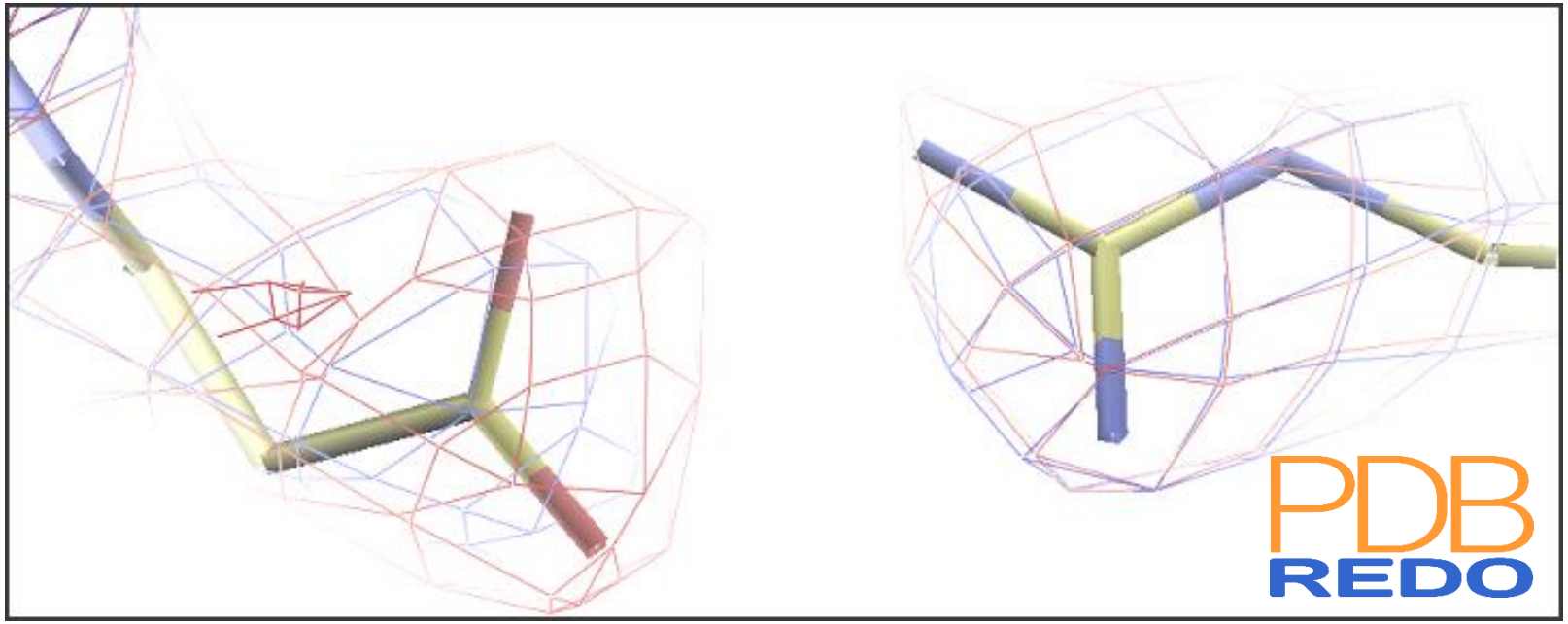
How good is my model? Can I improve it?

And how do I share it with the community?



Robbie P. Joosten

Netherlands Cancer Institute

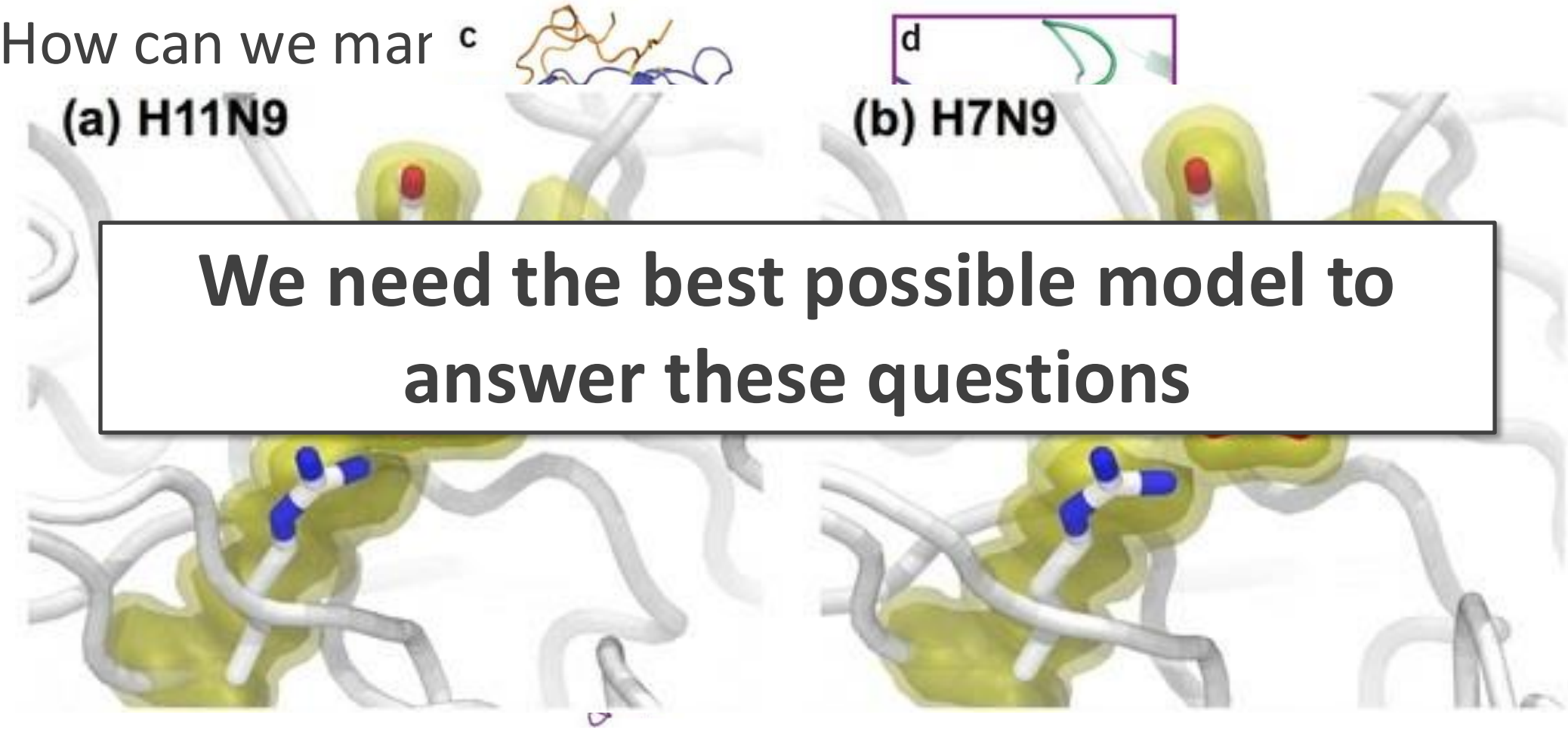
CCP4-DLS school 2025

Oncode Institute

NETHERLANDS CANCER INSTITUTE
ANTONI VAN LEEUWENHOEK

We want to know...

- What are a protein's function and mechanism?
- How can we mar



We need the best possible model to answer these questions

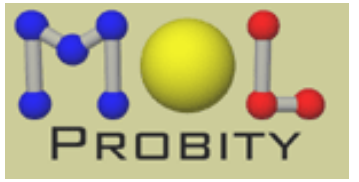
Is my model as good as it can be?

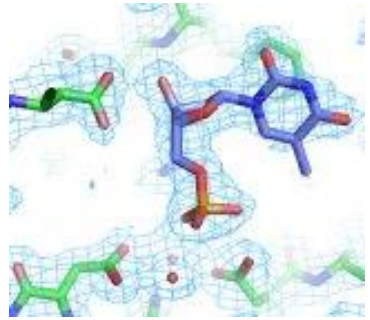
1. Use validation when making the model
   - Check model vs. data and vs. prior knowledge
   - Focus on outliers (fix or explain them)
   - Know the things that can go wrong
2. Optimise the model
   - Focus on what can be improved
   - Choose best refinement parameters and restraints
   - Rebuild parts of the model
   - PDB-REDO automates this
3. Deposit and publish the model
   - Quality of the model description and annotation

Validation

Need to know

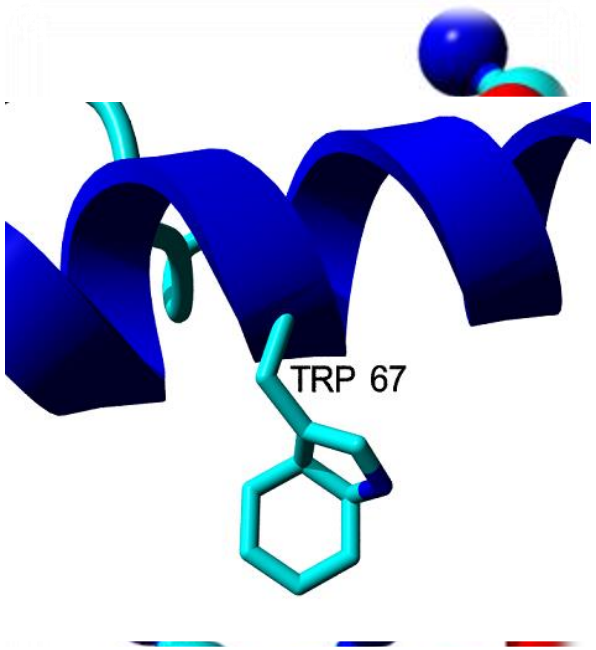
- Check the validity and value of a model
  - Accuracy and precision
- Many different software tools
  - General: MolProbity, PDB validation server
  - Non-protein: CheckMyMetal, Privateer, DNATCO
  - Tools may check the same things differently
- Not a substitute for common sense
  - False positives do occur
  - Conflicting results
  - Not all problems are detected (explicitly)

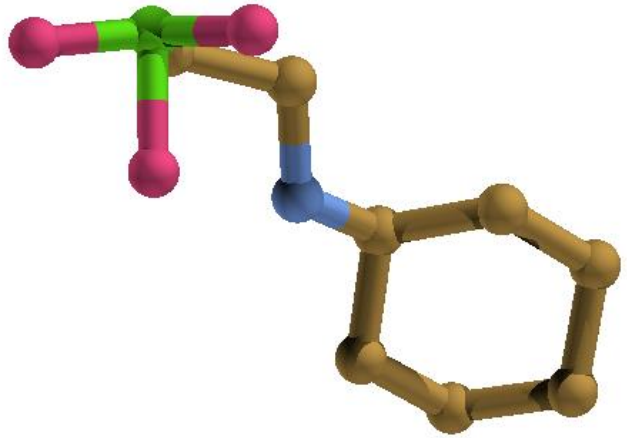




Things to validate

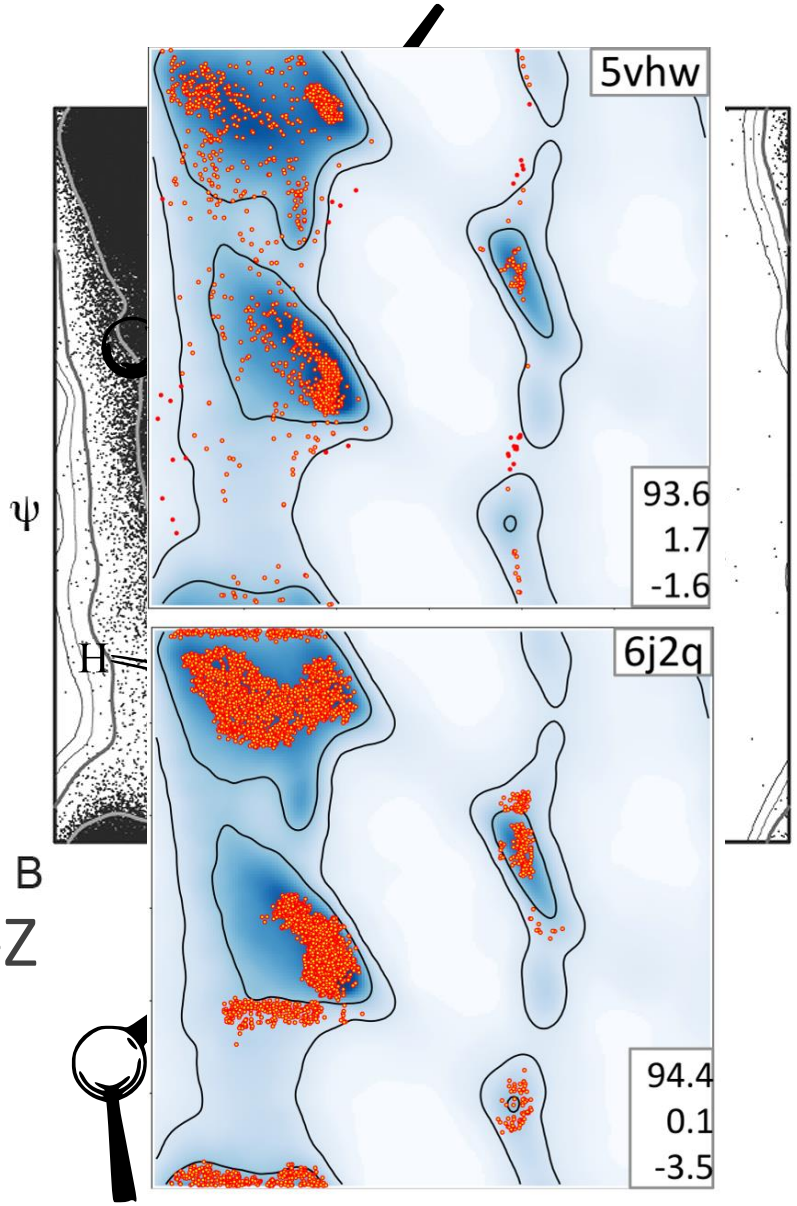
- Bond lengths and angles should be normal
  - Express deviation in terms of SD (Z-scores)
  - The higher the Z-score the more unlikely the bond/angle
    - Example: Z = 105
  - Individual outliers usually mark map fitting errors
  - Large overall deviation indicate poor model refinement
    - Express as rmsZ, not rmsd
    - Should be < 1.000 (increase restraint weight)
- Some things should be planar, some not
  - Rarely a problem in protein/nucleic acid nowadays
  - Still a problem in carbohydrates and ligands!





