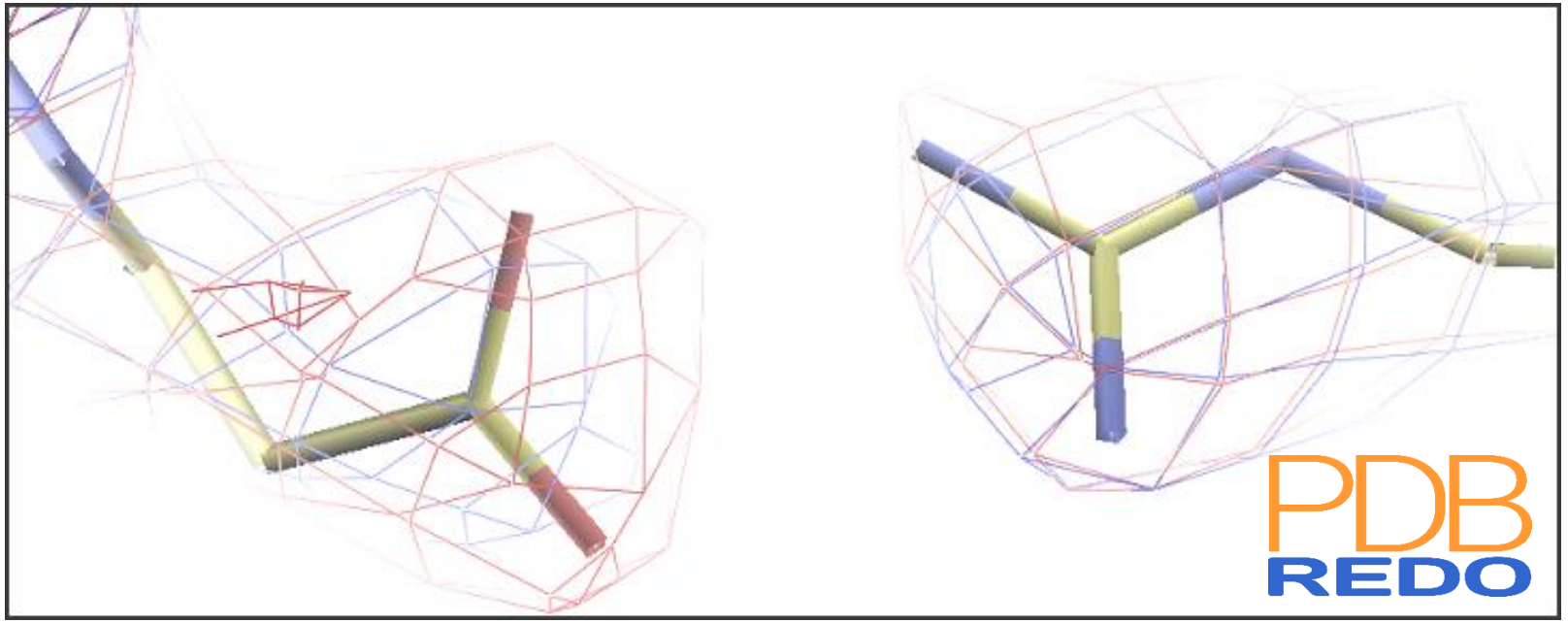
How good is my model?

And can it be improved?



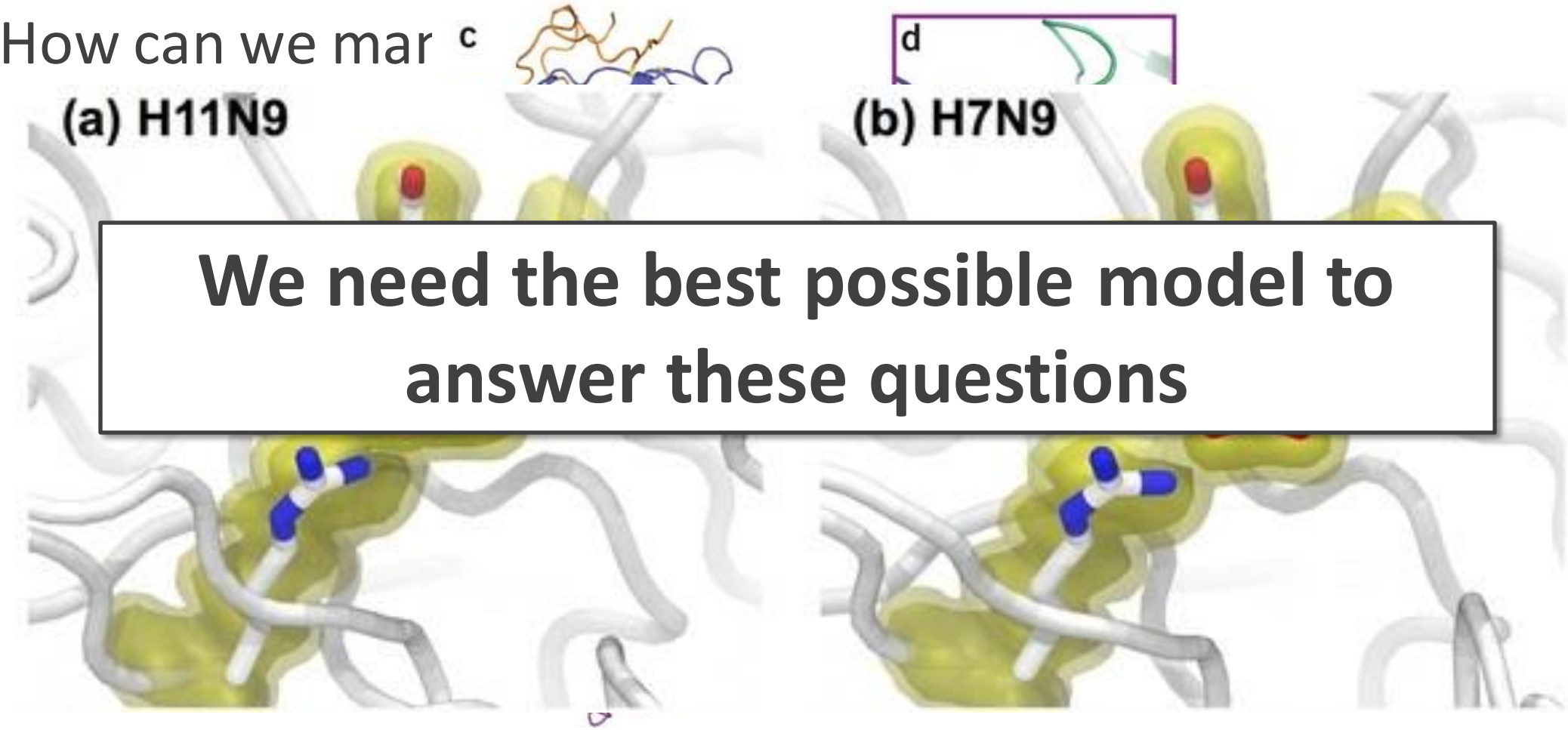
Robbie P. Joosten

Netherlands Cancer Institute

CCP4-DLS school 2023

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?

