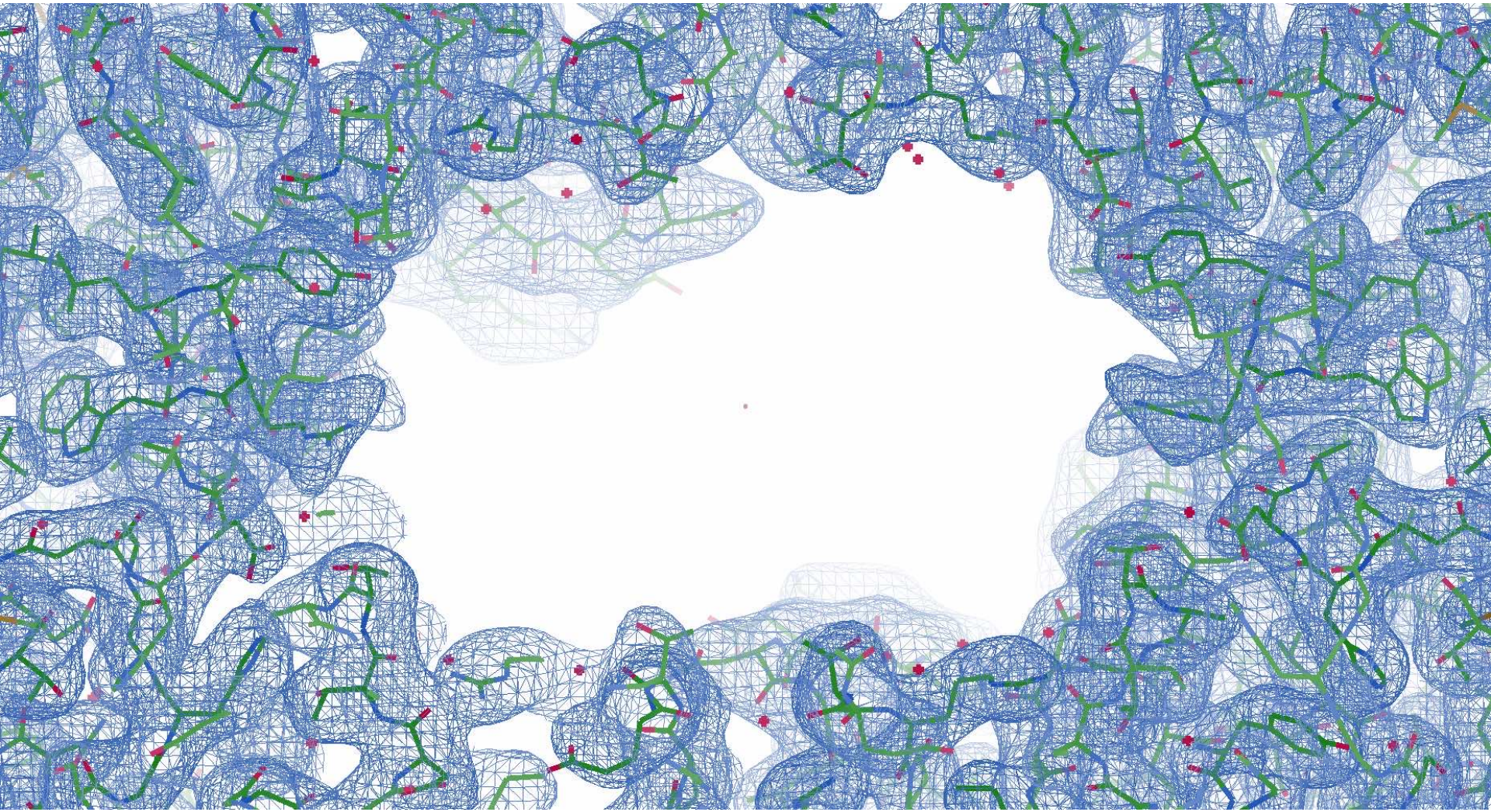
5a1a (2.2Å)



Map Sharpening/Blurring

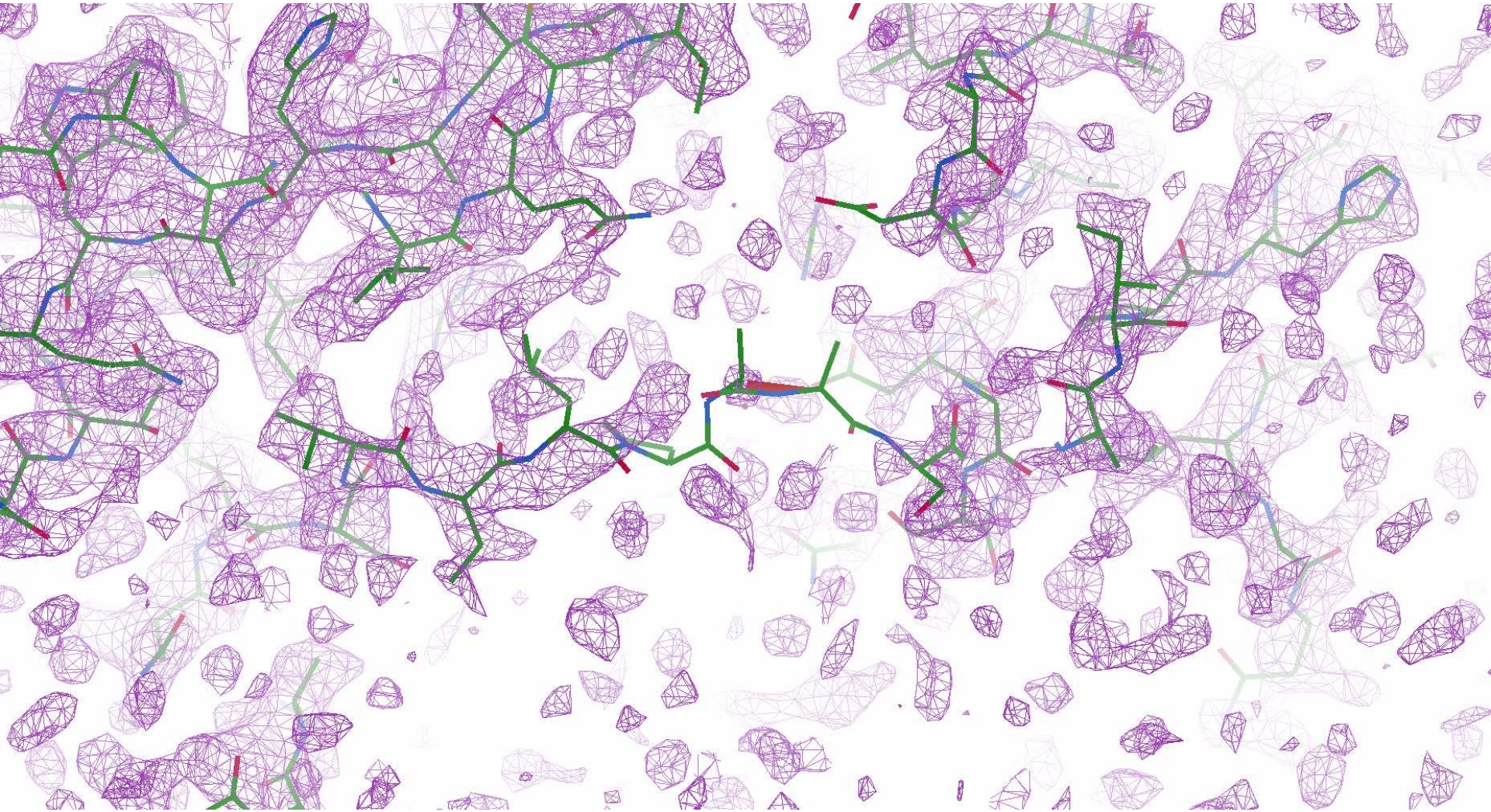
Blur 100 Å^2

5a1a (2.2Å)