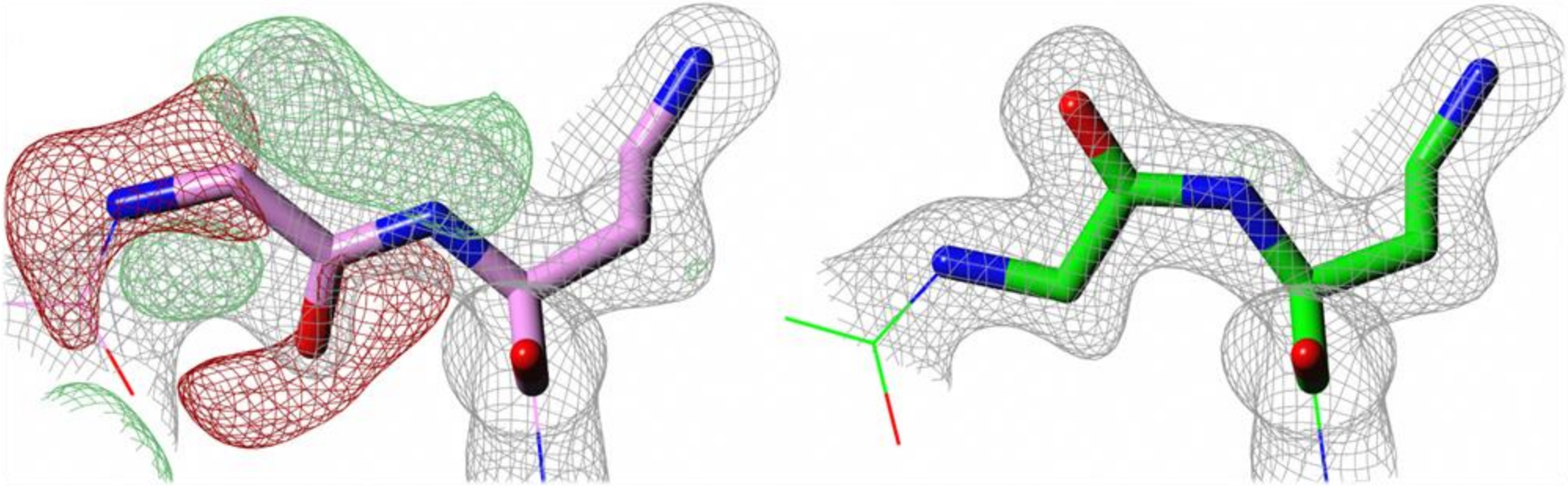
Backbone torsion angles

- Ramachandran plot
  - ϕ and ψ angles
  - Not all conformations are possible
  - Compare to the whole PDB or a subset
- Different implementations
  - MolProbity and COOT: preferred, okay, outlier
    - Good for finding specific individual problems
    - Check severity of outlier on the plot
  - Tortoize, MolProbity, Phenix: whole plot Rama-Z
    - Good for checking building and refinement progress
    - Sensitive to over-restraining



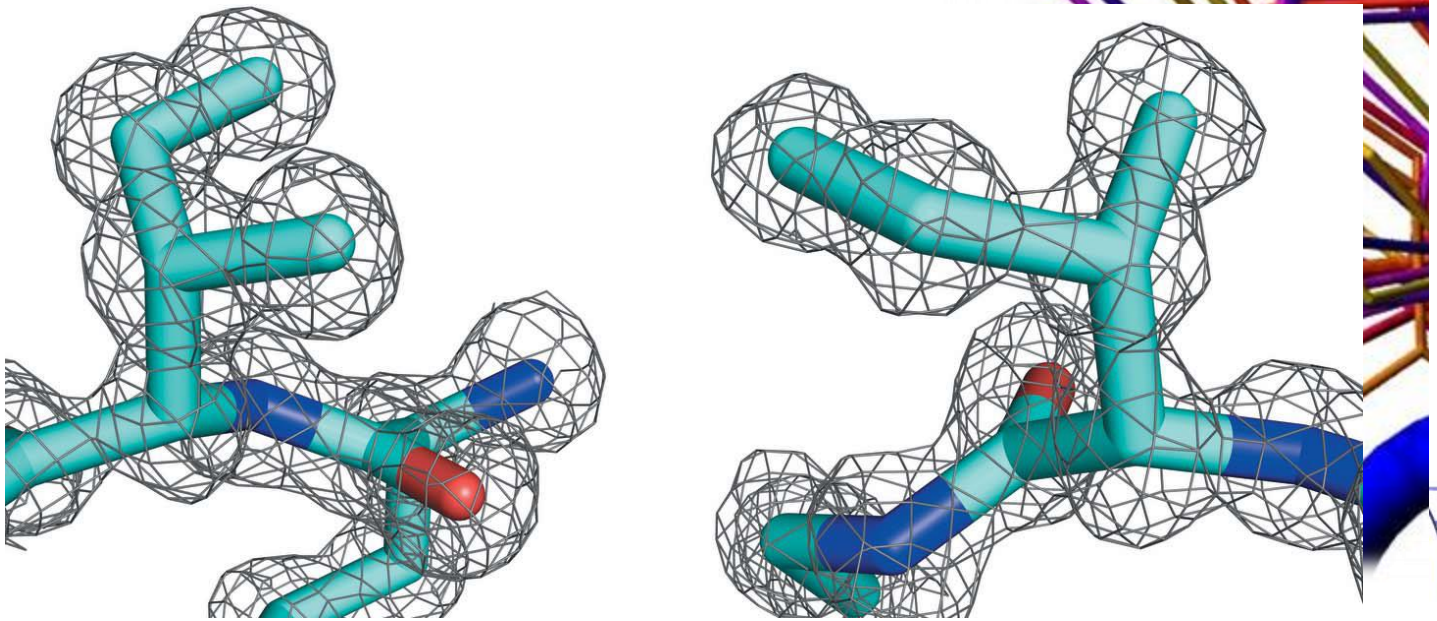
Backbone torsion angles

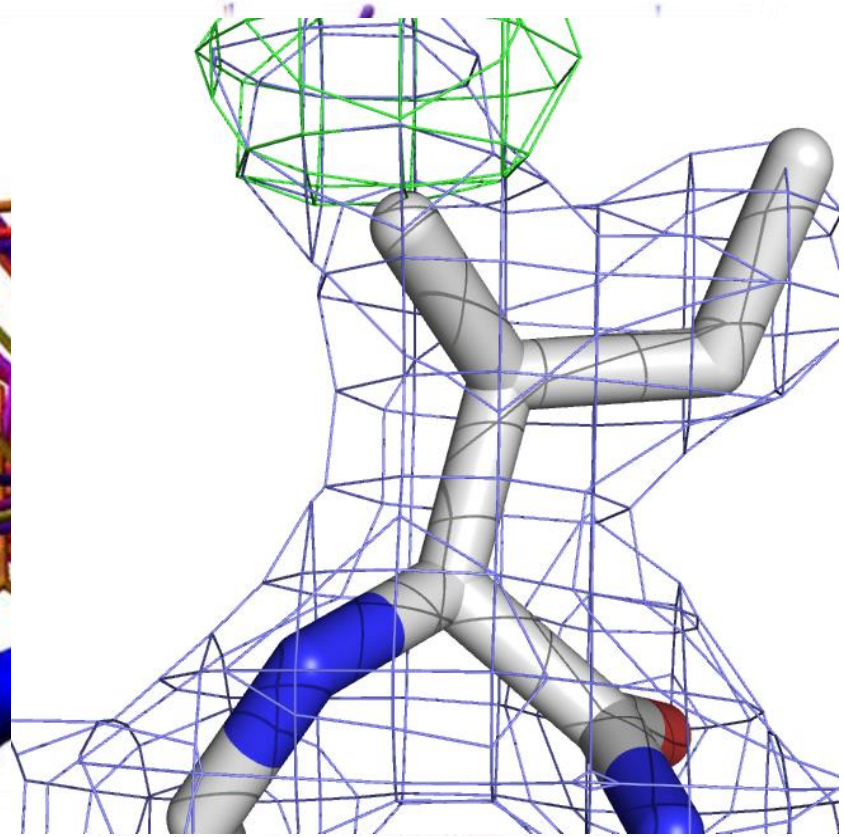
- Peptides are flat
  - ω angle is ~180° (*trans*) or ~0° (*cis*)
  - Fitting errors or the wrong restraints cause outliers
    - COOT treats all peptides as *trans* by default



Side chain torsion angles

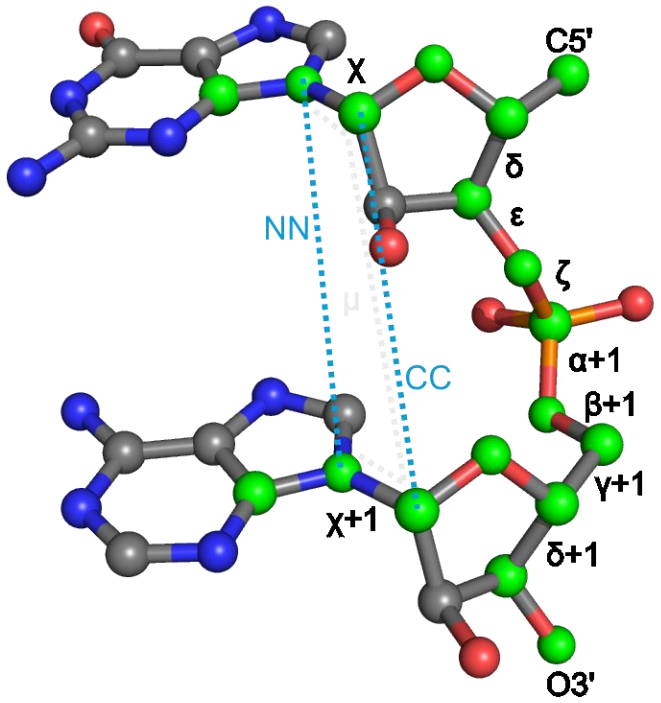
- Steric hindrance causes discrete rotamers
- Check against (backbone specific) distributions from the PDB
- Outliers are fitting errors or f

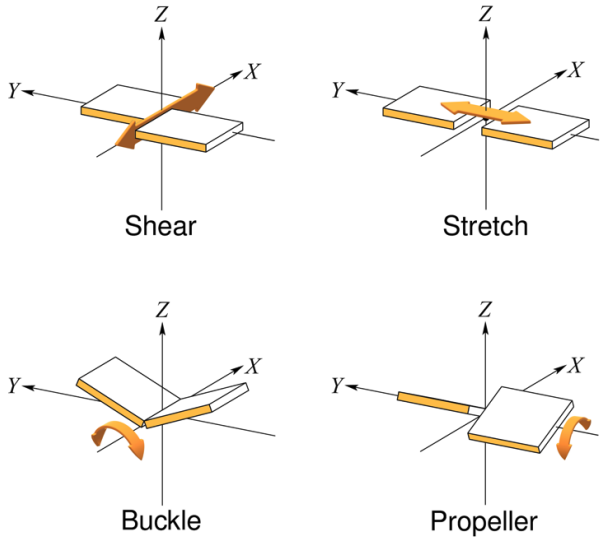




DNA and RNA

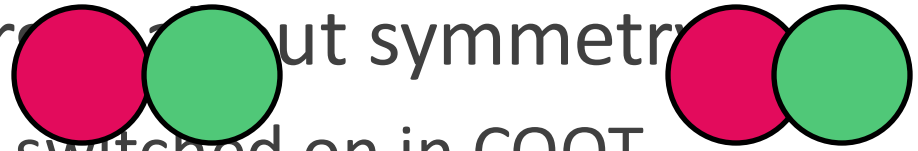
- Too many torsions to do Ramchandran-like things
  - MolProbity does conformational ‘suites’ (RNA only)
  - DNATCO used pairs of sequential residues
    - Conformation normality score CONFAL
- Base pairs also have conformations
  - PDB-REDO calculates normality of relative base orientation in Watson-Crick base pairs
    - Single Z-score per parameter, DNA/RNA-specific
  - Individual Z-scores are combined to overall Z_{bpG}
    - rmsZ score for the whole structure model
  - Will be extended to a broader set of base pairs





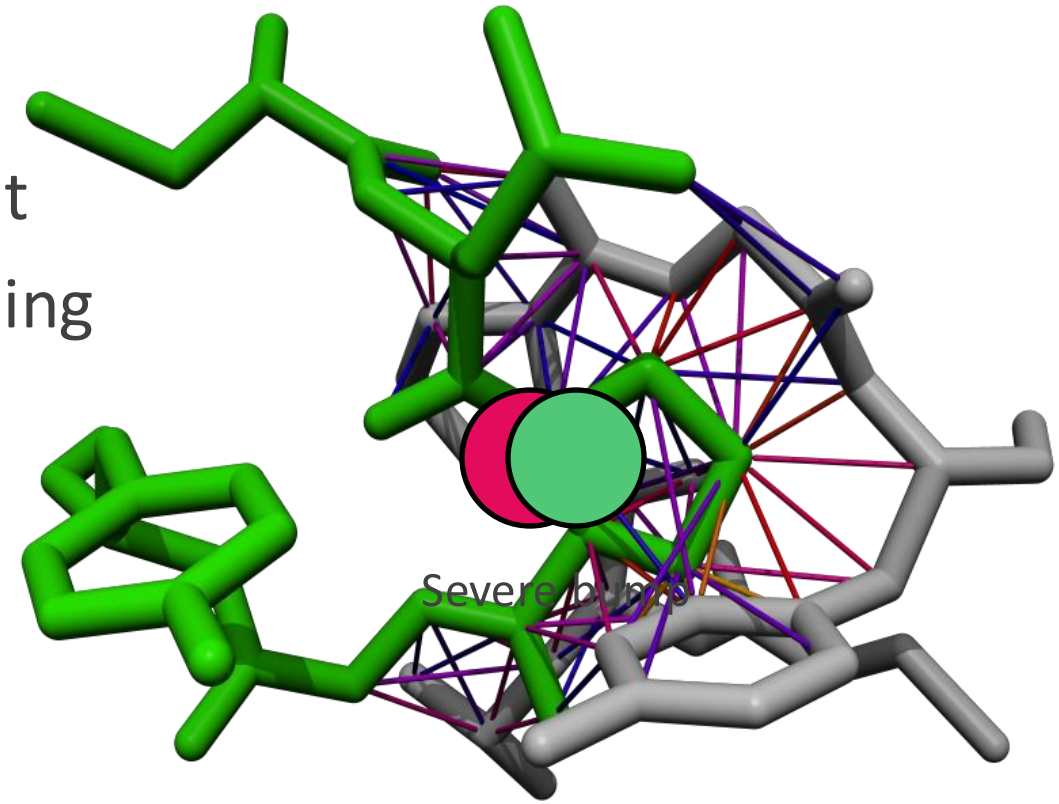
Bumps/clashes

- Two atoms cannot occupy the same space
- Average PDB entry > 100 bumps
- Bumps vary in severity
  - Mild bumps can be fixed by refinement
  - Severe bumps typically require rebuilding
- Don't forget about symmetry
  - Keep it switched on in COOT