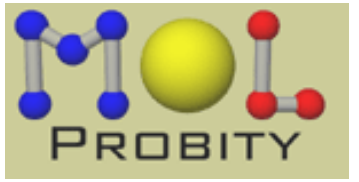
1. Use validation when making the model
   - Check model vs. data and vs. prior knowledge
   - Focus on outliers (fix or explain them)
   - Know the things that can go wrong

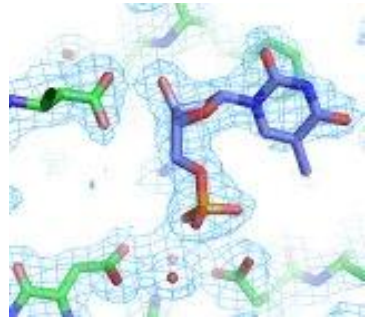
2. Optimise the model
   - Focus on what can be improved
   - Choose best refinement parameters and restraints
   - Rebuild parts of the model
   - PDB-REDO automates this

Validation

Need to know

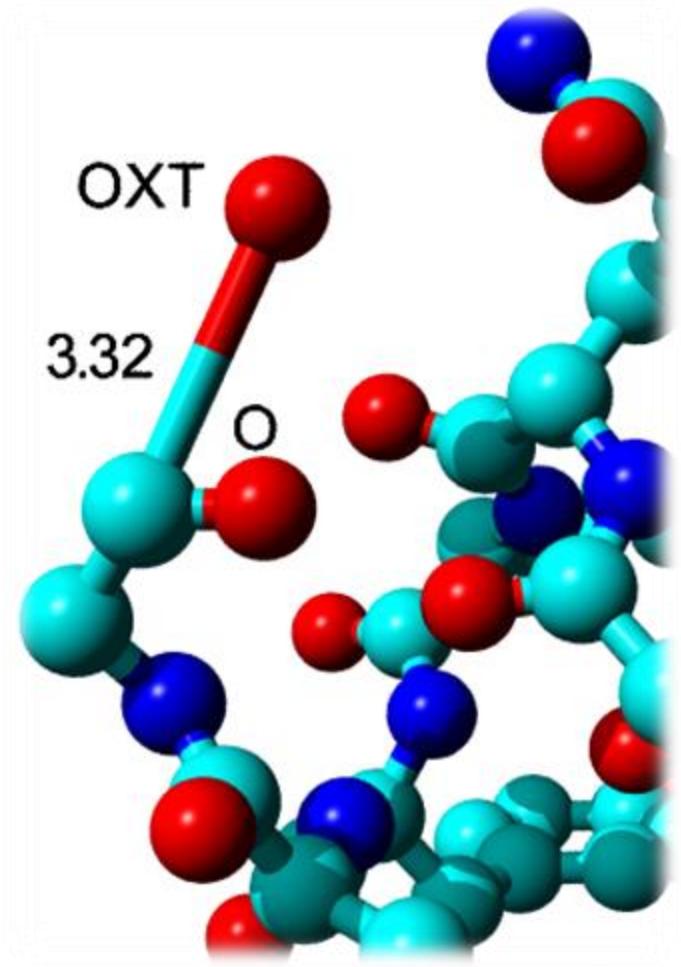
- Check the validity and value of a model
  - Accuracy and precision
- Many different software tools
  - General: MolProbity, PDB validation server
  - Non-protein: CheckMyMetal, Privateer, DNATCO
  - Tools may check the same things differently
- Not a substitute for common sense
  - False positives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





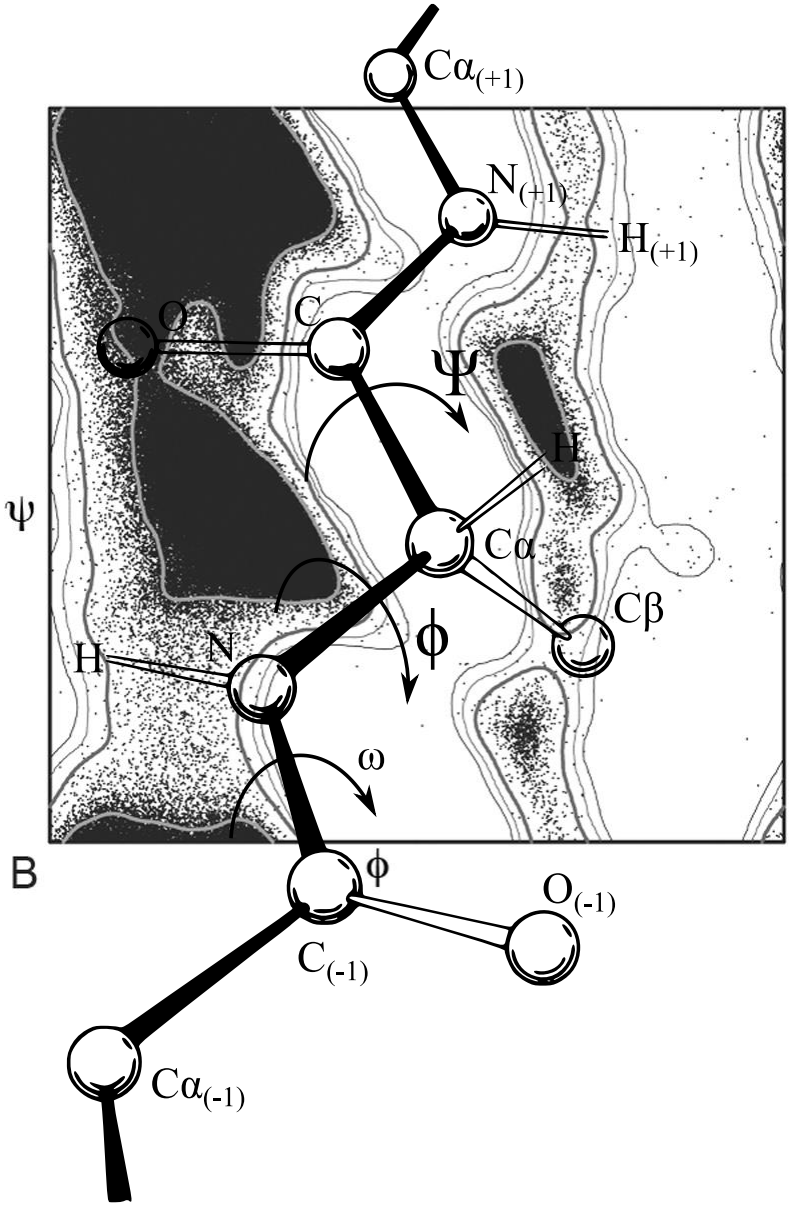
Bonds and angles

- Individual outliers
  - Usually fitting errors
  - Express deviation in terms of SD (Z-scores)
  - Example: Z = 105
- Large overall deviations from ideal values
  - Express as rmsZ, not rmsd
    - Should be < 1.000
  - Use tighter restraints