Blurring / Sharpening

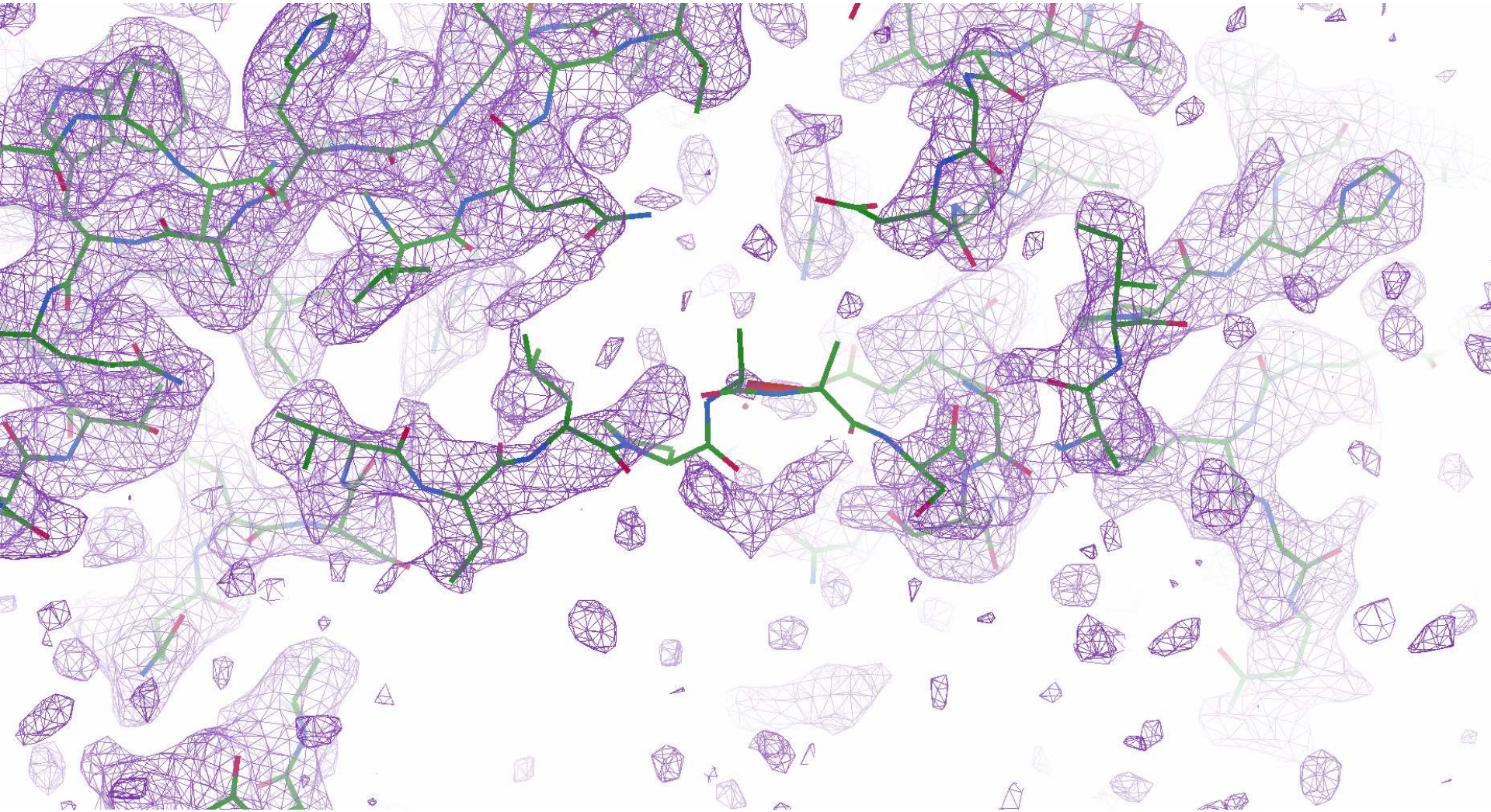
Default map **5a1a (2.2Å)**



Blurring / Sharpening

Blur 20 Å^2

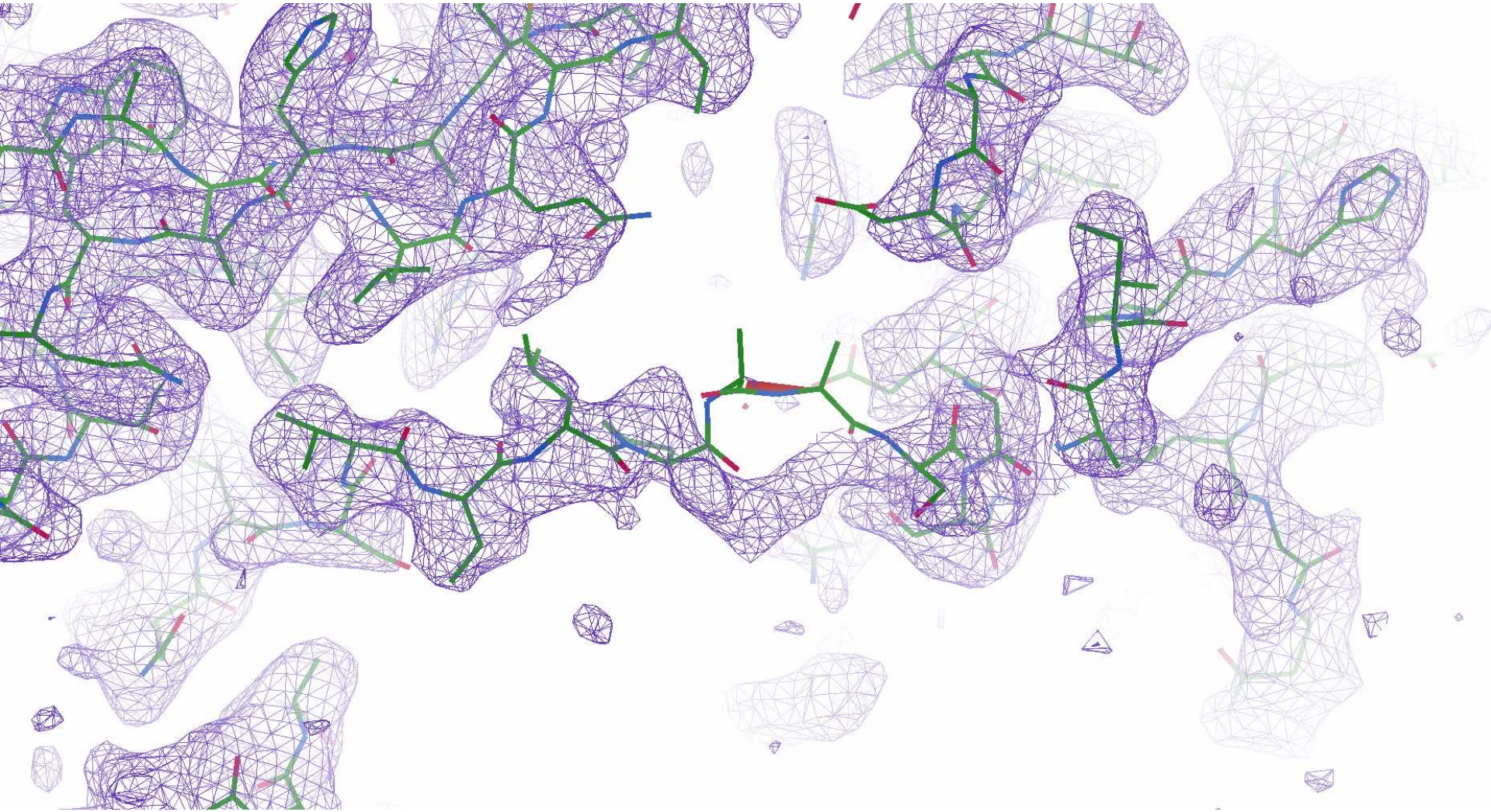
5a1a (2.2Å)