Normal contact

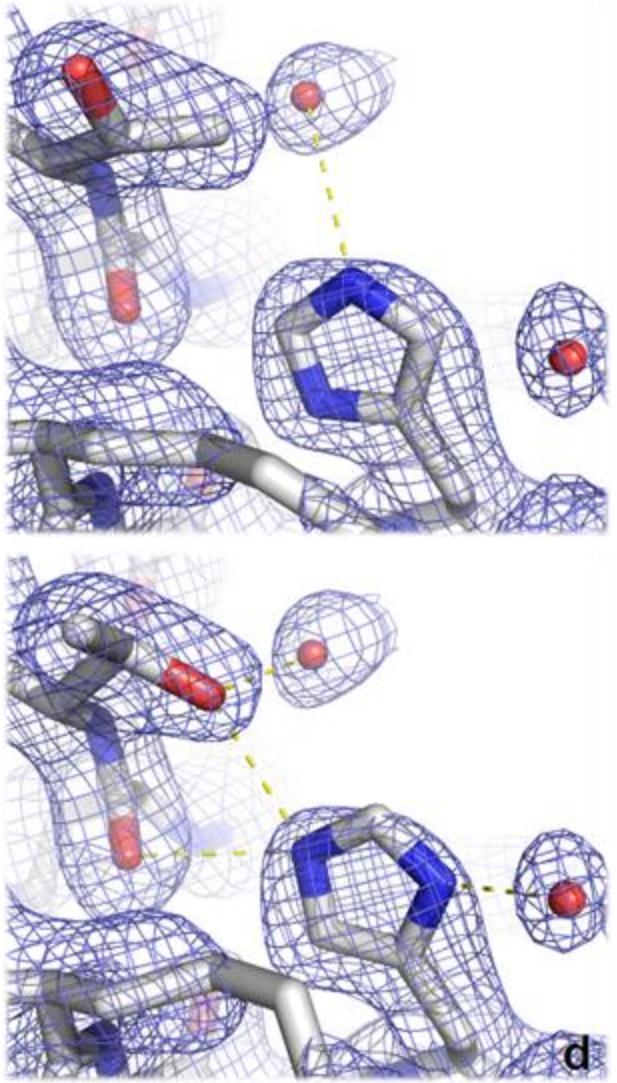
Mild bump



Severe bump

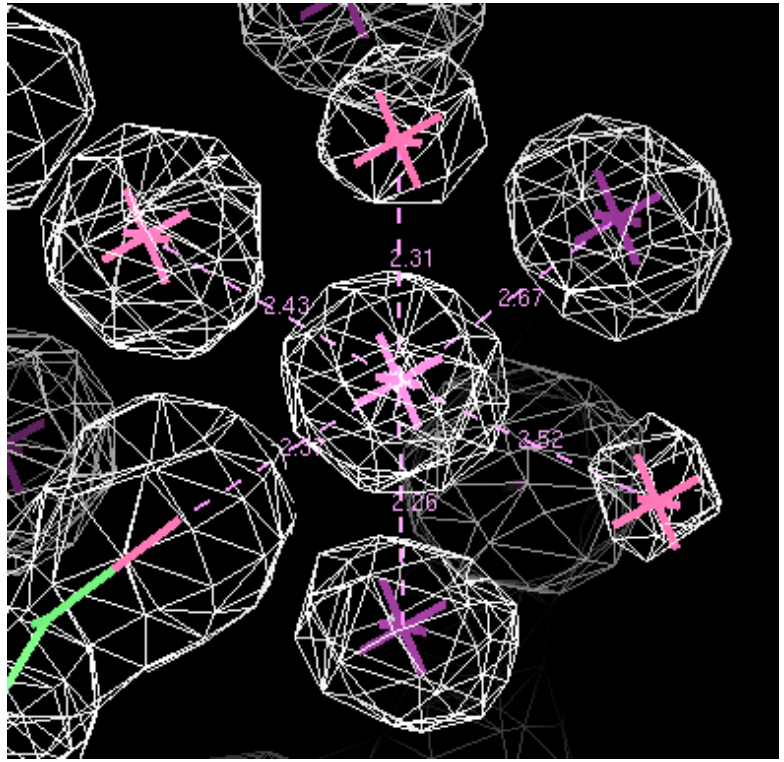
Hydrogen bonds

- Asn, Gln, and His flips
  - Detected by PDB-REDO, MolProbity & COOT
  - Also use common sense
- Buried unsatisfied H-bond donors and acceptors mark (subtle) errors
  - Use environment distances in COOT
  - Check your Arginines
- Waters should also make H-bonds
  - 3b3q has > 250 waters without H-bonds



Metal ions

- Light metal ions are easily overlooked
  - Water, Na^{+}, and Mg^{2+} have same number of electrons
- Detect and validate with COOT, Phenix, CheckMyMetal
  - All use the Bond Valence method
  - Depends on coordination distances
    - Be careful with restraints
  - Very different results

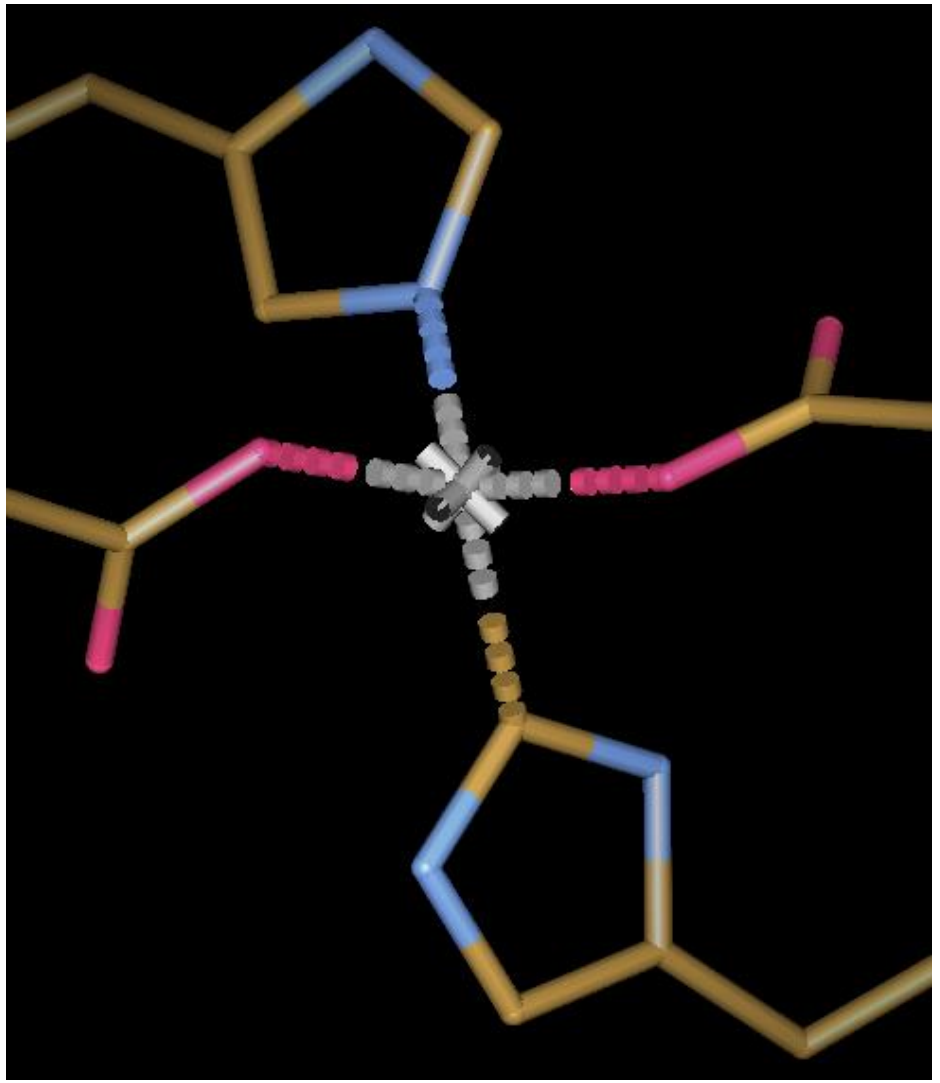


Metal ion identification and validation

- Use anomalous maps
- Use wavelength scan from synchrotron
- Keep your crystallisation conditions in mind
- Check site geometry with in MetalPDB
- If it is really important, do more experiments
- Validate metal site with CheckMyMetal

Metal ions

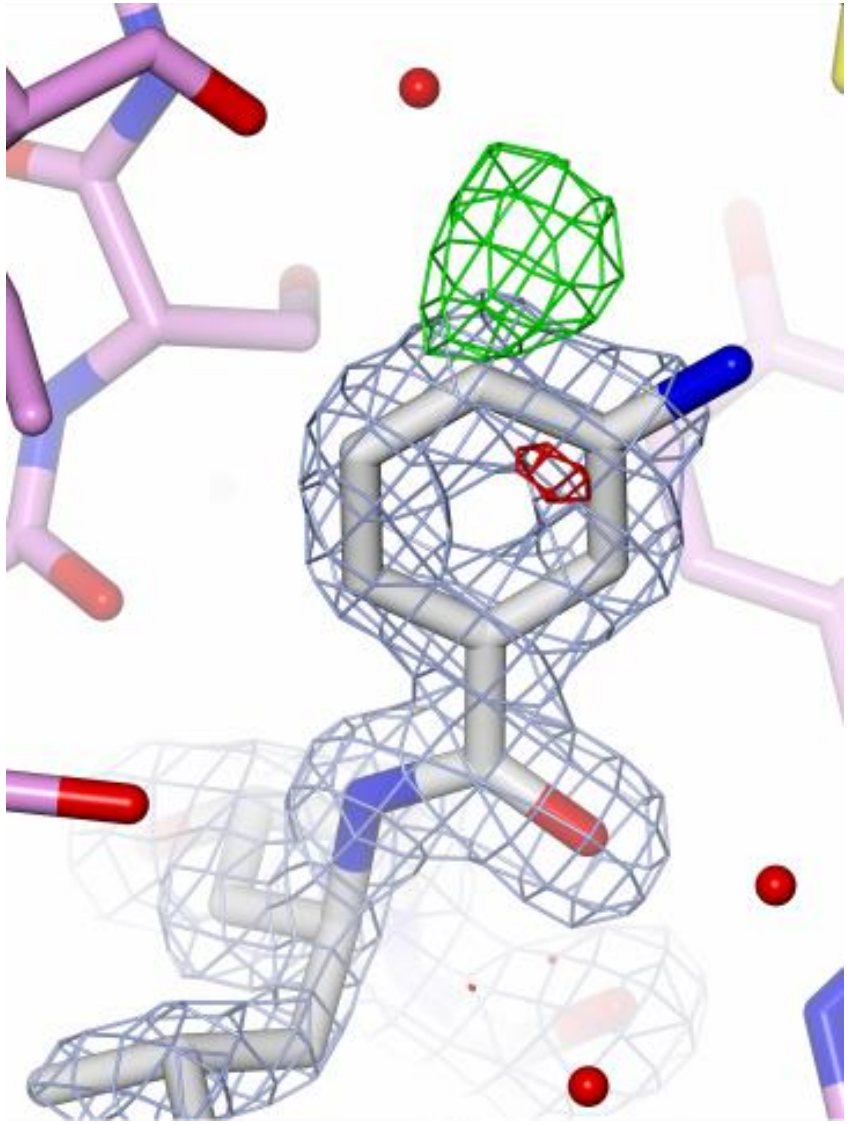
- Na, Mg, K, Ca prefer being coordinated by oxygen atoms
  - Flip Asn or Gln side chains if needed
- Carbons usually do not coordinate metals
  - Flip His side chains if needed
  - Cyanide and carbon-monoxide are exceptions



Ligands

Building and validation steps:

1. Is something there?
   - Check the (difference) density
2. Is it my ligand?
   - Check contacts
   - Remember crystallisation conditions
   - Check the density in detail
3. Is the geometry sensible?
   - First check the restraints themselves
     - Don’t forget about chirality
   - Then check model against restraints