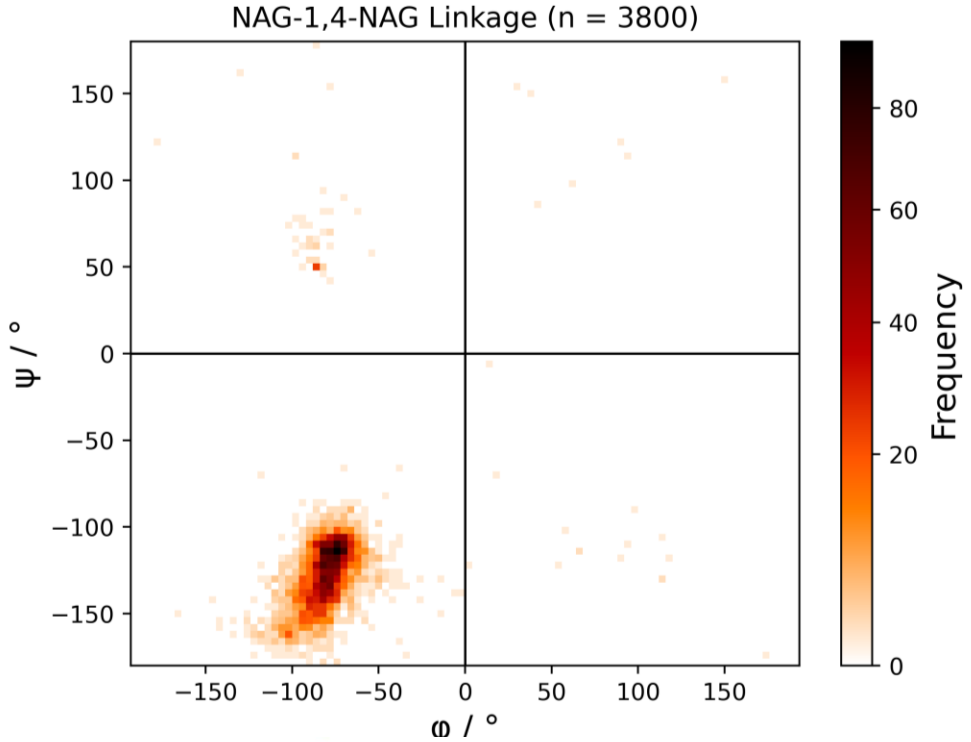
Backbone torsion angles

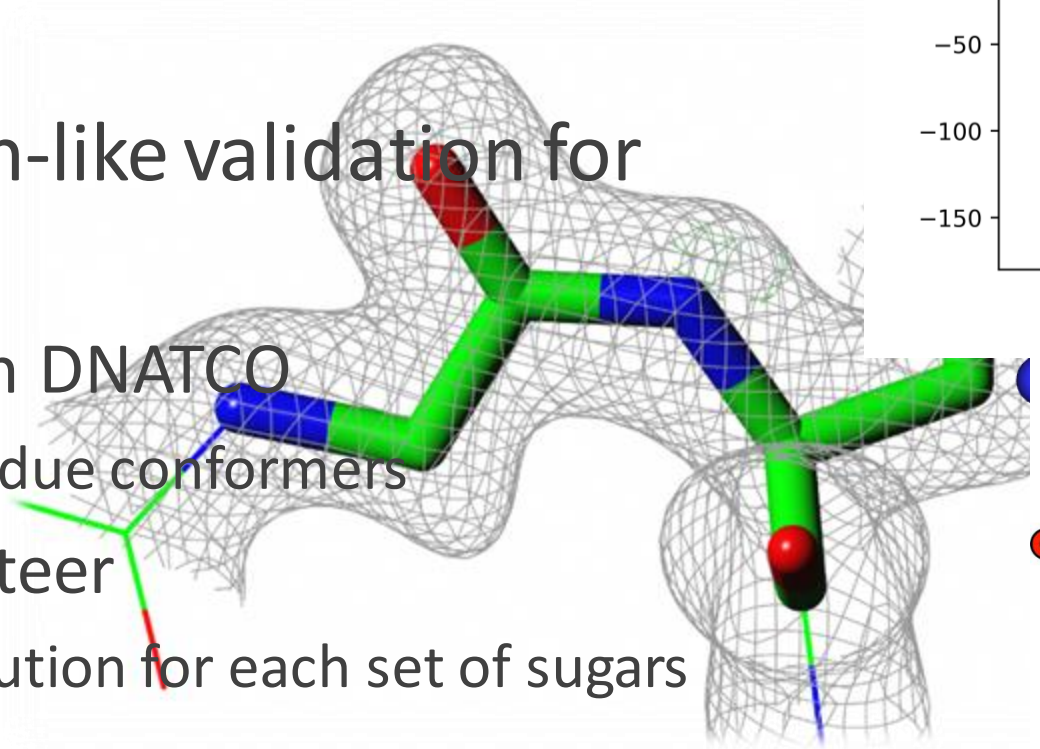
- Ramachandran plot
  - ϕ and ψ angles
  - Not all conformations are possible
  - Compare to the whole PDB or a subset
- Different implementations
  - MolProbity and COOT: preferred, okay, outlier
    - Good for finding specific individual problems
    - Check severity of outlier on the plot
  - Tortoize, MolProbity, Phenix: overall Z-score
    - Good for checking building and refinement progress
    - Watch out for scores that are too good to be true

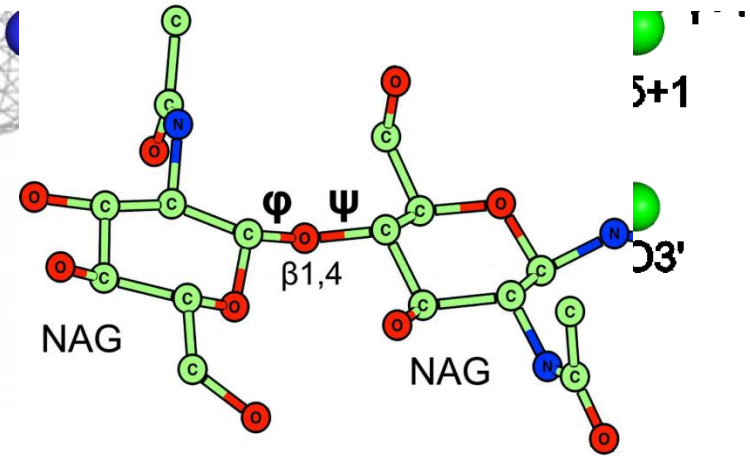


Backbone torsion angles

- Peptides are flat
  - ω angle is ~180° (*trans*) or ~0° (*cis*)
  - Fitting errors or the wrong restraints cause outliers
- Ramachandran-like validation for non-proteins
  - Nucleic acids in DNATCO
    - CONFAL: 2-residue conformers
  - Sugars in Privateer
    - Specific distribution for each set of sugars