Blurring / Sharpening

Blur 40 Å^2

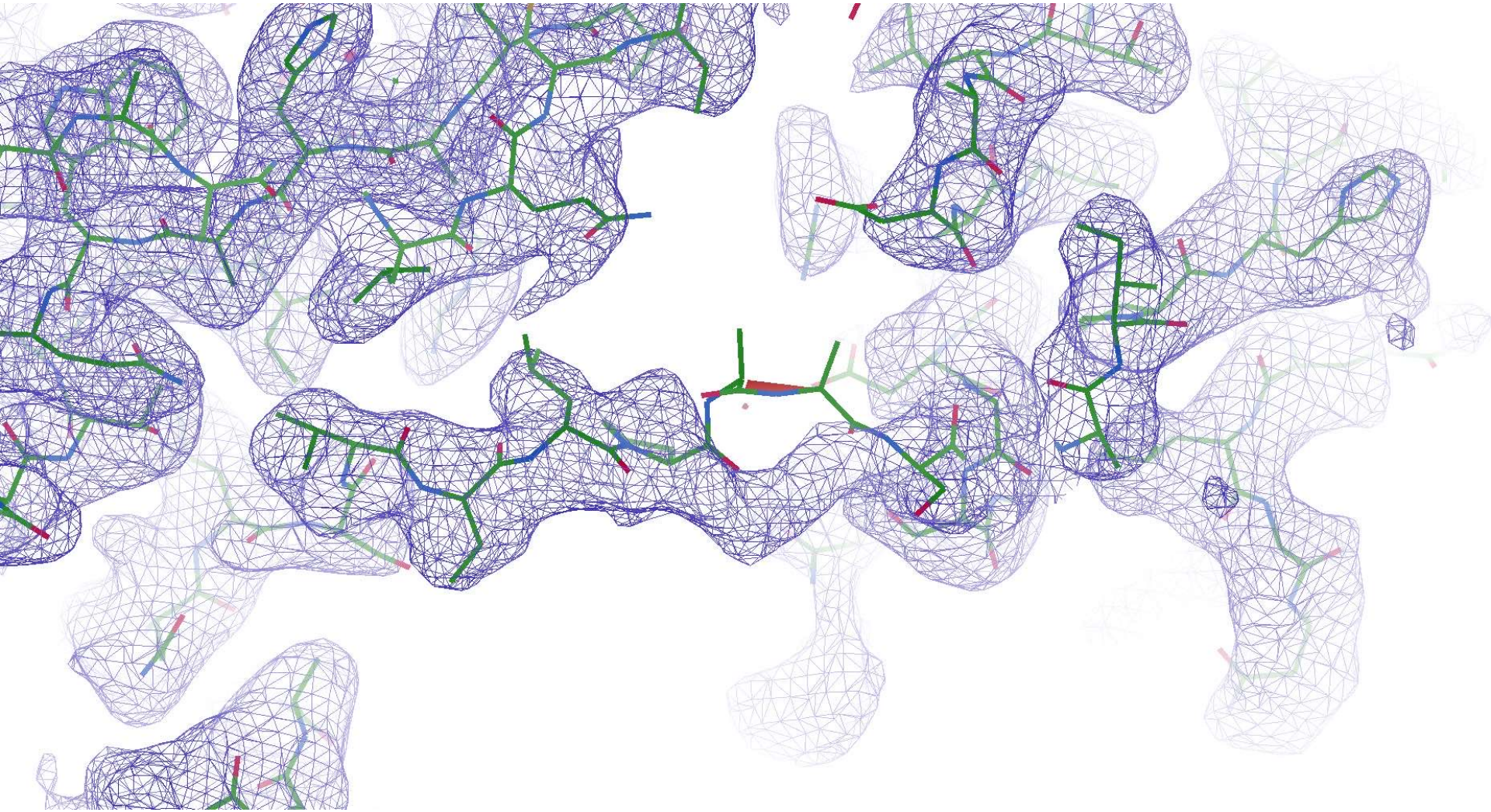
5a1a (2.2Å)



Blurring / Sharpening

Blur 60 Å^2

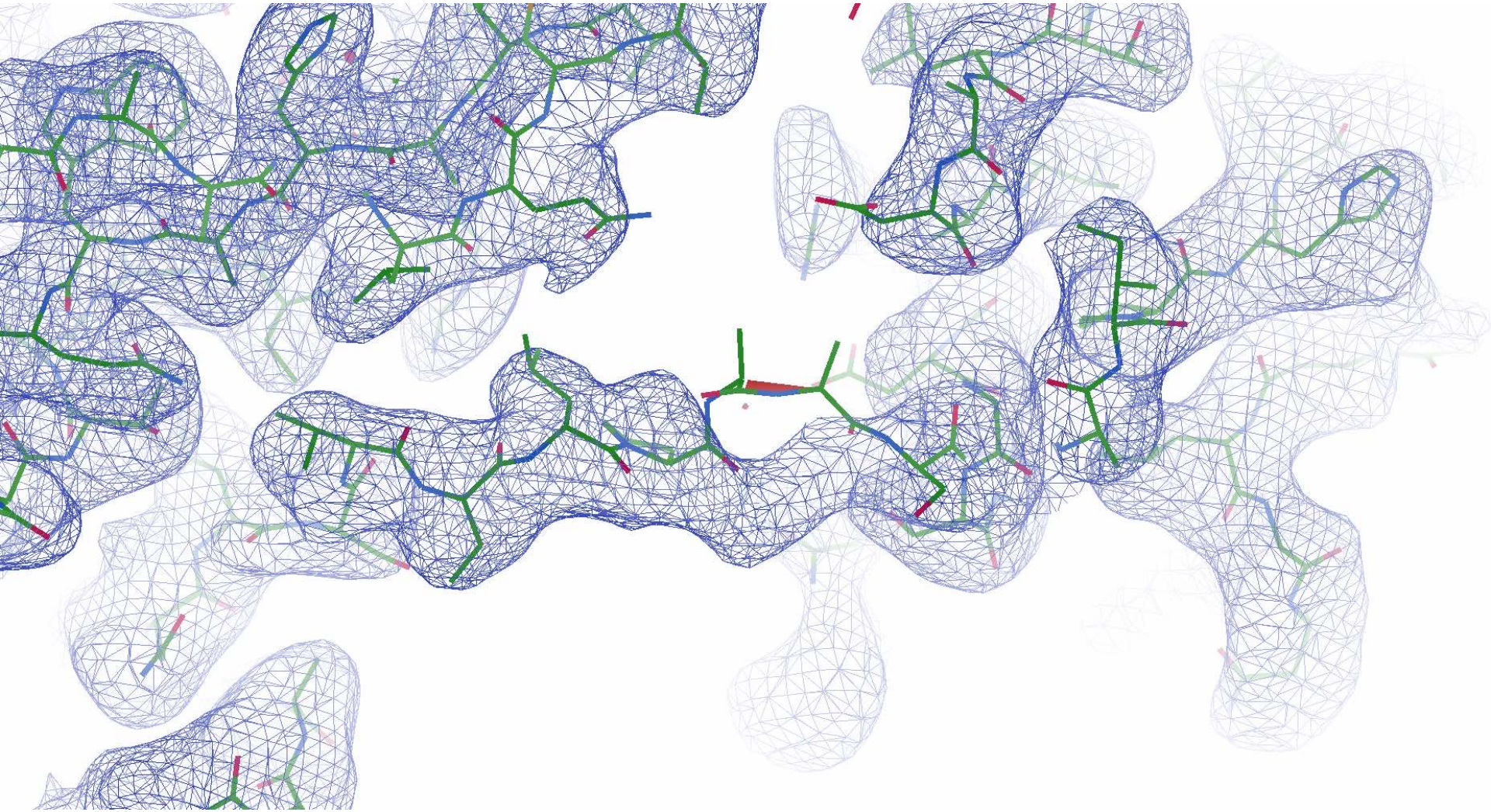
5a1a (2.2Å)