Validation is a lot of work, but it helps you make better models

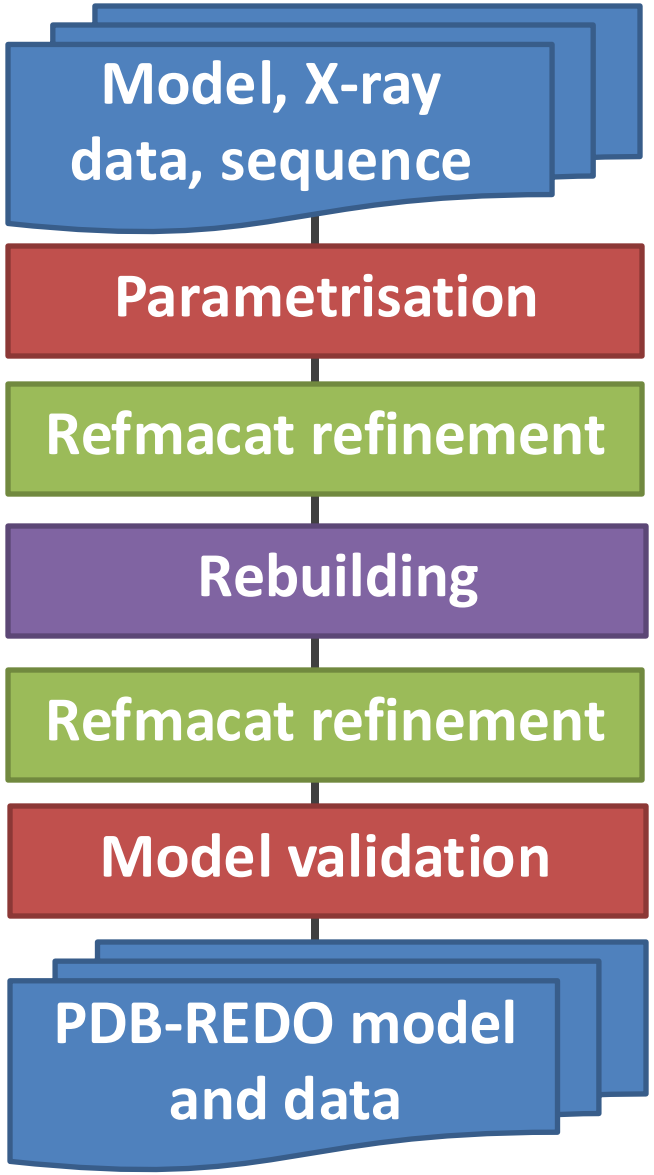
Model optimisation = making choices

- Refinement settings
  - Restraints and weights (geometry, B-factors, homology, jelly body)
  - Solvent model
  - High resolution cut-off
  - Special cases (NCS, twinning, occupancies)
- Number of model parameters
  - B-factor model
  - Number of TLS models
- Structure model
  - Main chain
  - Side chains
  - Hetero compounds

Automation speeds up optimisation

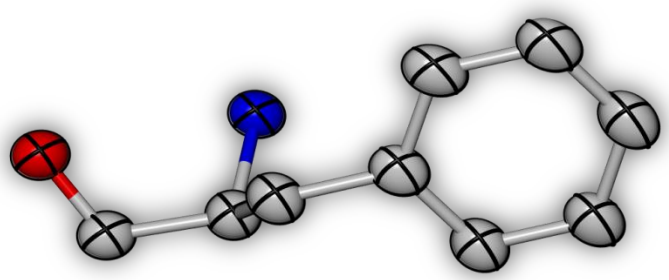
PDB-REDO

- Pipeline for X-ray & electron crystallography
  - Fully automated expert system
  - Refines, rebuilds, and validates your model
- Well-tested and high-throughput
  - Run on the entire PDB
  - Databank with weekly updates (pdb-redo.eu)
- Available as webserver and through CCP4
  - 1900 active users



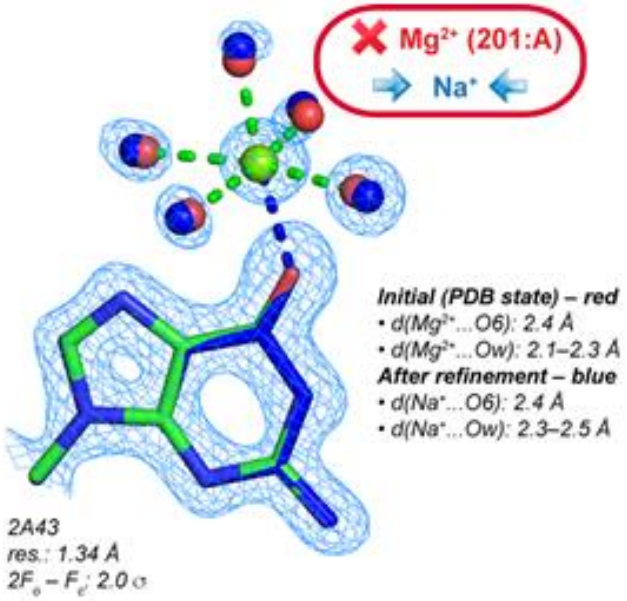
Features and algorithms

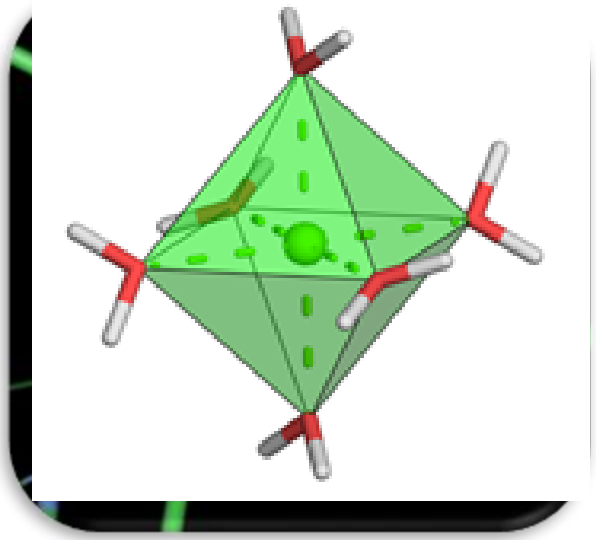
- NCS and twinning treated automatically
  - Warns for space group errors
- Paired refinement to find high resolution cut-off
- B-factor model selection:
  - Refine alternative models and select best one
    - Isotropic, anisotropic, or flat B-factors
    - One TLS group per chains, user-provided model, or no TLS
- Grid searches:
  - Optimise solvent mask parameters
  - Select weights for geometric and B-factor restraints



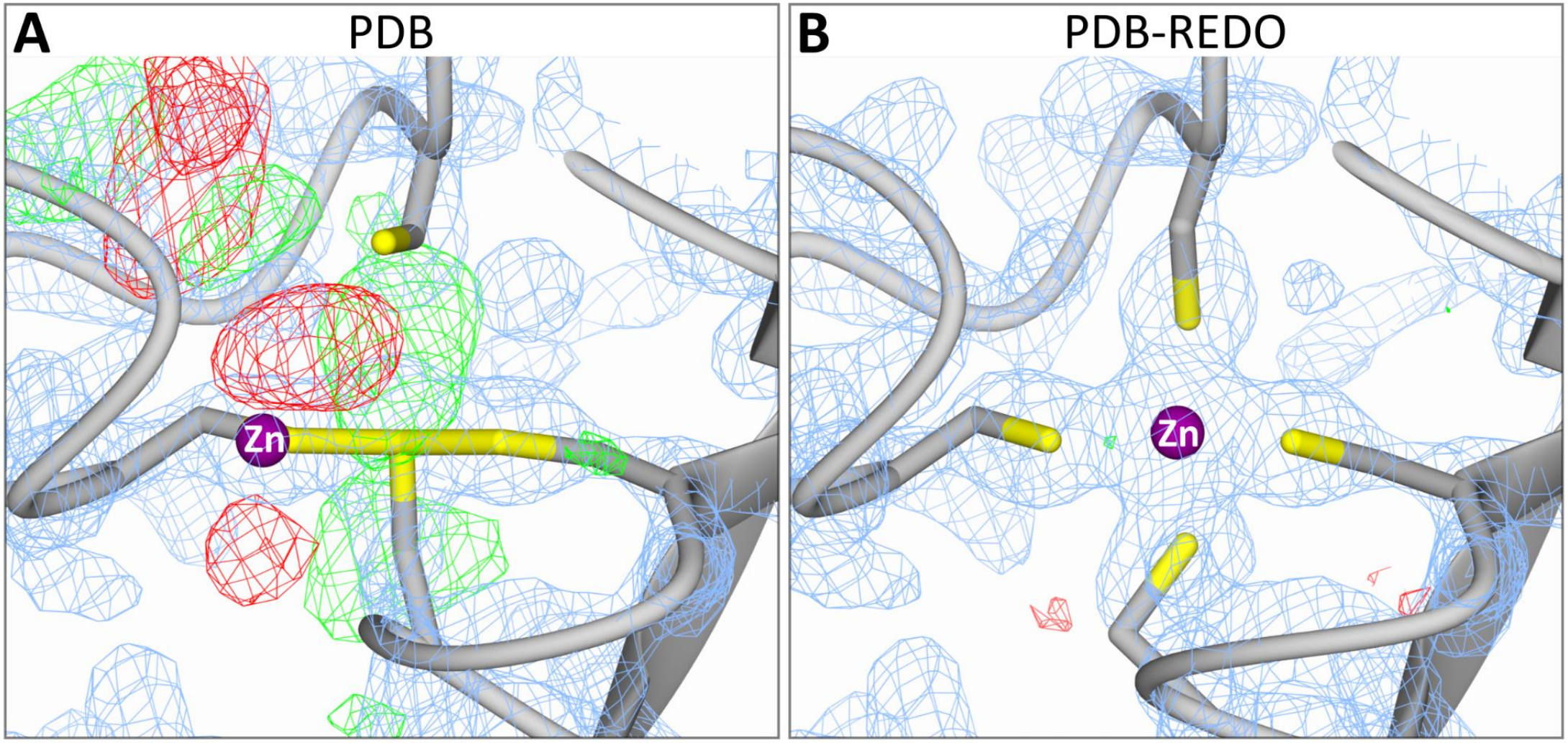
Features and algorithms

- Case specific restraints:
  - Jelly-body restraints (low-ish resolution)
  - Zinc sites (all structural zinc sites)
    - Many distorted sites in the PDB
    - Solution from *platonyzer*: cleaned set of restraints
  - Octahedral Sodium and Magnesium sites
    - Can only be distinguished by coordination distance
    - Distance biased by restraints, but restraints are needed
    - Solution: use angle restraints to define an octahedron
      - 18 angle restraints per site, no distance or VdW restraints



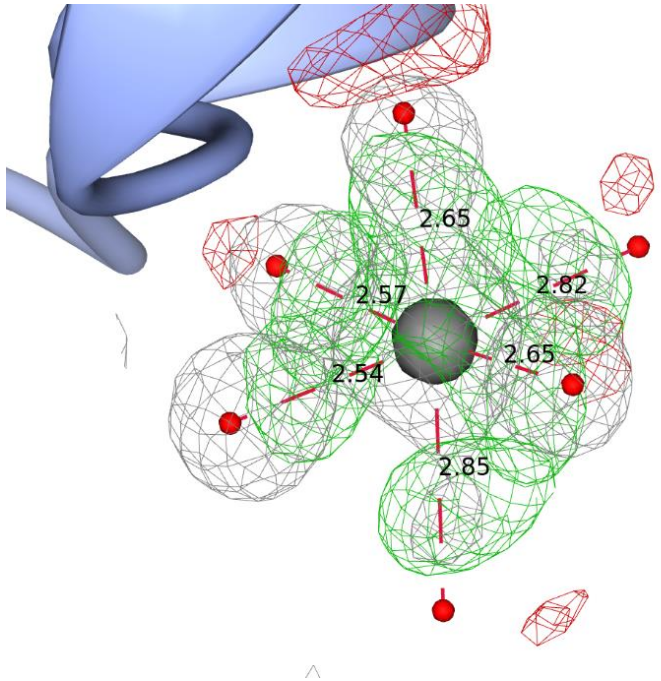


Added value of zinc restraints

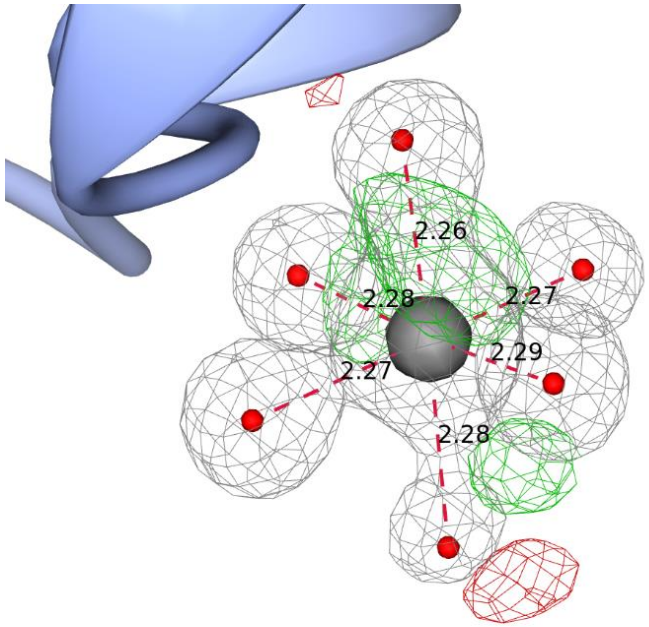


Added value of octahedral restraints

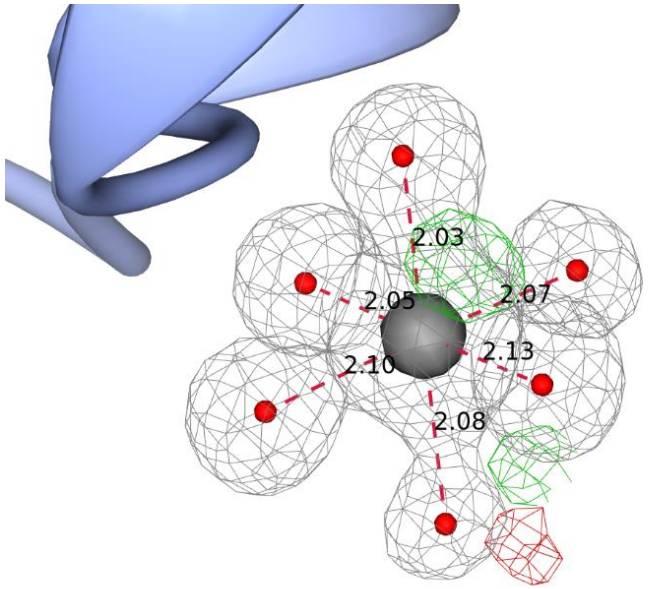
- Magnesium site with sodium modelled becomes Mg-like
  - Opposite can happen as well