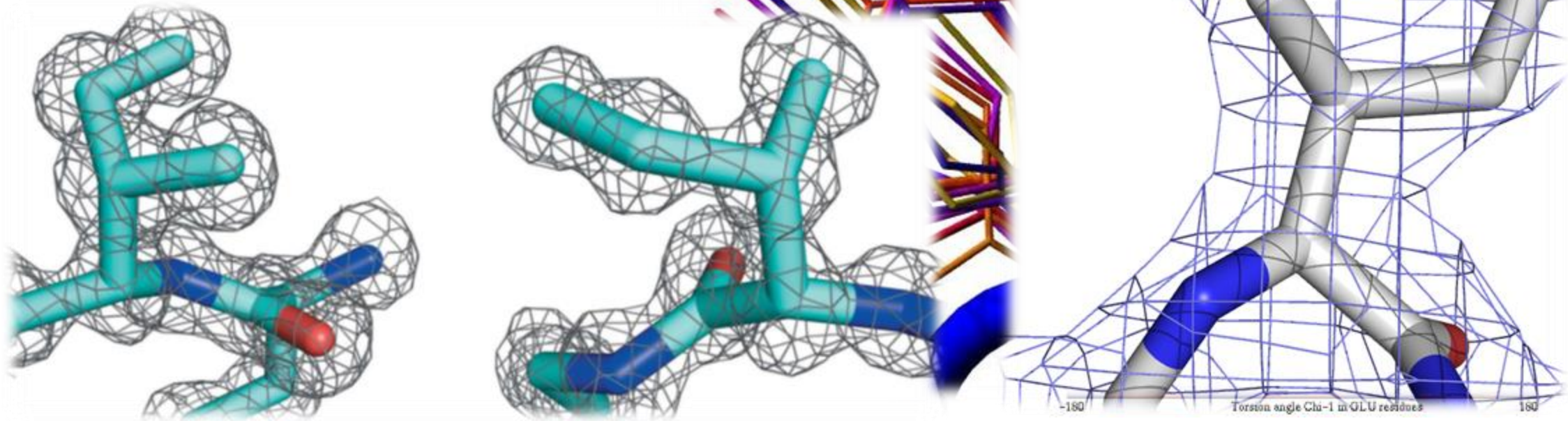






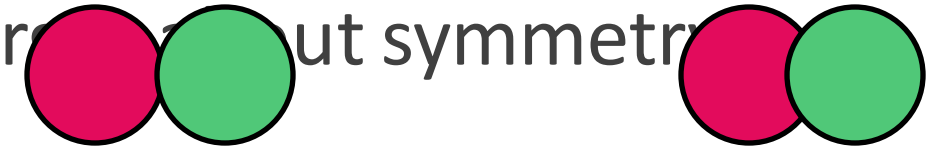
Side chain torsion angles

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



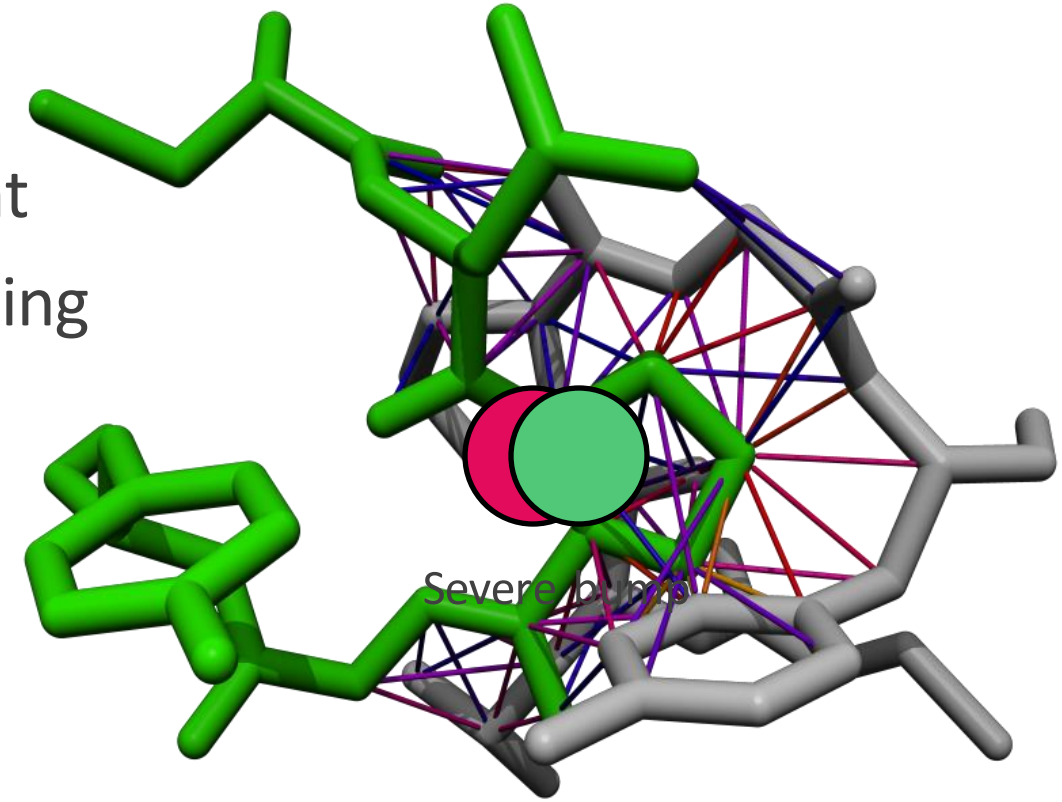
Bumps/clashes

- 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



Normal contact

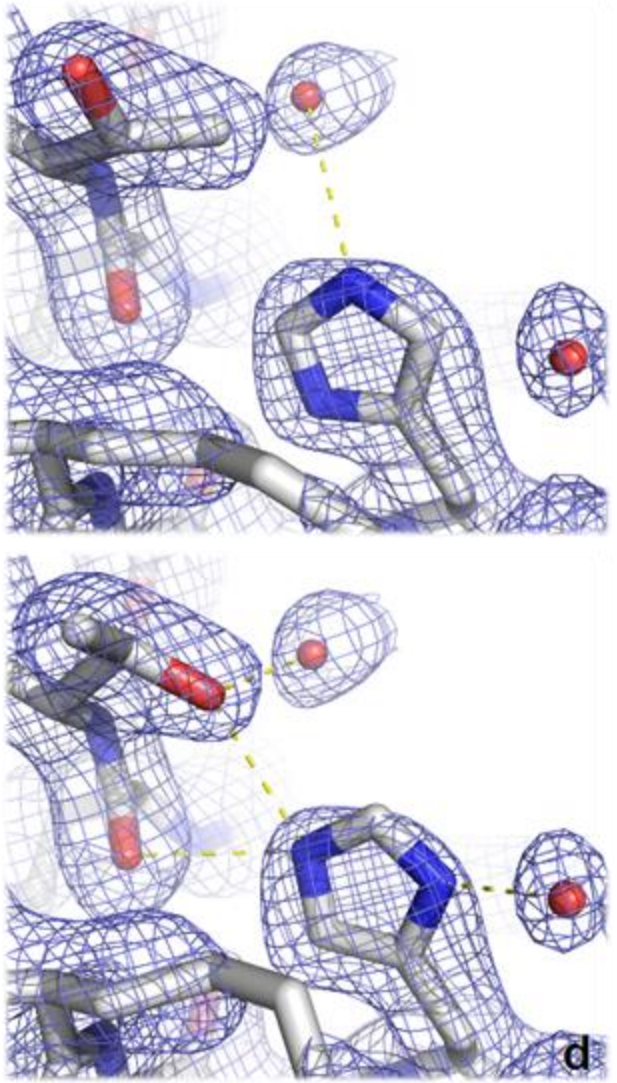
Mild bump



Severe bump

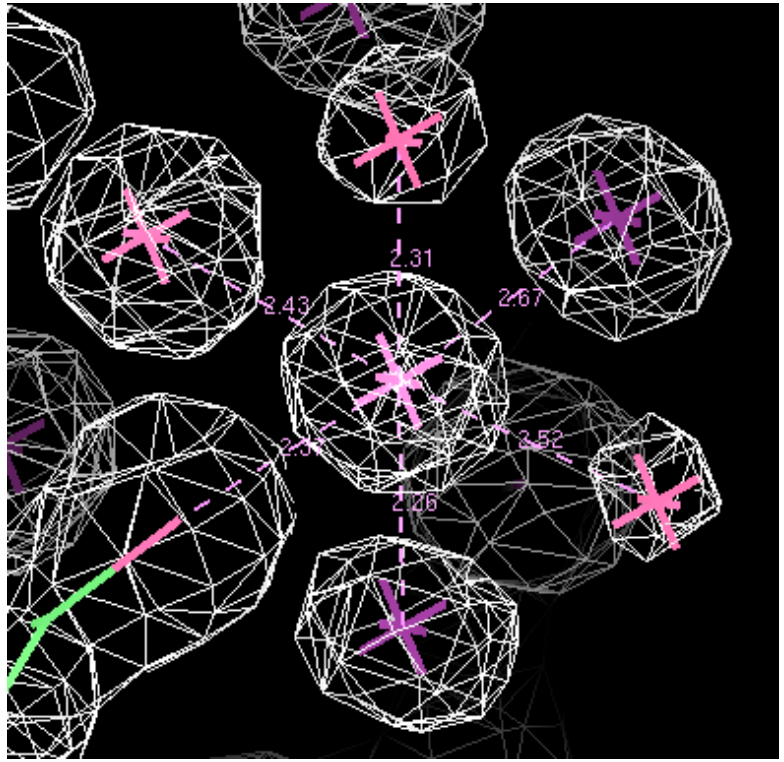
Hydrogen bonds

- 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