Blurring / Sharpening

Blur 80 Å^2

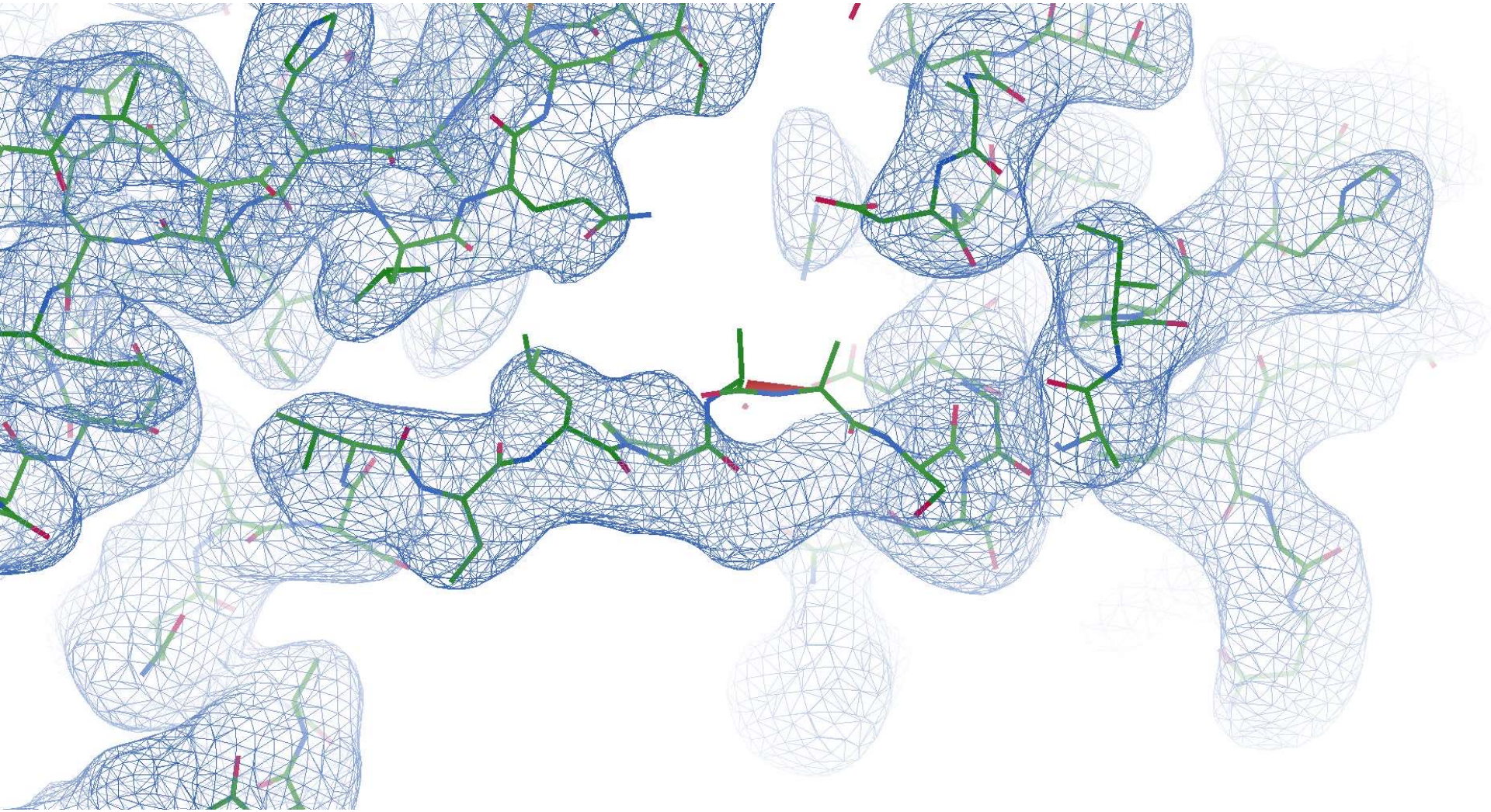
5a1a (2.2Å)



Blurring / Sharpening

Blur 100 Å^2

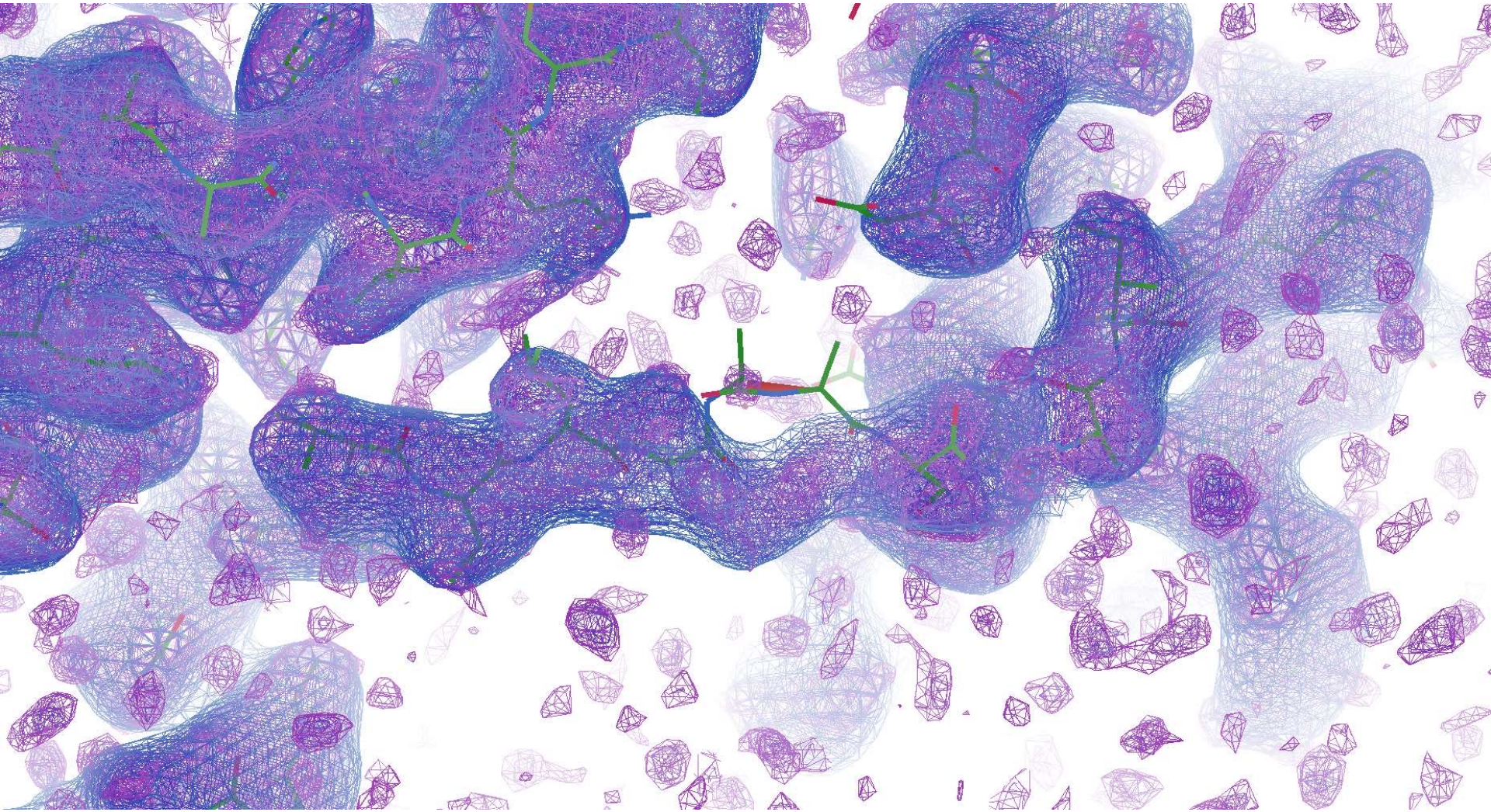
5a1a (2.2Å)



Blurring / Sharpening

Blur 0–100 Å²

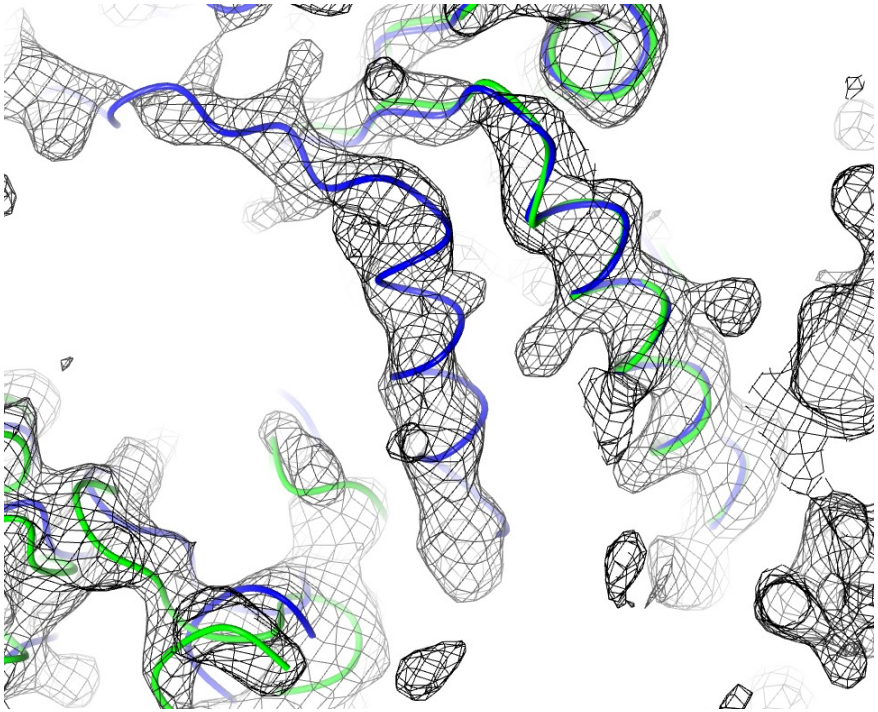
5a1a (2.2Å)