PDB



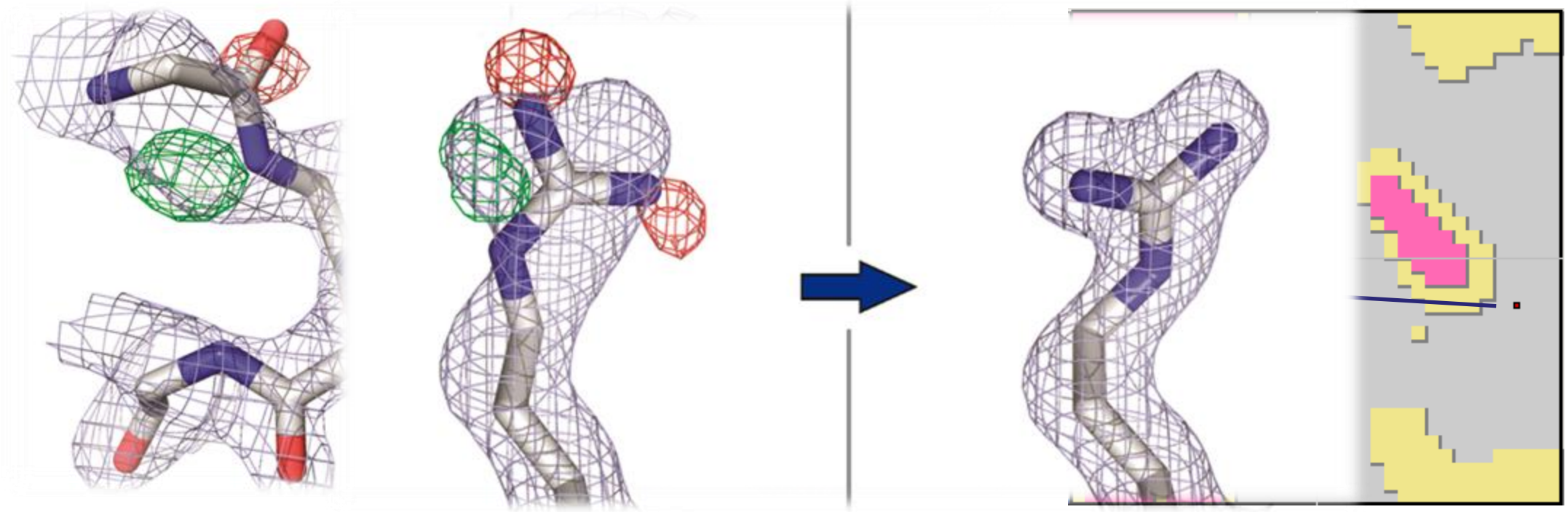
PDB-REDO



PDB-REDO + platonyzer

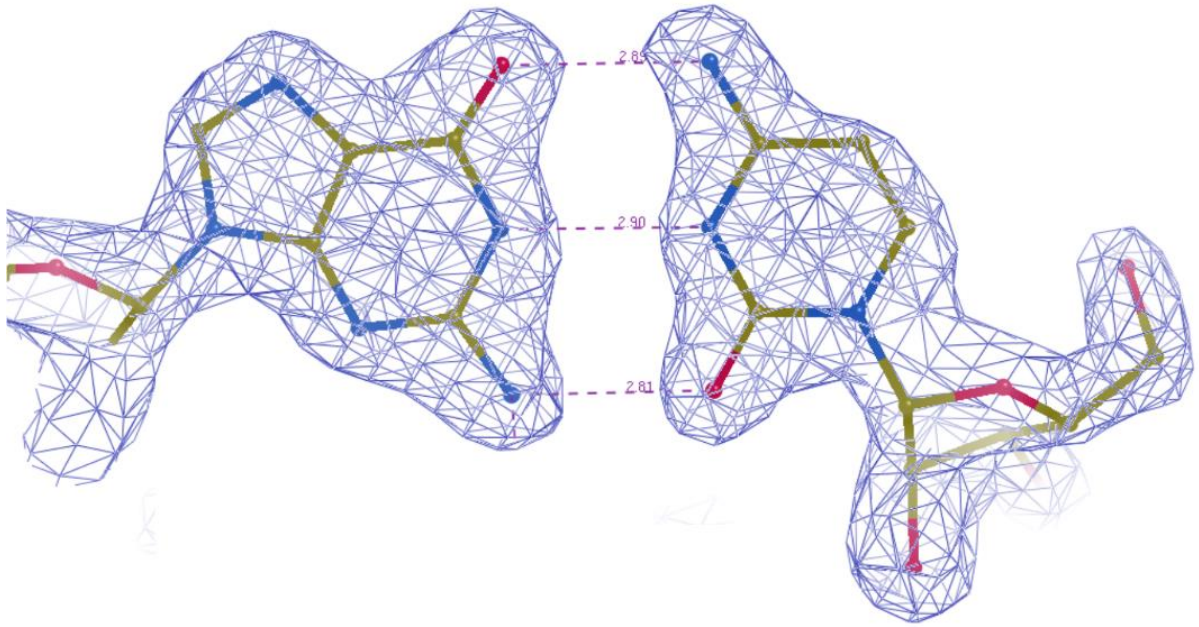
Model rebuilding

- Peptide flipping:
  - If a flip improves density fit and Ramachandran plot: accept flip
- Side-chain rebuilding and completion:
  - Add missing side-chains, rebuild existing ones, flip His/Asn/Gln if needed

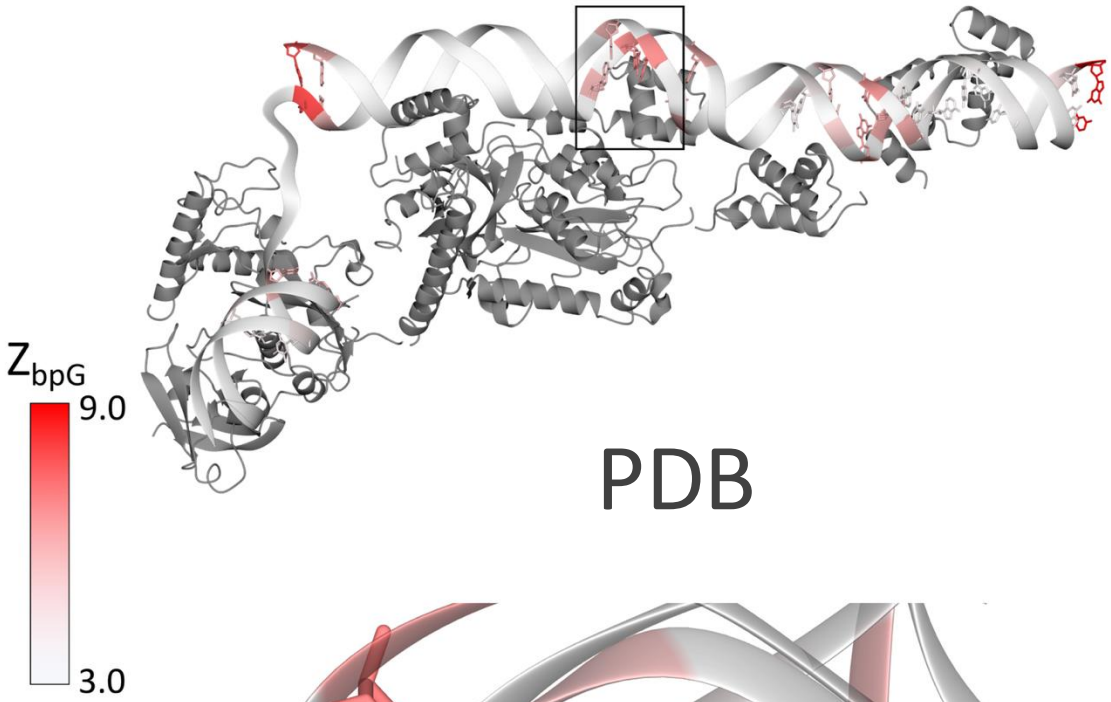


Nucleic acid features

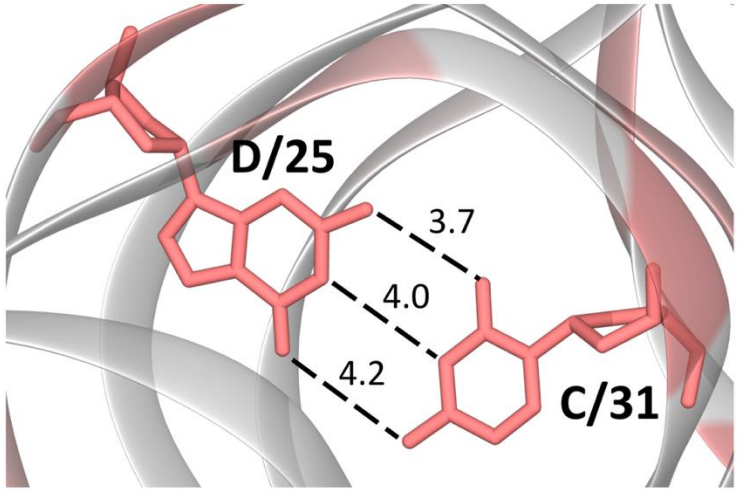
- Restraints for H-bond distances in WC base pairs
  - Targets mined from high-res PDB-REDO entries
- Base stacking restraints from LibG
- Geometric validation
  - Base pair geometry normality
    - Z_{bpG} and $rmsZ_{bpG}$
  - CONFAL scores from DNATCO