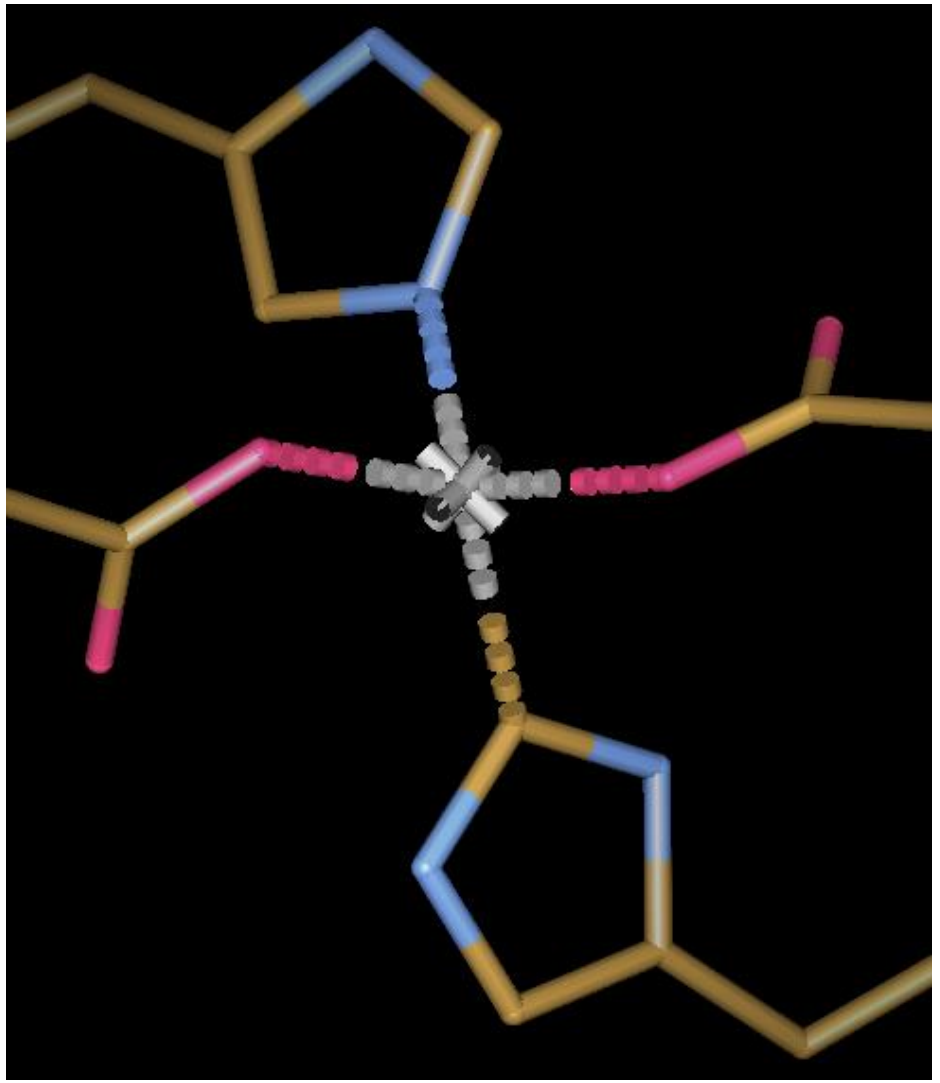


Metal ion validation

- 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

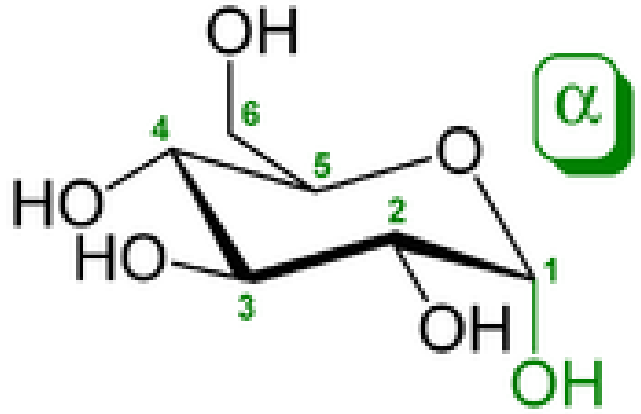
Metal ions

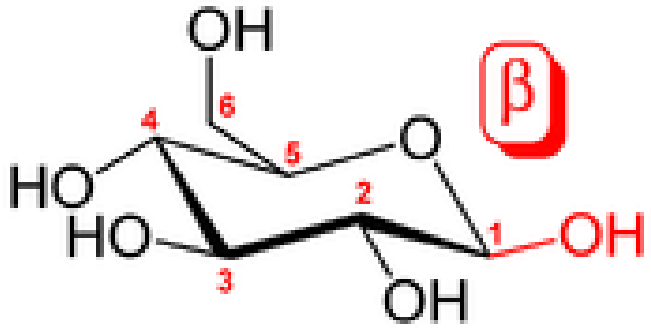
- Na, Mg, K, Ca prefer being coordinated by oxygen atoms
  - Flip Asn or Gln side chains if needed
- Carbons usually do not coordinate metals
  - Flip His side chains if needed
  - Cyanide and carbon-monoxide are exceptions