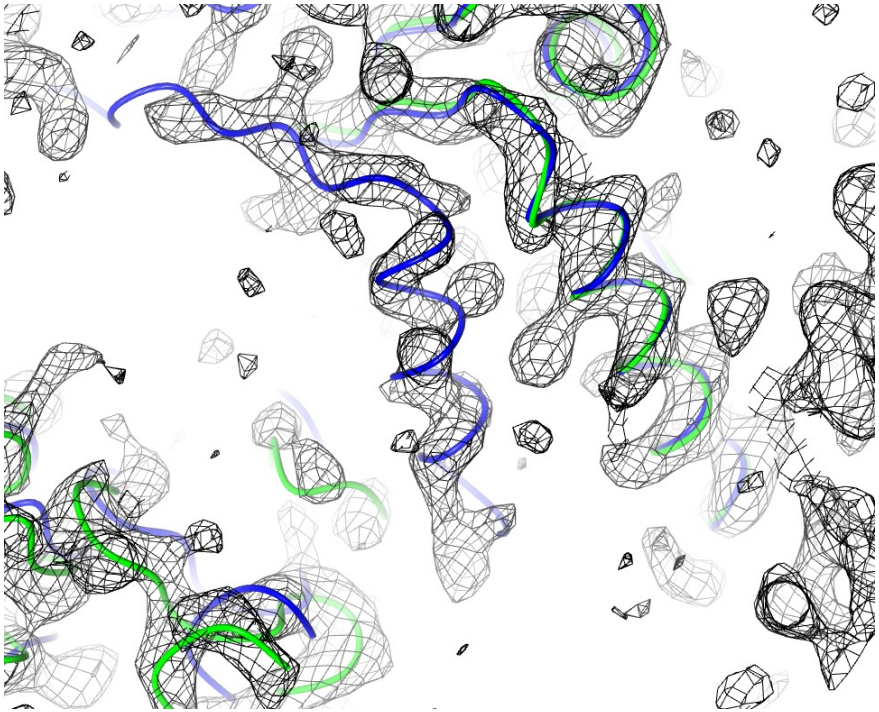
Regularised Anisotropic Map Sharpening

Idea – remove an overall B value

Original Map



Sharpened map from REFMAC5



Green: original structure 2r6c (4.0Å) – helix unmodelled

Blue: homologous structure 2r6a (2.9Å)

Blurring / Sharpening

Blurring/Sharpening is useful for visual interpretation

- In MX, map blurring/sharpening does not affect refinement

- In cryo-EM, map blurring/sharpening does affect refinement

Servalcat Difference Maps

Map Input: Unweighted half-maps – for refinement

Map Output: Weighted sharpened maps – for visualisation

Difference map calculation:

Variance of noise:

Calculated from half-maps in res bins

$$\sigma_n^2 = \frac{\mathrm{var}(F_{o1} - F_{o2})}{4}.$$

Used to calculate SF weightings:

Proportion of unexplained signal vs signal+noise

(unexplained signal estimated using ML)

$$w = \frac{\sigma_{U,T}^2}{\sigma_{U,T}^2 + \sigma_n^2}.$$

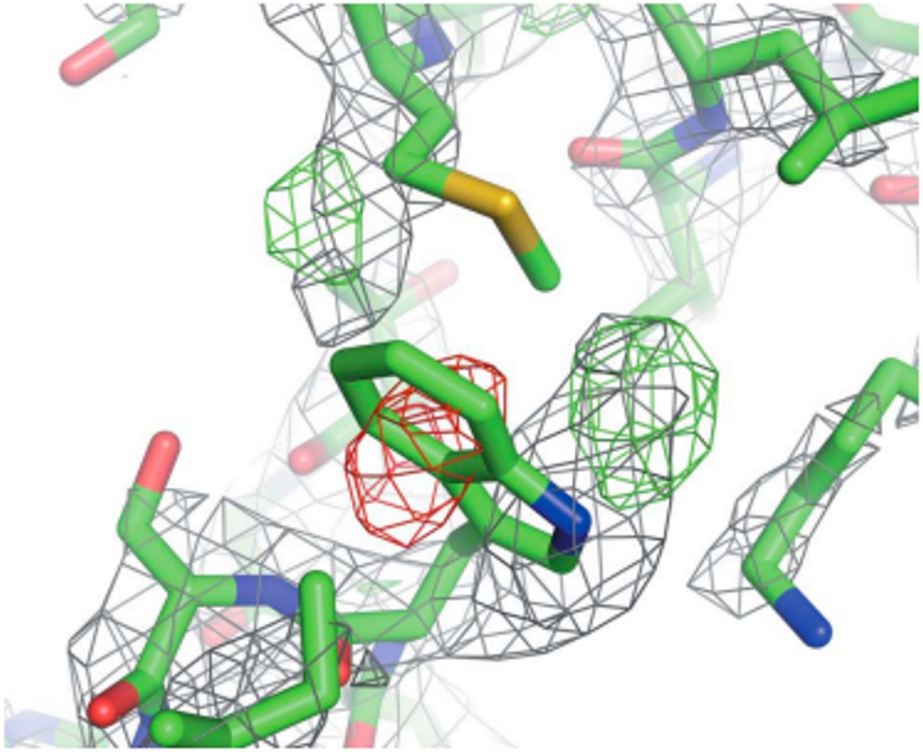
Final scaling sharpens the map (~removes B):

$$F_{diff} = \frac{w}{(\mathrm{FSC}_{full}\langle|F_o|^2\rangle)^{1/2}}(F_o - DF_c).$$

Yamashita *et al.* (2021) Cryo-EM single-particle structure refinement and map calculation using Servalcat. Acta Cryst D77, 1282-91.

Servalcat Difference Maps

Model Improvement – F_o–F_c map:

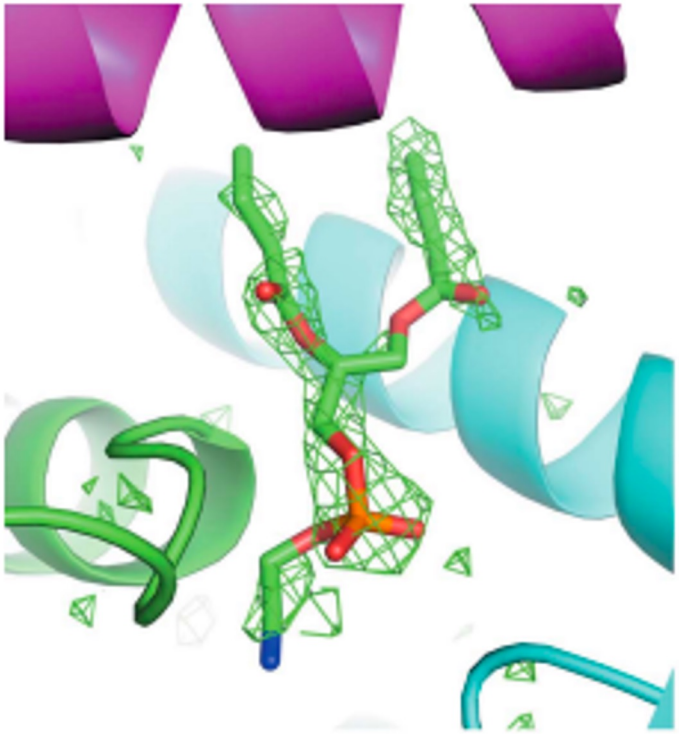


Killifish Calcium homeostasis modulator

EMD-0919 (2.7 Å)

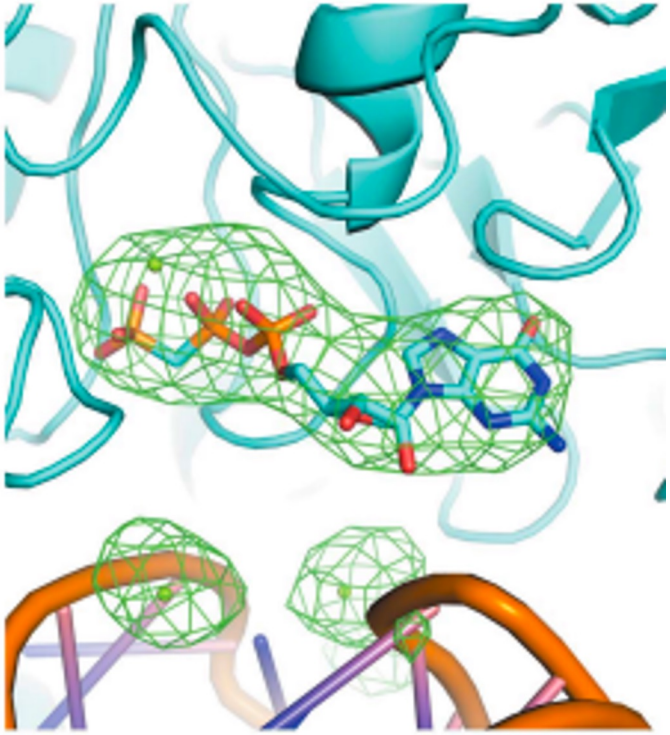
Servalcat Difference Maps

Ligand Inspection – F_o–F_c omit maps:



SARS-CoV-2 3a ion channel

EMD-22898 (2.1 Å)

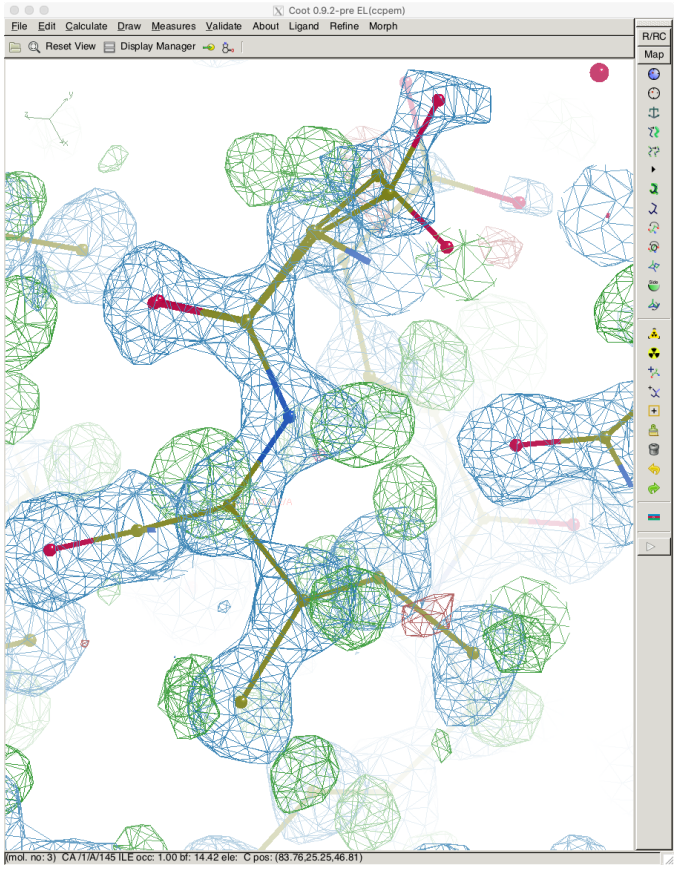


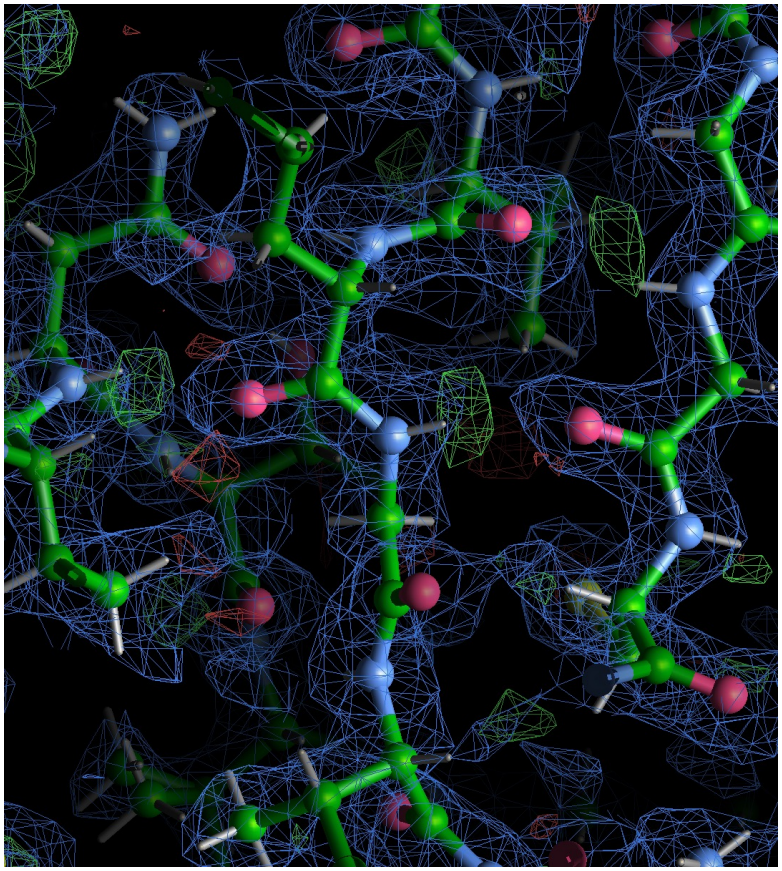
Kluyveromyces lactis 80S ribosome

EMD-8123 (3.6 Å)

Servalcat Difference Maps

Hydrogen inspection – F_o–F_c hydrogen omit map:





Apoferritin

EMD-11638, PDB: 7a4m (1.22 Å)

β-galactosidase

EMD-10109, PDB: 6s6z (2.0 Å)

EMDA Validation Tools

Fourier Shell Correlation (FSC) calculation: *EMDA internally masks around the input map*

- Half map FSC
- Map-to-map FSC
- Map-to-model FSC

Local real space correlation: *Correlation within a sphere*

- Map-to-map *E.g. "full map correlation" – between half maps*
- Map-to-model

Map calculation: *Involves likelihood-based scaling*

- Map averaging
- Difference map calculation

Other utilities:

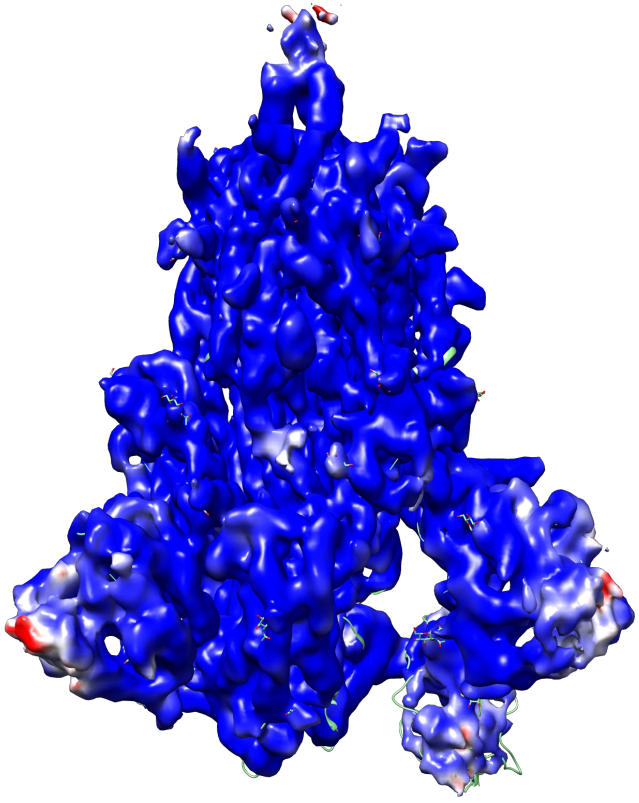
- Map fitting
- Magnification refinement

EMDA Local Real Space Correlation

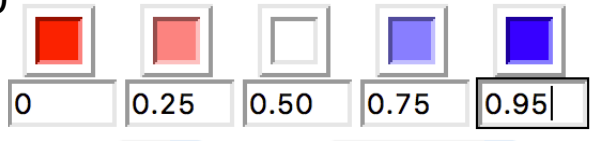
Visualised by colouring the "standard" full map by calculated correlation