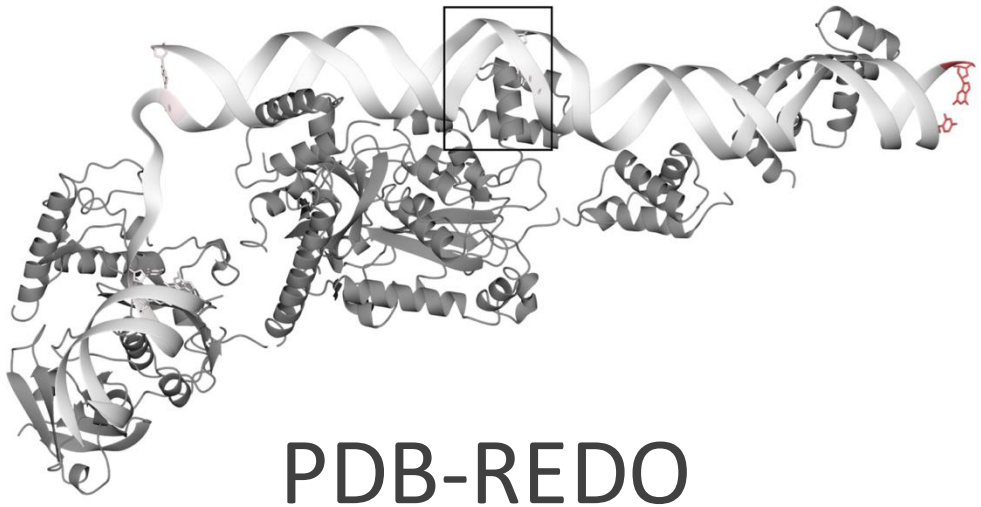


Effect of nucleic acid restraints

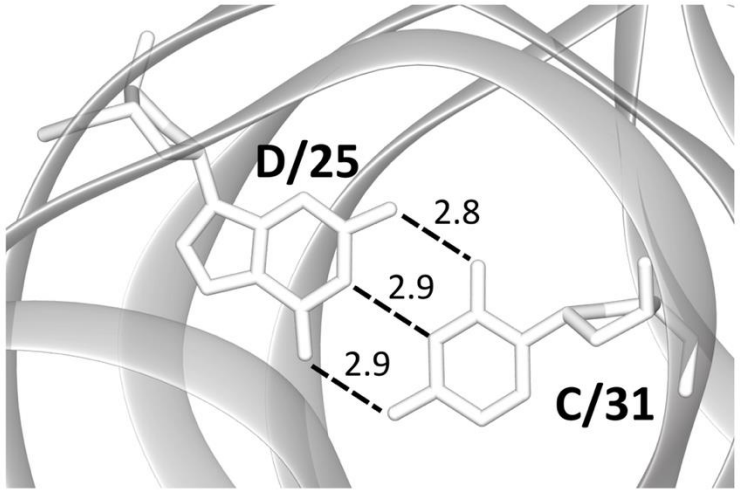


PDB



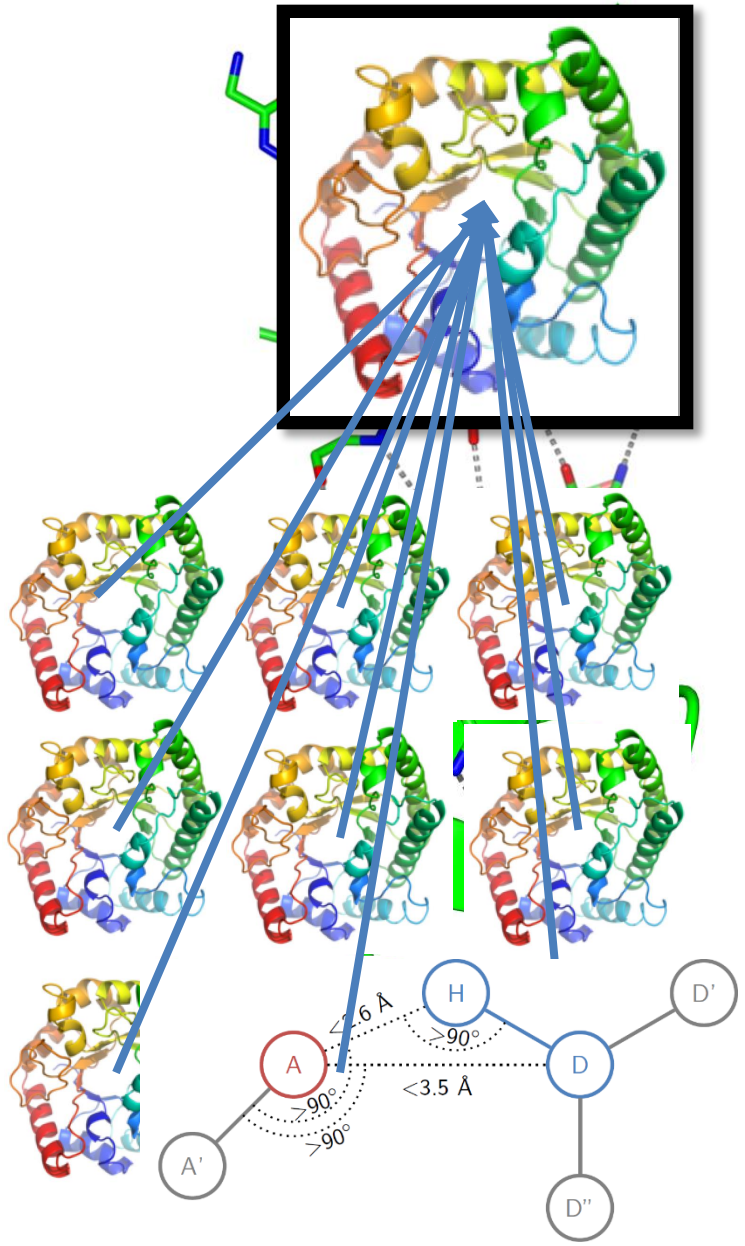


PDB-REDO



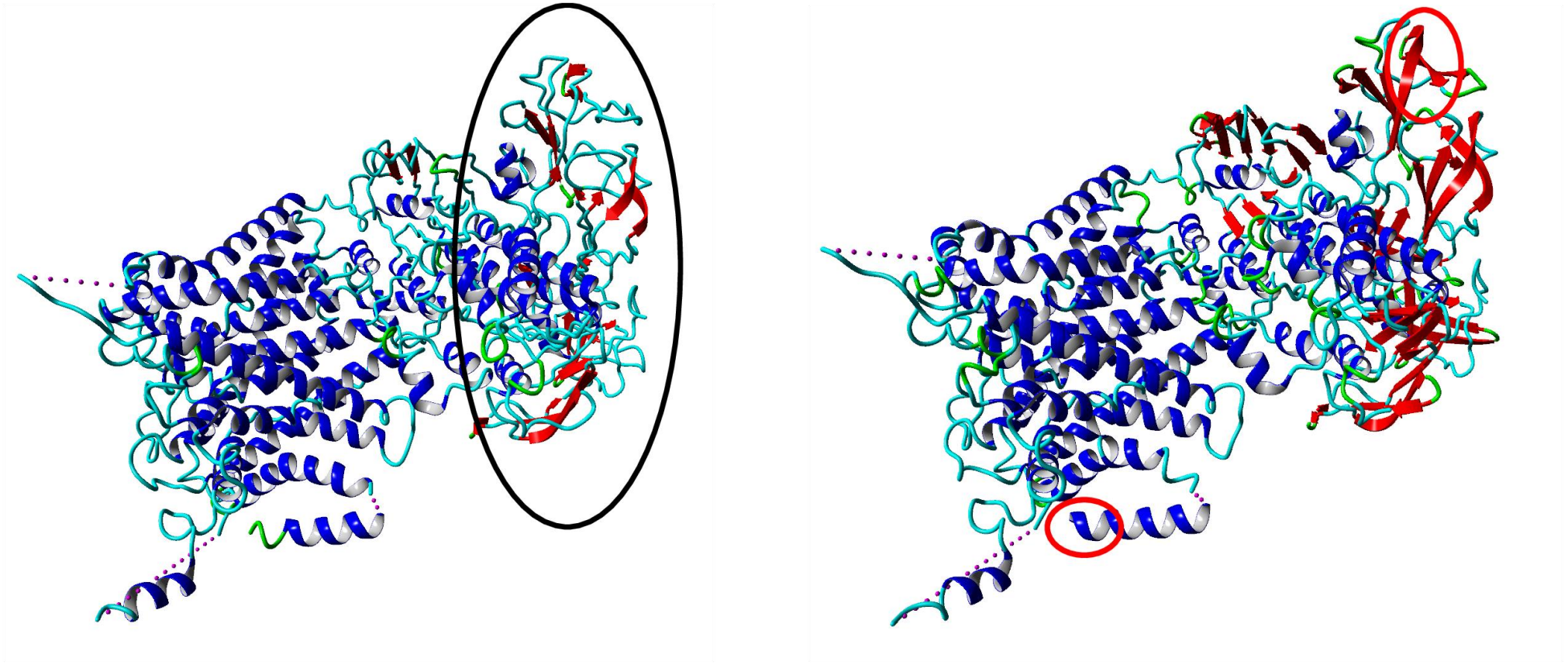
Homology in PDB-REDO

- Homology: more knowledge, better model
  - Average PDB entry has > 10 homologs
- Combine knowledge from all homologs
  - High throughput screening gives a lot of data
  - Highlights true structural differences in models
    - Important for drug development
- Hydrogen bond restraints in refinement
  - Many restraints that can be filtered reliably
  - Use local structural variability for weighting



Added effect of homology restraints

E. Coli maltose transporter (3fh6, 4.5Å)



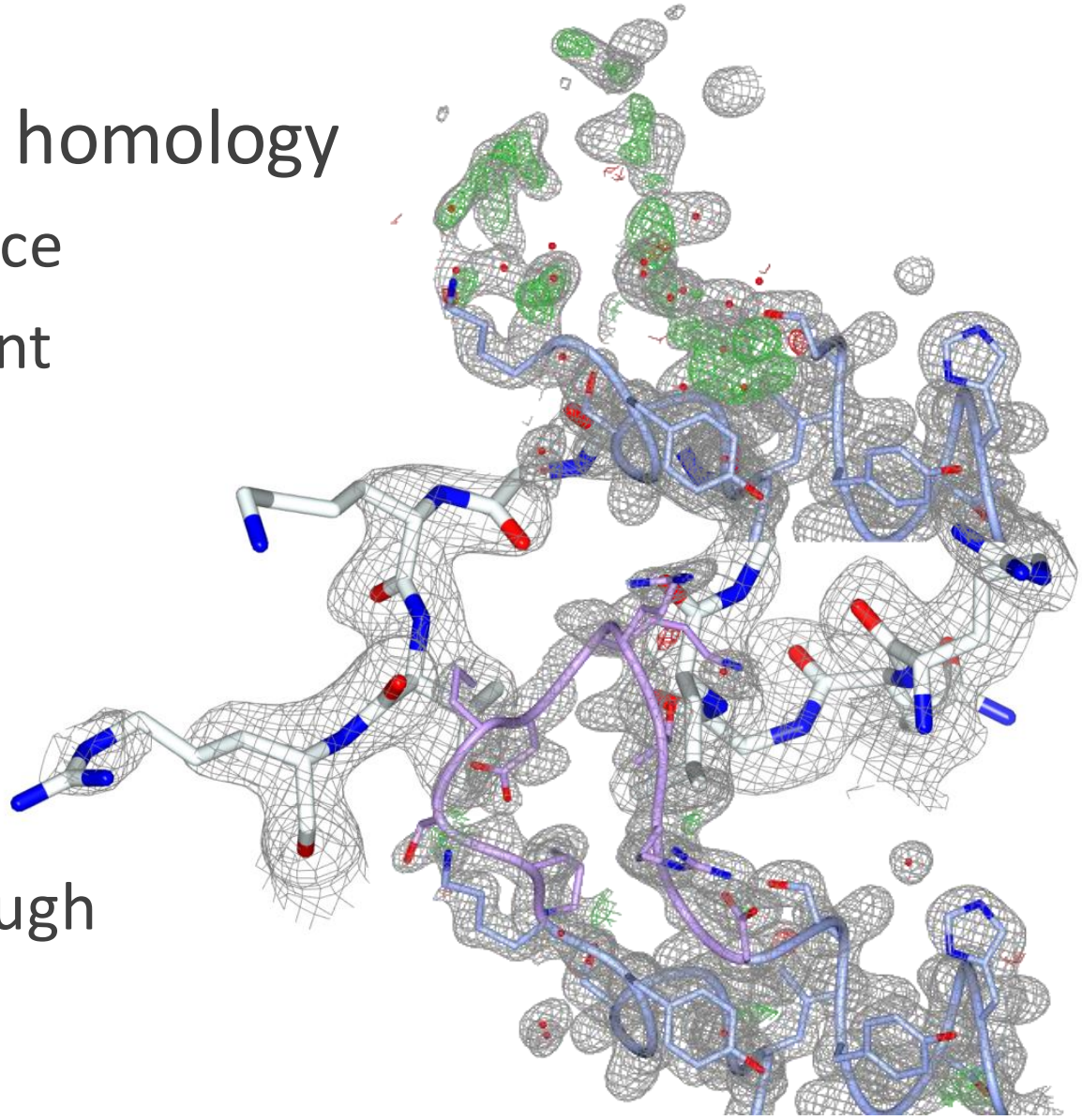
PDB

PDB-REDO

Homology-based loop building

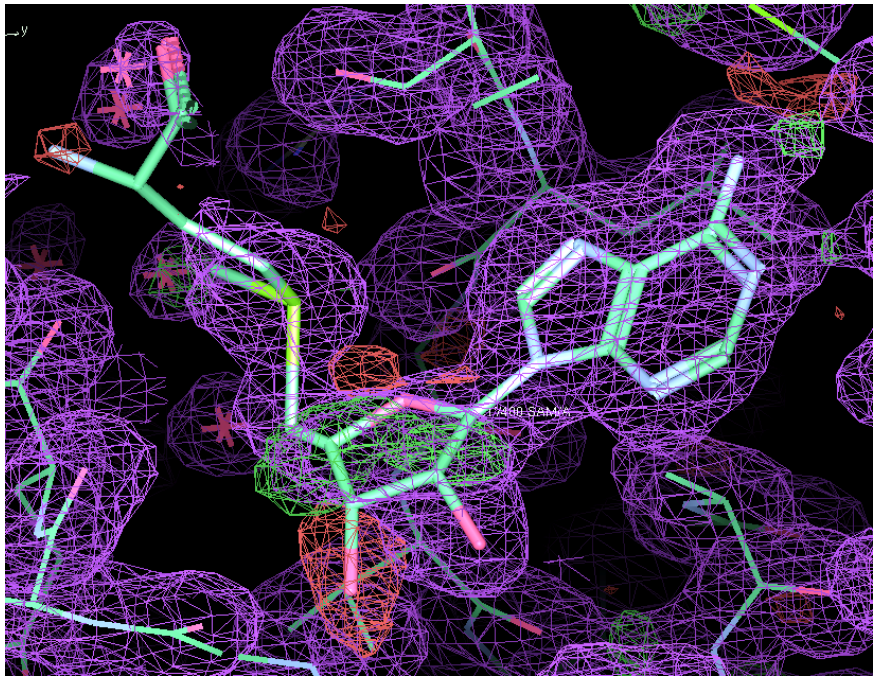
Loopwhole adds missing loops by homology

1. Find missing loop based on sequence
2. Find homologs with the loop present
3. Prepare for loop transfer
4. For all homologous loops:
   - Align loop borders
   - Copy over loop
   - Refine with coot-mini-rsr
5. Keep the best loop if it is good enough
   - Filter on geometry and map fit

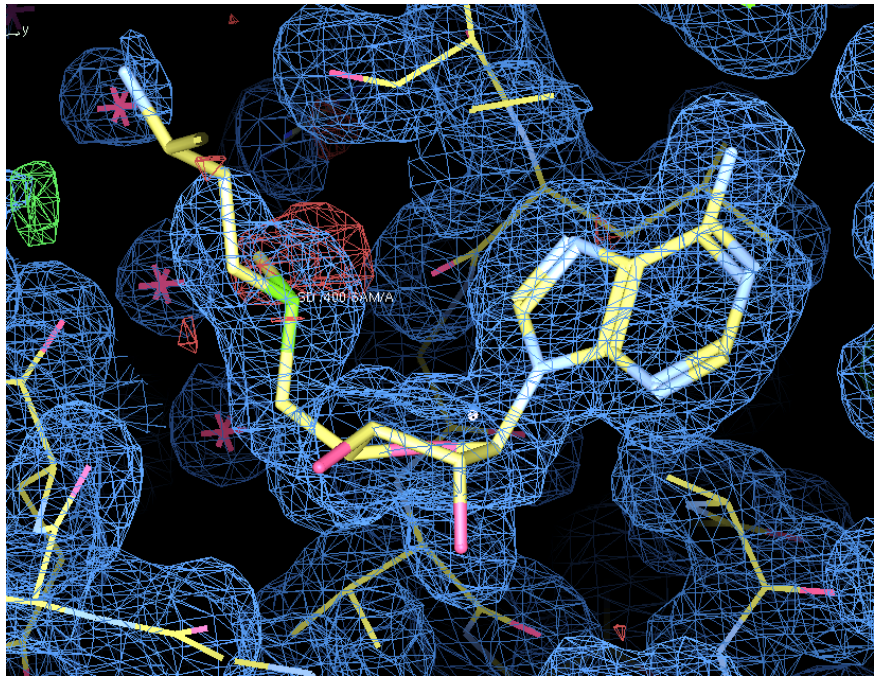


Validation (before and after scores)