Sugars are complicated

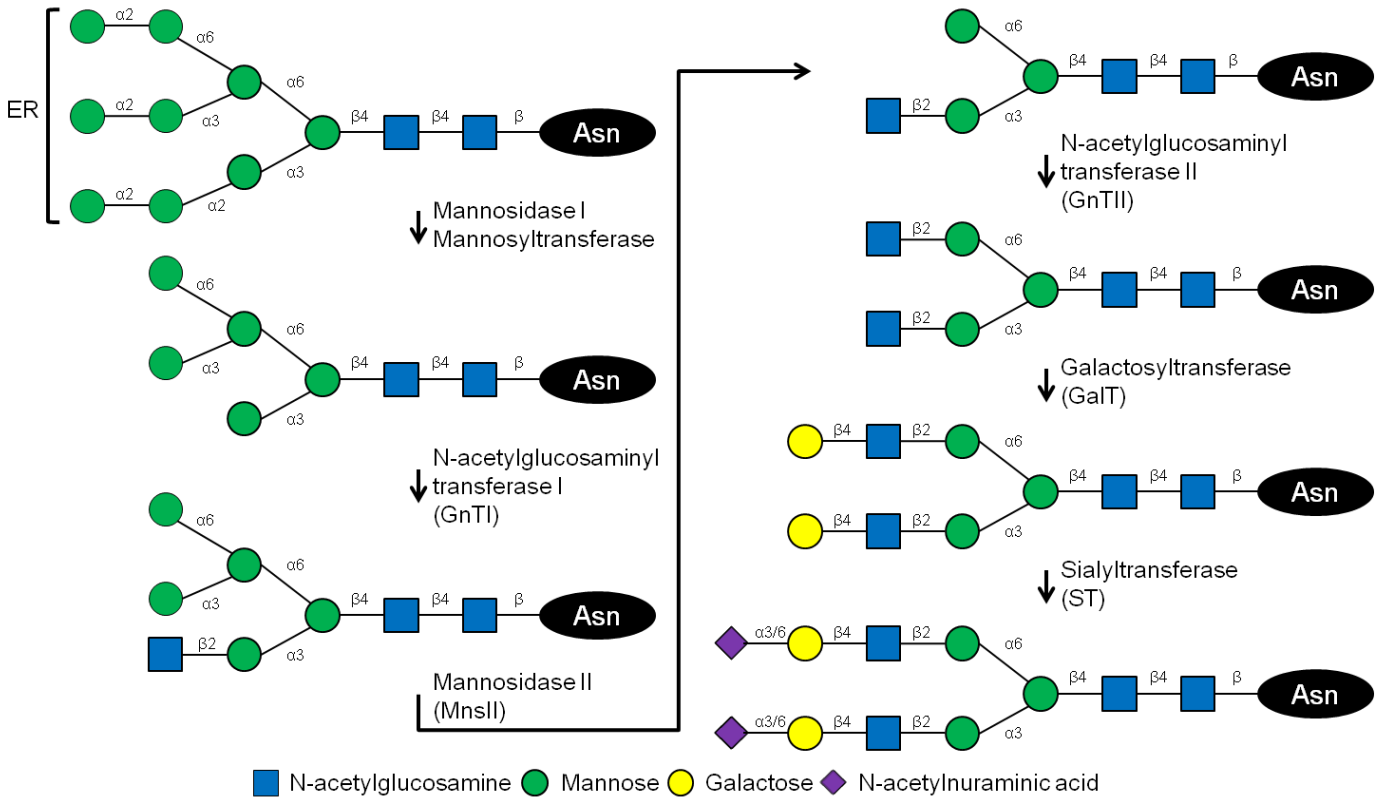
- Small differences matter in biology
- Maps are often difficult to interpret
- Coordinates and residue name must match sugar identity
  - Or your refinement will go wrong
- Bonds between sugars are common
  - ‘Always’ from C1 to an oxygen (O1 is lost)
  - Original position of the O1 describes the linkage type (α or β)
    - Should be reflected in the link restraints





Sugar validation

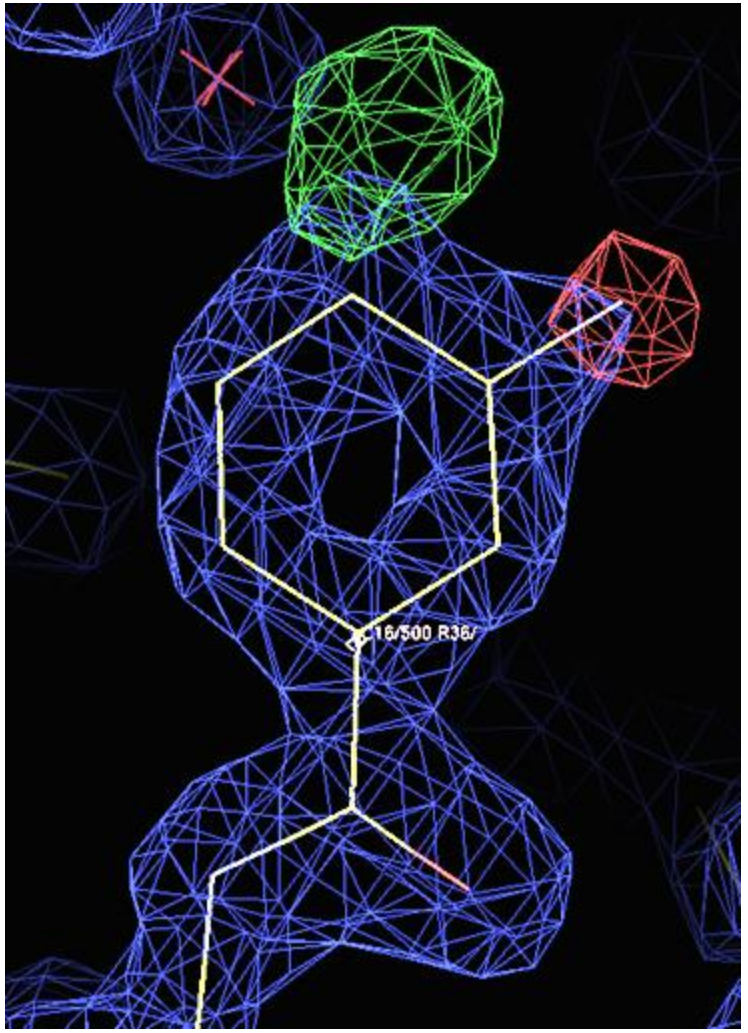
- Privateer checks conformations and overall tree consistency with glycomics databases



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?
   - Check the restraints themselves
   - Check against restraints
   - Check against chemistry
     - Especially chirality and planarity