Full map correlation

local correlation between half maps



Average CC: 0.70



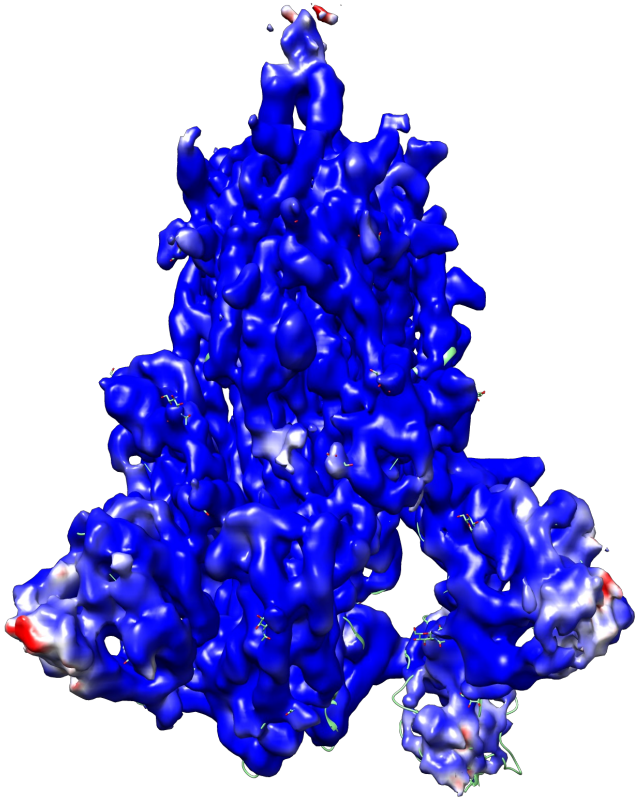
EMD-22162 (4.0 Å)

EMDA Local Real Space Correlation

Visualised by colouring the "standard" full map by calculated correlation

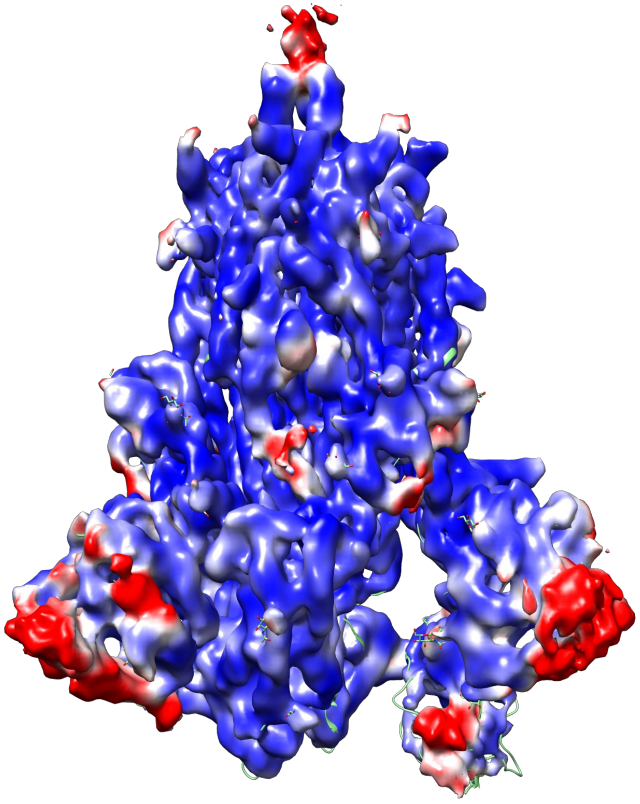
Full map correlation

local correlation between half maps



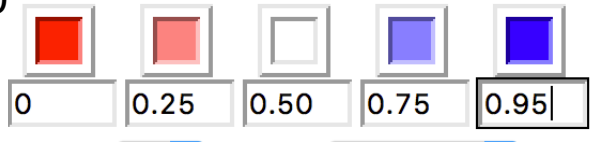
Average CC: 0.70

Map-model correlation



Average CC: 0.48

EMD-22162 (4.0 Å)

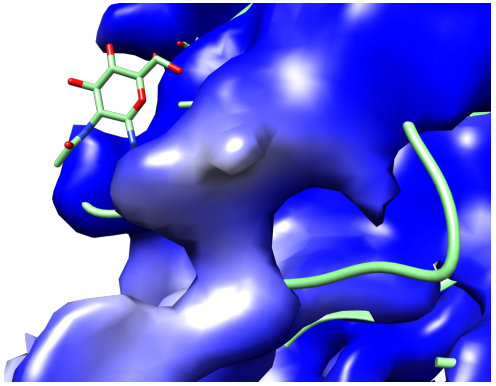


EMDA Local Real Space Correlation

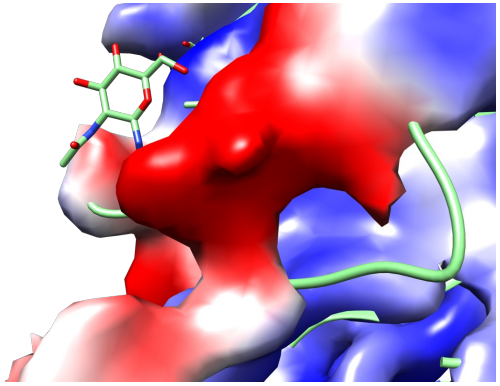
Visualised by colouring the "standard" full map by calculated correlation

Full map correlation

local correlation between half maps

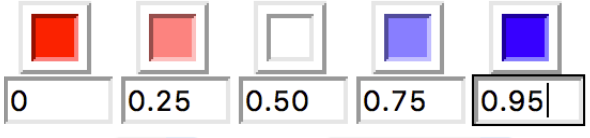


Map-model correlation



C/135 - 140

EMD-22162 (4.0 Å)

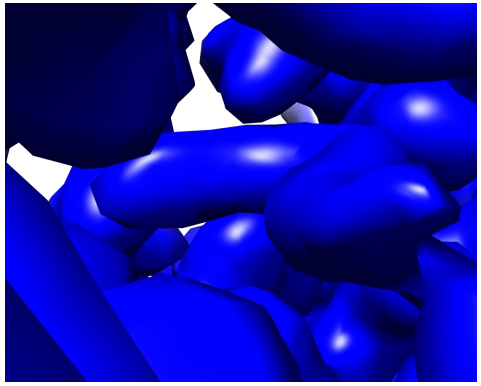


EMDA Local Real Space Correlation

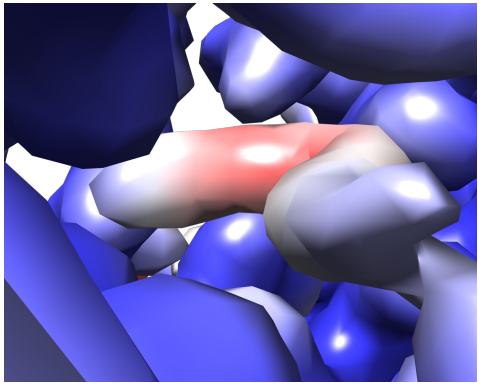
Visualised by colouring the "standard" full map by calculated correlation

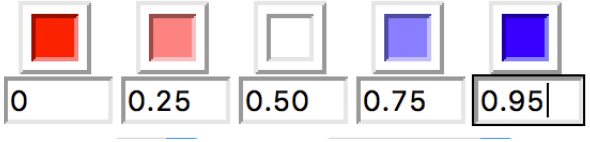
Full map correlation

local correlation between half maps



Map-model correlation





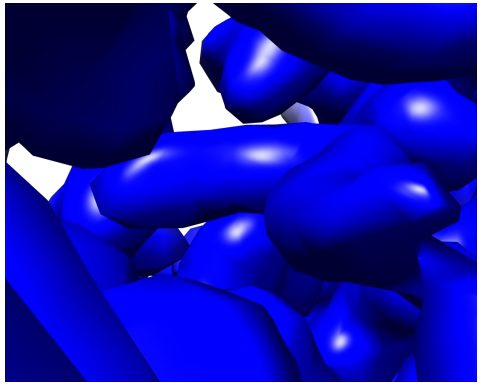
EMD-11203 (2.6 Å)

EMDA Local Real Space Correlation

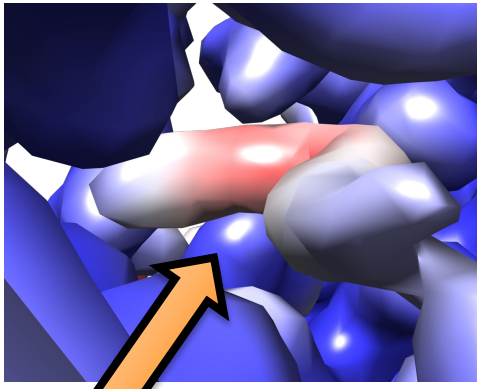
Visualised by colouring the "standard" full map by calculated correlation