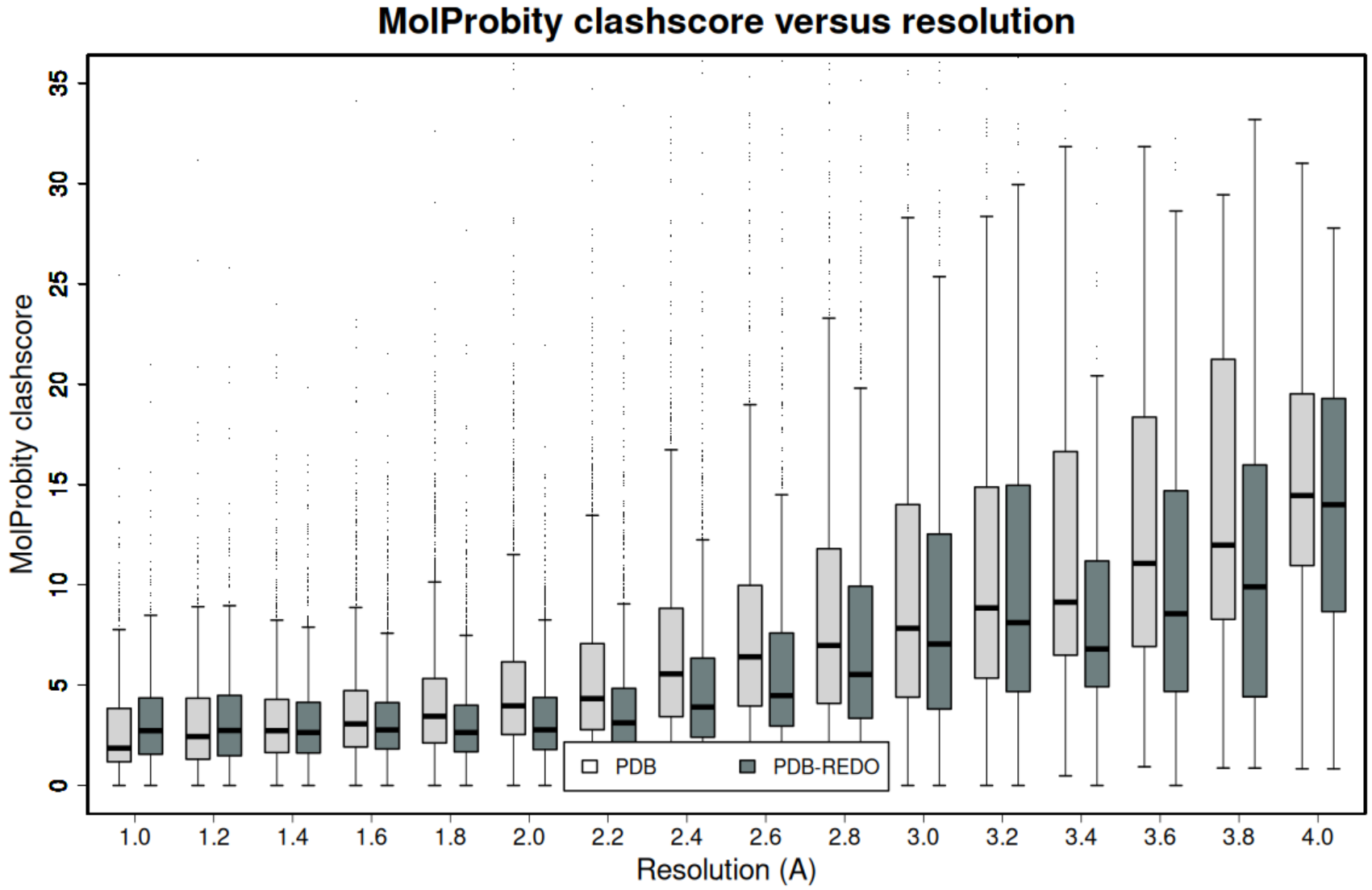
- Geometry: packing, rotamers, Ramachandran plot, H-bonds, bumps, DNA/RNA/carbohydrate quality
- Density fit: per-residue RSR, RSCC, EDIAm and OPIA
- Ligands: density fit, interactions and Heat-of-Formation



	PDB	REDO
RSCC	0.91	0.96
Bumps	22	5
H-bonds (kJ/mol)	-70	-95
HoF (kJ/mol)	11817	322



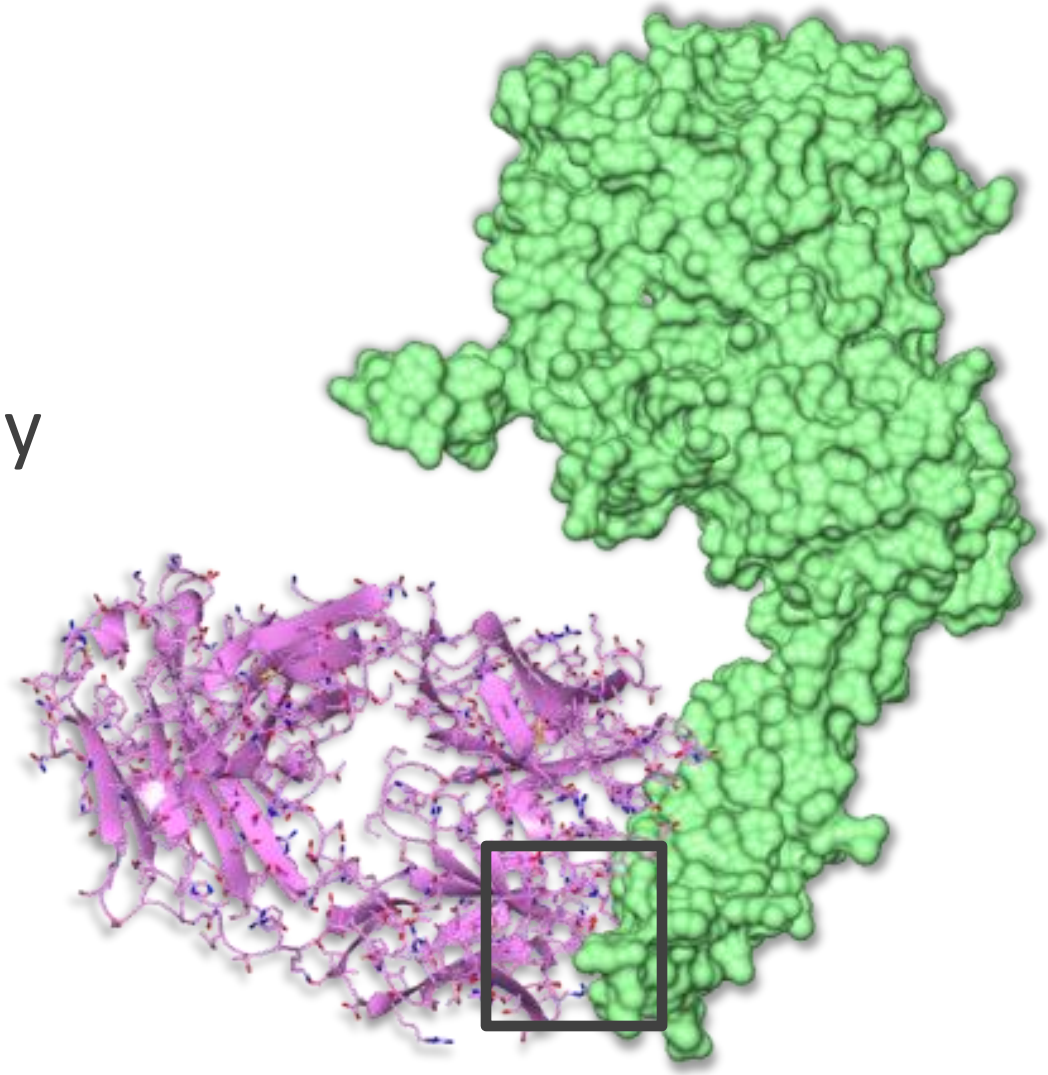
PDB-REDO model quality (n = 189k)



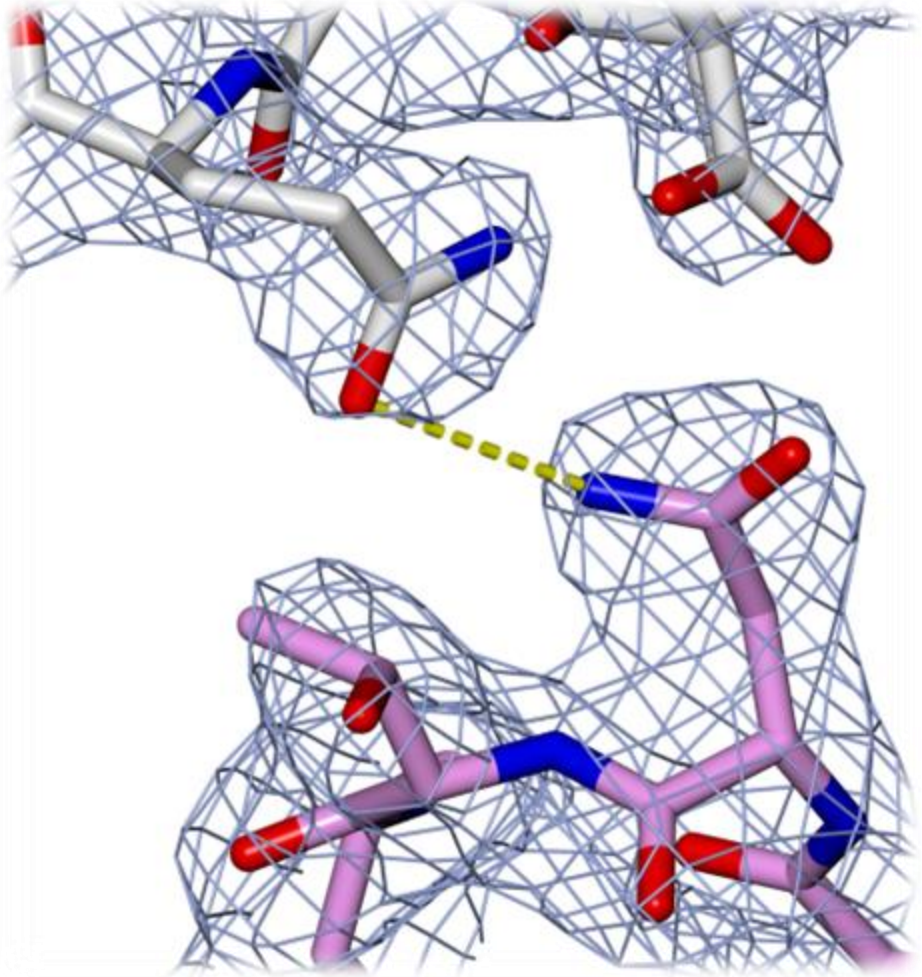
Herceptin – HER2 interface

After PDB-REDO:

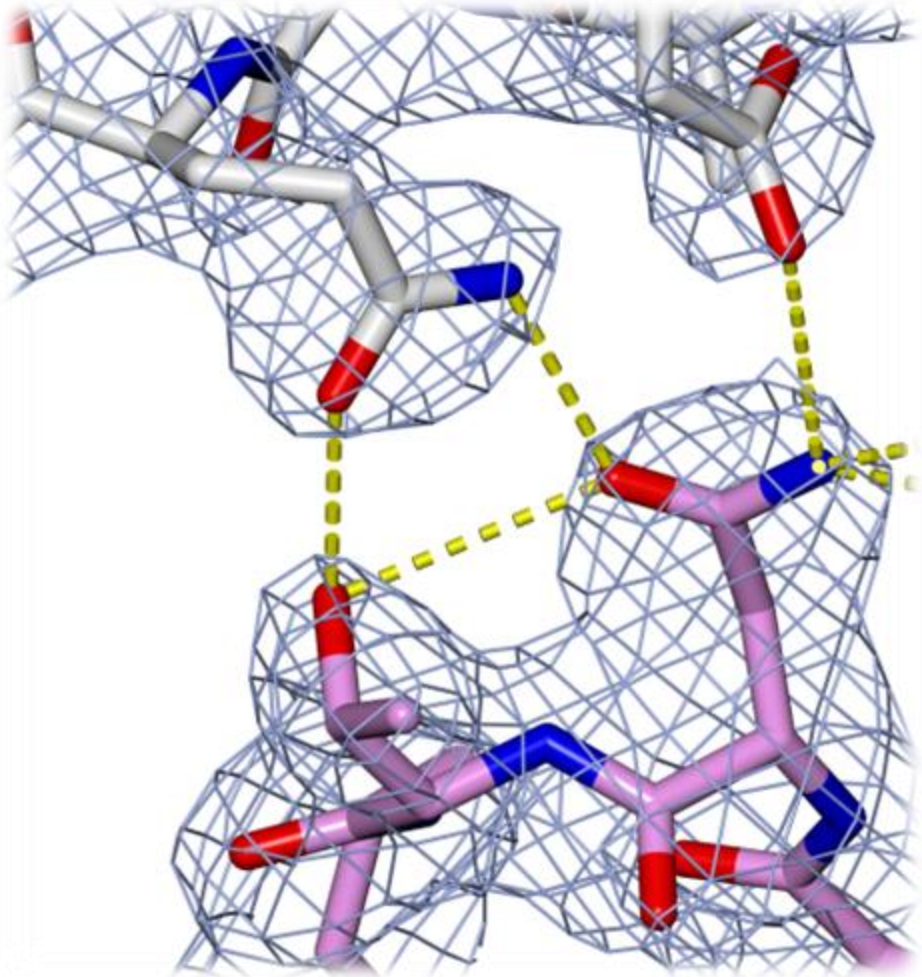
- R-free from 31.6% to 26.7%
  - 7σ improvement
- Moved from 34th to the 99th quality percentile in MolProbity