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

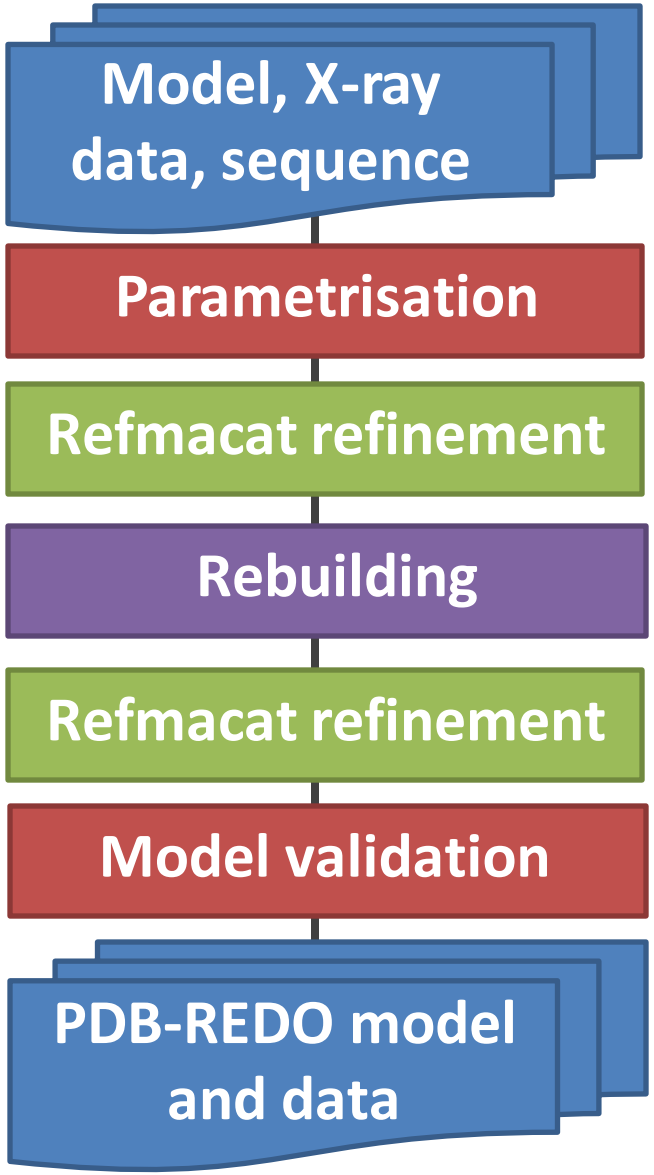
Model optimization = making choices

- 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

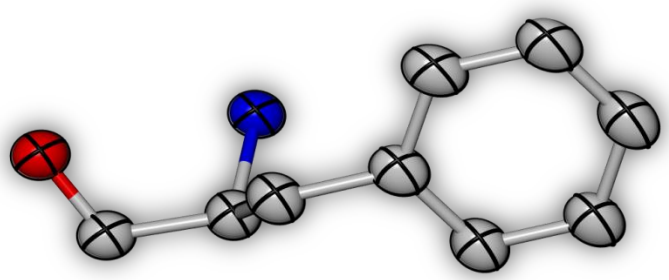
PDB-REDO

- 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



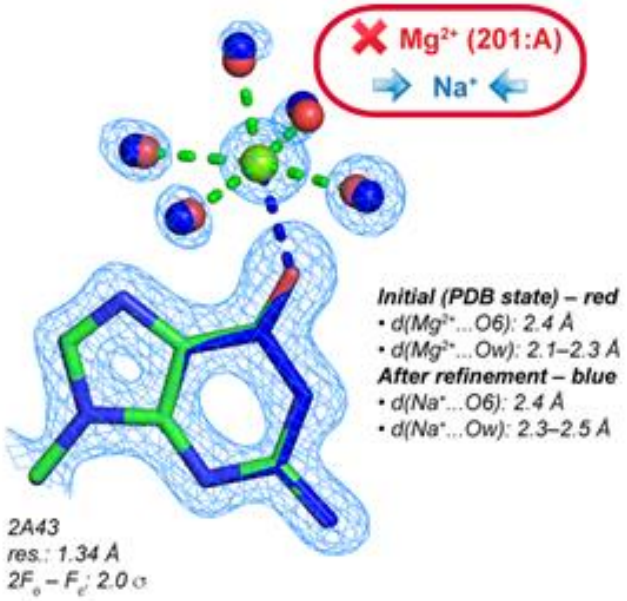
Features and algorithms

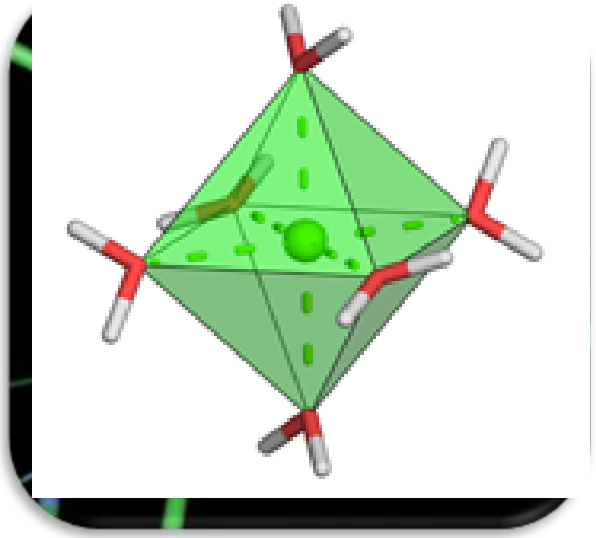
- NCS and twinning treated automatically
  - Warns for space group errors
- Establish high resolution cut-off through paired refinement
- B-factor model selection:
  - Refine alternative models and select best one
    - Isotropic, anisotropic, or flat B-factors
    - One TLS group per chains, user-provided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



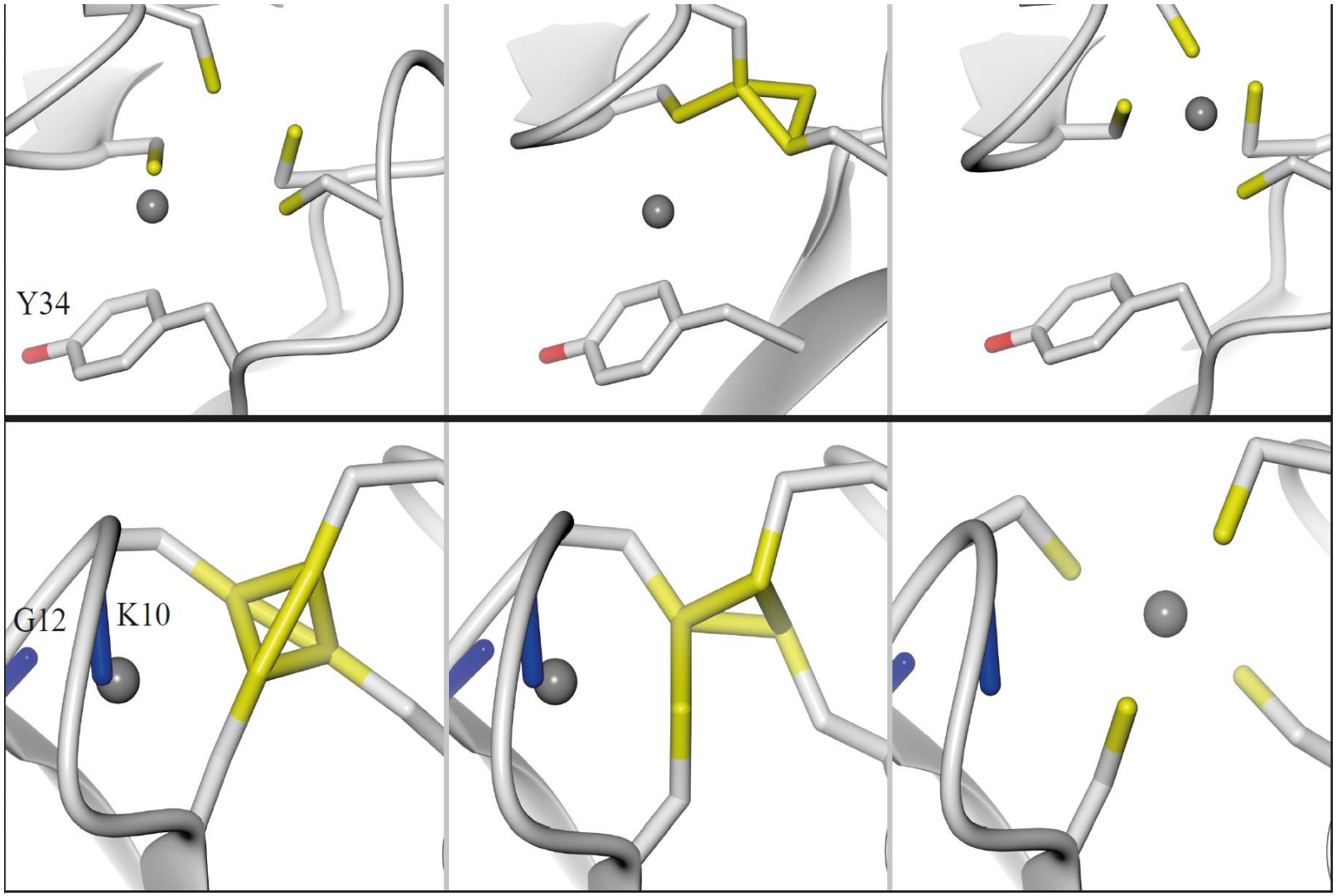
Features and algorithms

- Case specific restraints:
  - Jelly-body restraints (low 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