Full map correlation

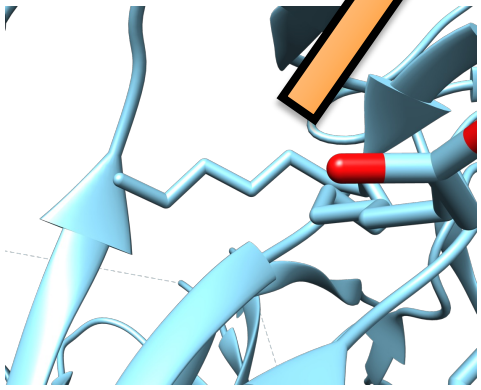
local correlation between half maps



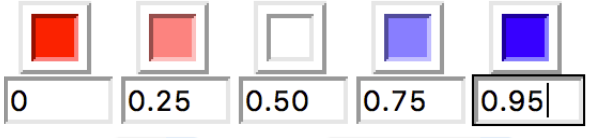
Map-model correlation



Homolog: 6z5b (1.9 Å)
Linoleic acid



EMD-11203 (2.6 Å)

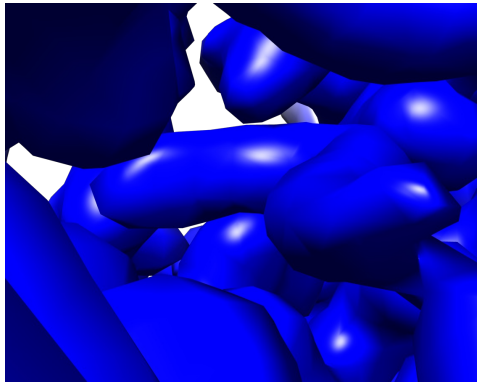


EMDA Local Real Space Correlation

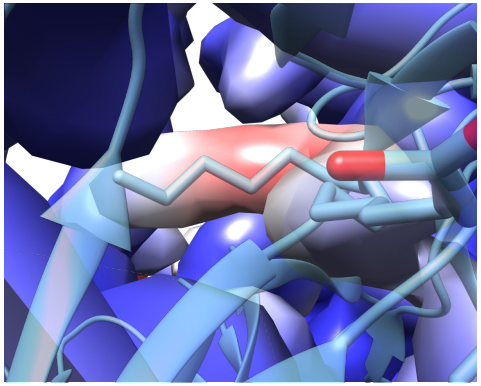
Visualised by colouring the "standard" full map by calculated correlation

Full map correlation

local correlation between half maps



Map-model correlation



Homolog 6z5b superposed
Linoleic acid seems to fit

EMD-11203 (2.6 Å)

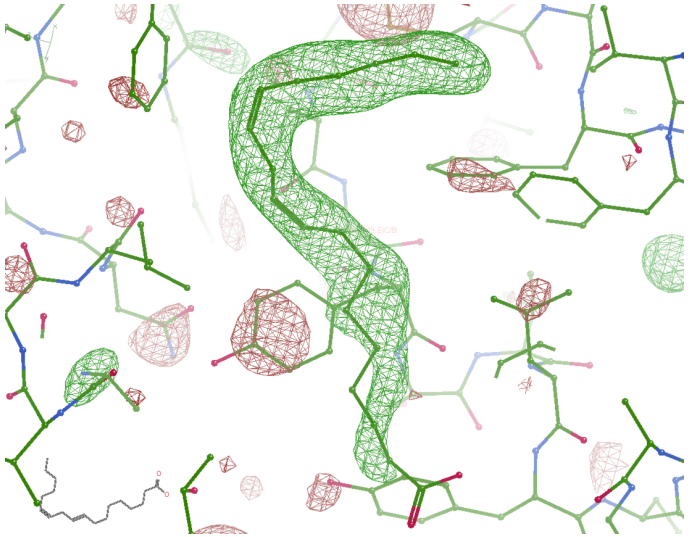
EMDA Difference Map Calculation

Aim: to represent differences between a map and a model

- Analogous to the F_o-F_c map in MX
- Structure factors are normalised so that they are on the same scale (by default)
- Structure factors are weighted according to FSC – avoid high-resolution noise
- Can weight according to the signal-to-noise ratio (requires half maps)

Difference map between
EMD-11203 and PDB-6zge

Homolog 6z5b superposed



Linoleic acid

EMD-11203 (2.6 Å)

Importance of Phases

"Phases are more important than amplitudes"

ρ_C = current map

ρ_t = ideal (true) map

$$\mathrm{cor}(\rho_t, \rho_C) = \mathrm{cor}(F_t, F_C) = \frac{\sum |F_t||F_C| \cos(\varphi_t - \varphi_C)}{\sqrt{\sum |F_t|^2 \sum |F_C|^2}}$$

Importance of Phases

"Phases are more important than amplitudes"

ρ_C = current map

ρ_t = ideal (true) map

$$\mathrm{cor}(\rho_t, \rho_C) = \mathrm{cor}(F_t, F_C) = \frac{\sum |F_t||F_C| \cos(\varphi_t - \varphi_C)}{\sqrt{\sum |F_t|^2 \sum |F_C|^2}}$$

In a given shell: $$FSC = \frac{1}{N} \sum |E_t||E_C| \cos(\varphi_t - \varphi_C)$$

FSC : Fourier Shell Correlation

$$FSC = \left\langle |E_t||E_C| \cos(\varphi_t - \varphi_C) \right\rangle$$

(under certain assumptions)

Importance of Phases

Case 1: *phases are random; amplitudes are exact:*

$$FSC = \left\langle |E_t||E_C|\cos(\varphi_t - \varphi_C) \right\rangle \rightarrow 0$$

$$= 0$$

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

Importance of Phases

Case 1: *phases are random; amplitudes are exact:*

$$FSC = \left\langle |E_t||E_C|\cos(\varphi_t - \varphi_C) \right\rangle \rightarrow 0$$
$$= 0$$

Case 2: *phases are exact; amplitudes are random:*

$$FSC = \left\langle |E_t||E_C|\cos(\varphi_t - \varphi_C) \right\rangle$$
$$= \left\langle |E_t||E_C| \right\rangle$$
$$\approx 0.785$$

(Under simple assumptions)

Conclusions:

1. Phases are more important than amplitudes

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

Importance of Phases

Case 1: *phases are random; amplitudes are exact:*

$$FSC = \left\langle |E_t||E_C|\cos(\varphi_t - \varphi_C) \right\rangle \quad (\cos(\varphi_t - \varphi_C) \to 0)$$
$$= 0$$

Case 2: *phases are exact; amplitudes are random:*

$$FSC = \left\langle |E_t||E_C|\cos(\varphi_t - \varphi_C) \right\rangle$$
$$= \left\langle |E_t||E_C| \right\rangle$$
$$\approx 0.785$$

(Under simple assumptions)

Case 3: *phases are exact; amplitudes are random – replace with expectation:*

$$|E_C| = \left\langle |E_C| \right\rangle$$
$$FSC \approx 0.886$$

Conclusions:

1. Phases are more important than amplitudes
2. Justification for "free lunch" approach

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

Correlation between F_{obs} and F_{calc}

More practically… a warning:

There is an upper limit: $$\mathrm{cor}\left(F_{\mathrm{obs}}, F_{\mathrm{calc}}\right) \leq \left(\frac{2FSC_{1/2}}{1+FSC_{1/2}}\right)^{1/2}$$

E.g. if $FSC_{1/2} = 0.5$ then $\mathrm{cor}\left(F_{\mathrm{obs}}, F_{\mathrm{calc}}\right) \leq 0.82$

Observe correlation higher than this -> further investigation is required

(denoted $CC_{1/2}$ for SFs in Cryo-EM, CC* for intensities in MX)

Karplus & Diederichs (2012) Linking crystallographic model and data quality. Science 336(6084)
Nicholls *et al.* (2018) Current approaches for the fitting and refinement of atomic models into cryo-EM maps using CCP-EM. Acta Cryst D74

Relevant Publications

Yamashita *et al.* (2021) Cryo-EM single-particle structure refinement and map calculation using Servalcat. Acta Cryst D77, 1282-91.

Casanal *et al.* (2020) Current developments in Coot for macromolecular model building of electron cryo- microscopy and crystallographic data. *Protein Science* 29(4), 1055-64.

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.

Burnley *et al.* (2017) Recent developments in the CCP-EM software suite. Acta Cryst D73.

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.

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 - *Servalcat*
Fei Long - *AceDRG, LibG*
Paul Emsley - *Coot*
Lucrezia Catapano - *Coot*

CCP-EM

Tom Burnley
Colin Palmer
Rangana Warshamanage - *EMDA*
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