Herceptin – HER2 interface



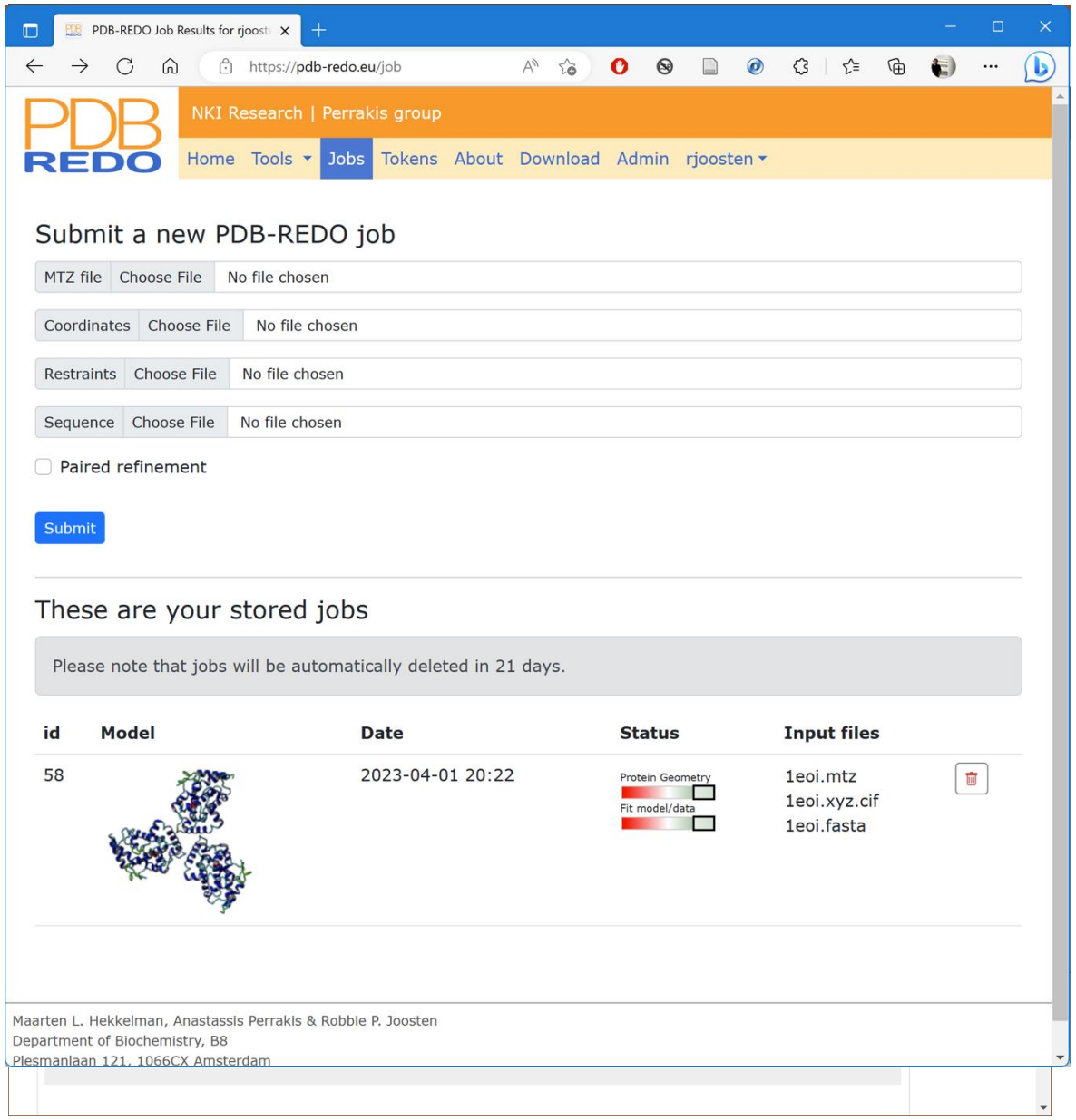
PDB



PDB-REDO

Using PDB-REDO

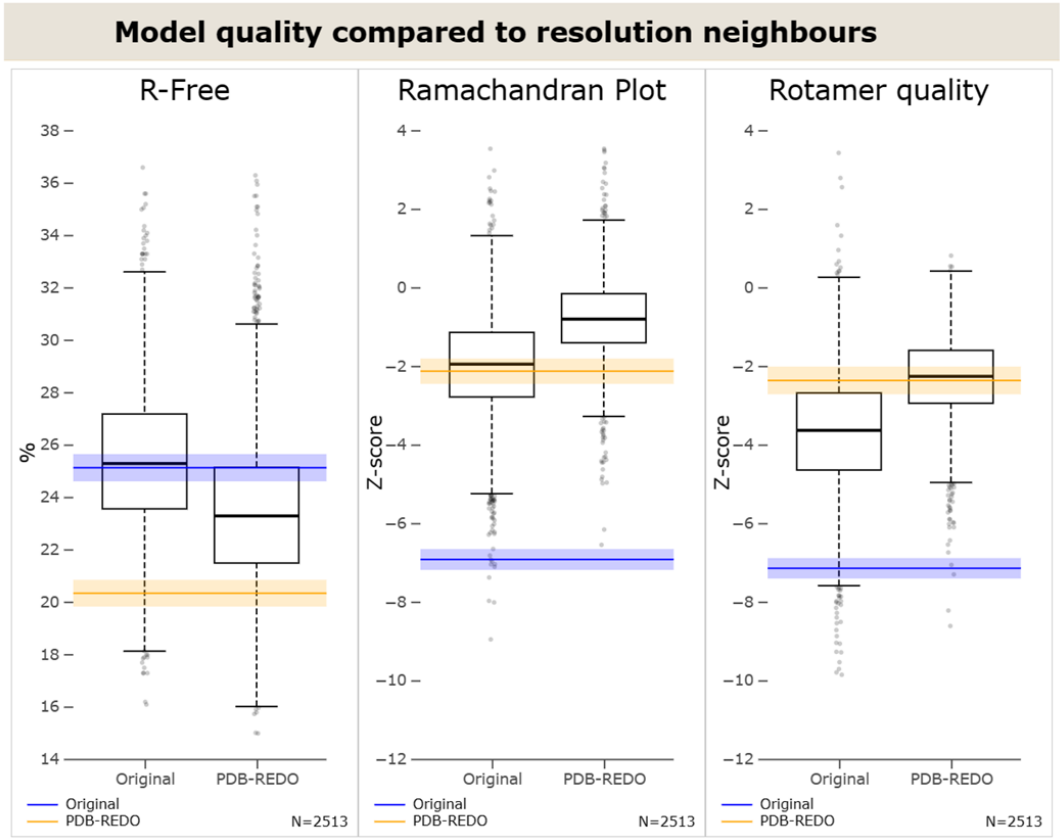
- Use CCP4 or pdb-redo.eu website
- Password protected accounts
  - Users from all over the world
- Dashboard with previous jobs
  - Summary sliders
- Submit model, MTZ, sequence
  - Restraints are optional
  - 8500 jobs run in 2024



PDB-REDO output

- New model + new map coefficients
- Optimised settings for Refmacat + ready-made extra restraints
- Model quality indicators

Validation metrics from PDB-REDO		
	PDB	PDB-REDO
Crystallographic refinement		
R	0.2094	0.1660
R-free ⚠	0.2512	0.1988
Bond length RMS Z-score	0.526	0.423
Bond angle RMS Z-score	0.748	0.668
Model quality raw scores | percentiles		
Ramachandran plot normality	1	35
Rotamer normality	4	52
Coarse packing	94	99
Fine packing	20	82
Bump severity	10	42



Using PDB-REDO is little work, but it helps you make better models

Writing up

PDB deposition and publication

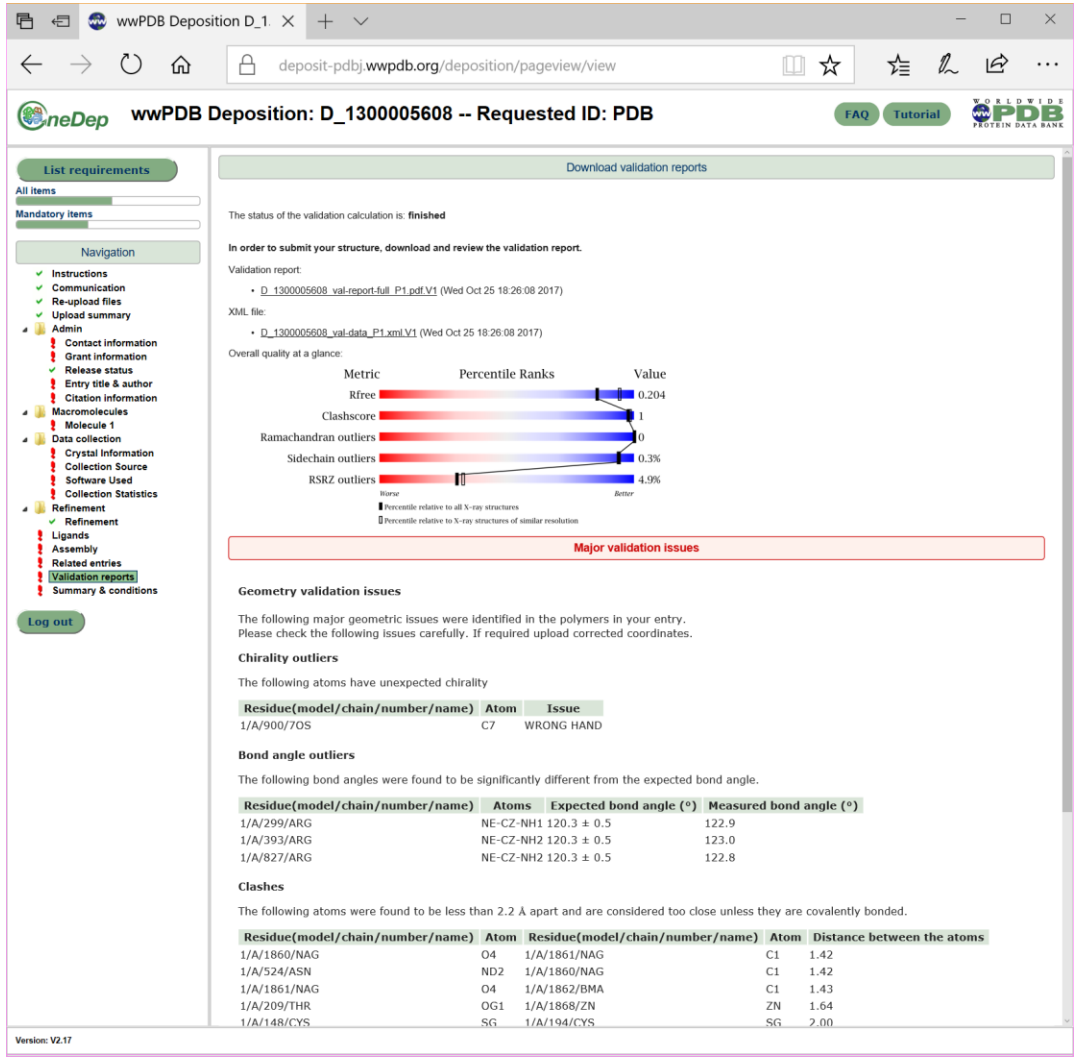
- Share your model with the entire community
  - Get credit for your work
  - The community paid for your research
- Prepare for deposition:
  - Finish your model
  - Check it with the PDB validation server
  - Fix real problems before starting the deposition
- Be comprehensive in describing your model and its creation:
  - Both at the PDB and in your manuscript
  - Better description gives more useful model for others

Preparing for PDB deposition

- Use ‘Prepare files for deposition’ task in CCP4 cloud
  - Takes final model, reflection data, target sequences, and if available logs from XIA2 and Aimless
- Collect additional log files from software you used
  - Data integration, scaling and merging, structure solving
- Check reflection data: you should have all reflections + anomalous data
- Collect experimental info
  - Protein expression, sequence of construct, crystallisation conditions
  - Diffraction experiment: how, where, and when
- Other needed info
  - Names and ORCID of depositors (not authors), funding information
  - Title of model (not the title of the paper), release plan

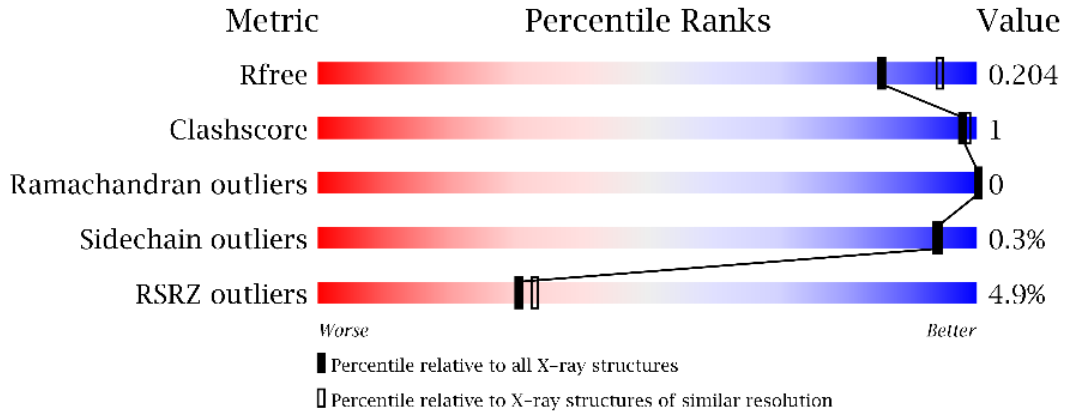
PDB deposition

1. Register in the OneDep system
   - You can use your ORCID
2. Upload your data and fill out the forms
   - Hold model for publication, hide sequence and title if you want
   - Pay extra attention to ligands
     - Check the interpretation of your ligand
3. Check preliminary validation report
4. If no solvable issues, finish deposition



PDB annotation

- You will get an email that your annotated model is ready
  - You now have a PDBid for your model
- Answer all questions that the annotator has
  - If you disagree on something, explain why you think you are right
- Once the annotation is done you get the final validation report
  - Having a few outliers is not a problem
    - PDB may use different targets for model geometry
    - RSRZ outliers is a point of discussion
  - You can still replace your model if there is something you need to fix
    - Even after release



Things to have in your manuscript

- Science and stuff
- A comprehensive ‘(supplemental) methods’ section
  - Say what you did and which tools you used
  - Be prepared to explain why you did certain things
- A finished, deposited model in the PDB
- A PDB validation report to submit with the manuscript
- A comprehensive (supplemental) Table 1 with
  - Key data quality indicators (overall and high-resolution shell)
  - Model quality metrics (but which ones…?)

Summary

- There are a lot of different validation programs and metrics
- None of them are perfect; use common sense
- No model is perfect; do the best you can; let PDB-REDO help
- Always clearly describe what you did, and know why

Acknowledgements

NETHERLANDS CANCER INSTITUTE
ANTONI VAN LEEUWENHOEK

- Ida de Vries
- Maarten Hekkelman
- Bart van Beusekom
- Krista Joosten
- Anastassis Perrakis
- RHPC facility

Radboudumc
university medical center

- Gert Vriend

UNIVERSITY of York

- Jon Agirre

- CCP4 developers
- PDB annotators

MRC Laboratory of Molecular Biology

- Garib Murshudov
- Paul Emsley
- Rob Nichols
- Keitaro Yamashita

x3DNA DSSR

- Xiang-Lun Ju

Fragment Screen
CCP4
Oncode Accelerator Foundation
iNEXT DISCOVERY
EOSC-Life
janssen
PHARMACEUTICAL COMPANIES OF Johnson&Johnson
NWO