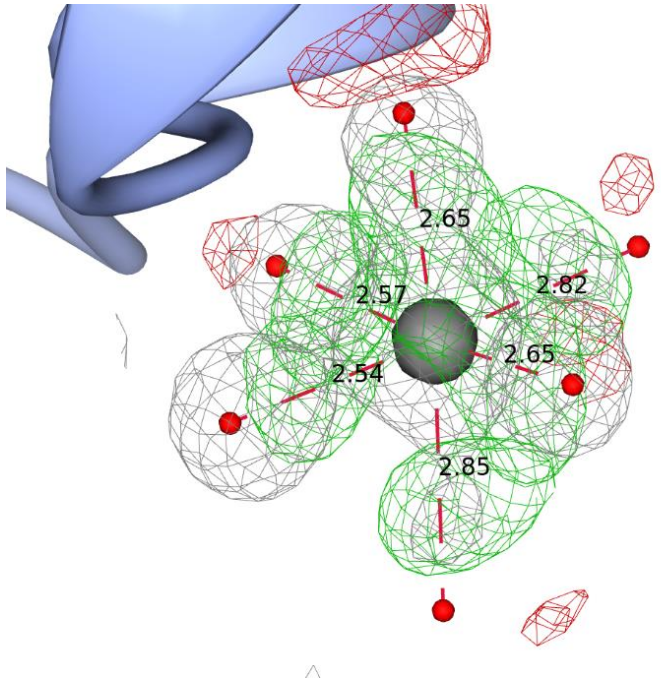
Added value of zinc restraints



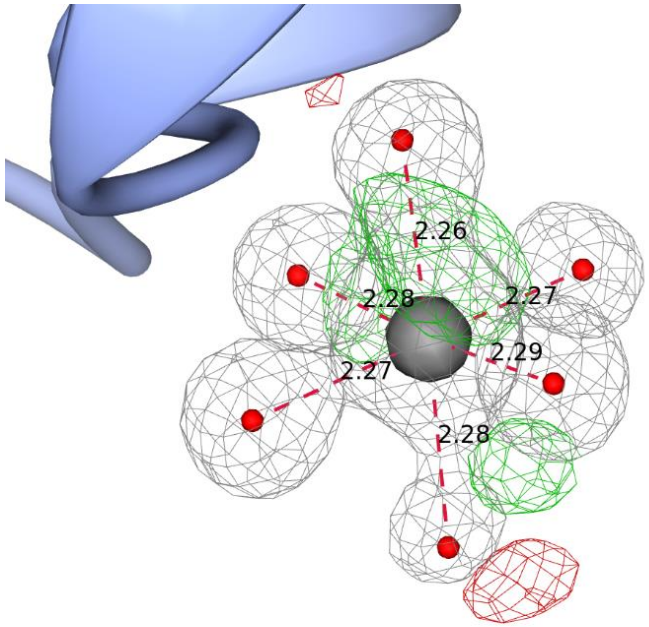
PDB PDB-REDO PDB-REDO & platonyzer

Added value of octahedral restraints

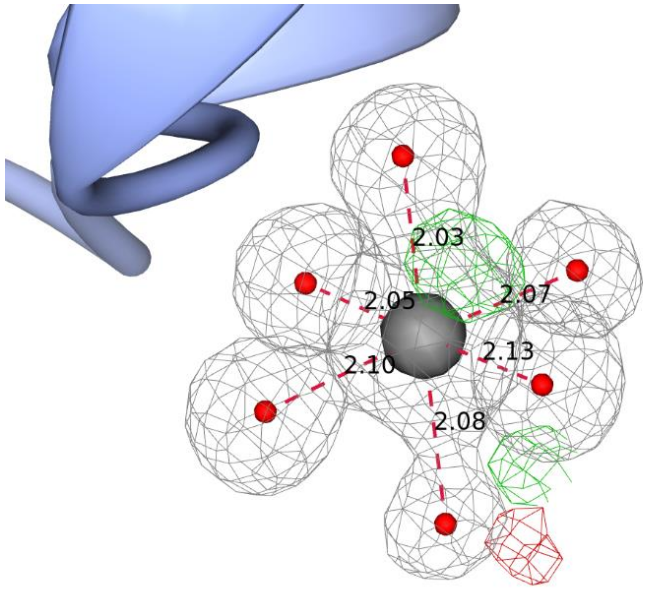
- Magnesium site with sodium modelled becomes Mg-like
  - Opposite can happen as well



PDB



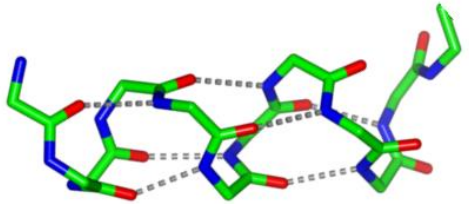
PDB-REDO

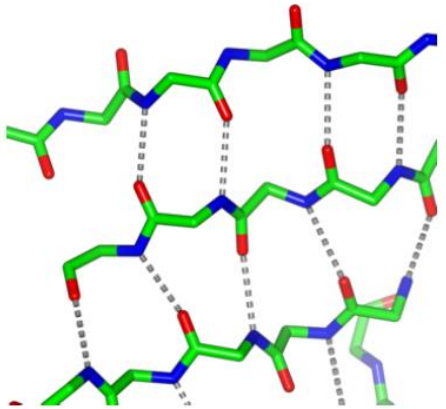


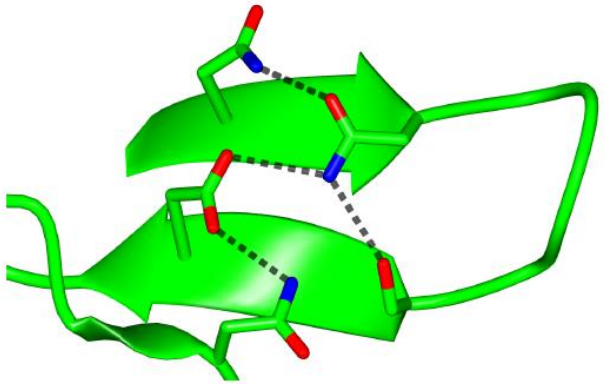
PDB-REDO + platonyzer

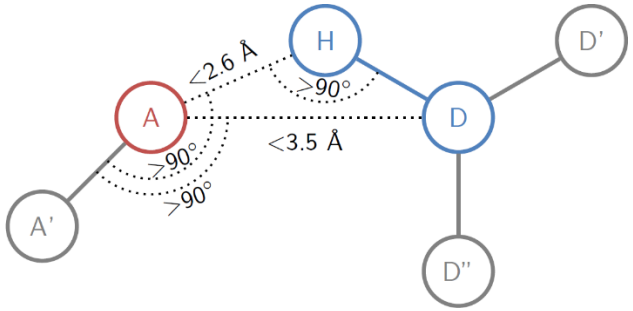
Homology-based H-bond restraints

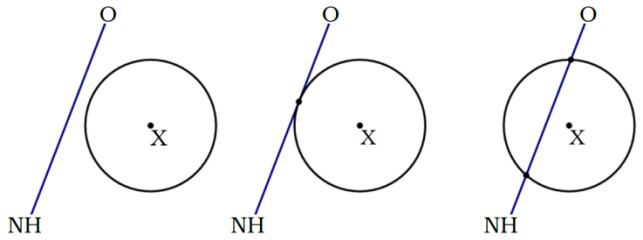
- H-bonds are real ‘long-distance’ interactions
  - Can be used as distance restraints
- Omnipresent in protein structures
  - Enough extra information
  - Side-chain H-bonds are very structure specific
- We know what they should look like
  - Detect and filter real H-bonds





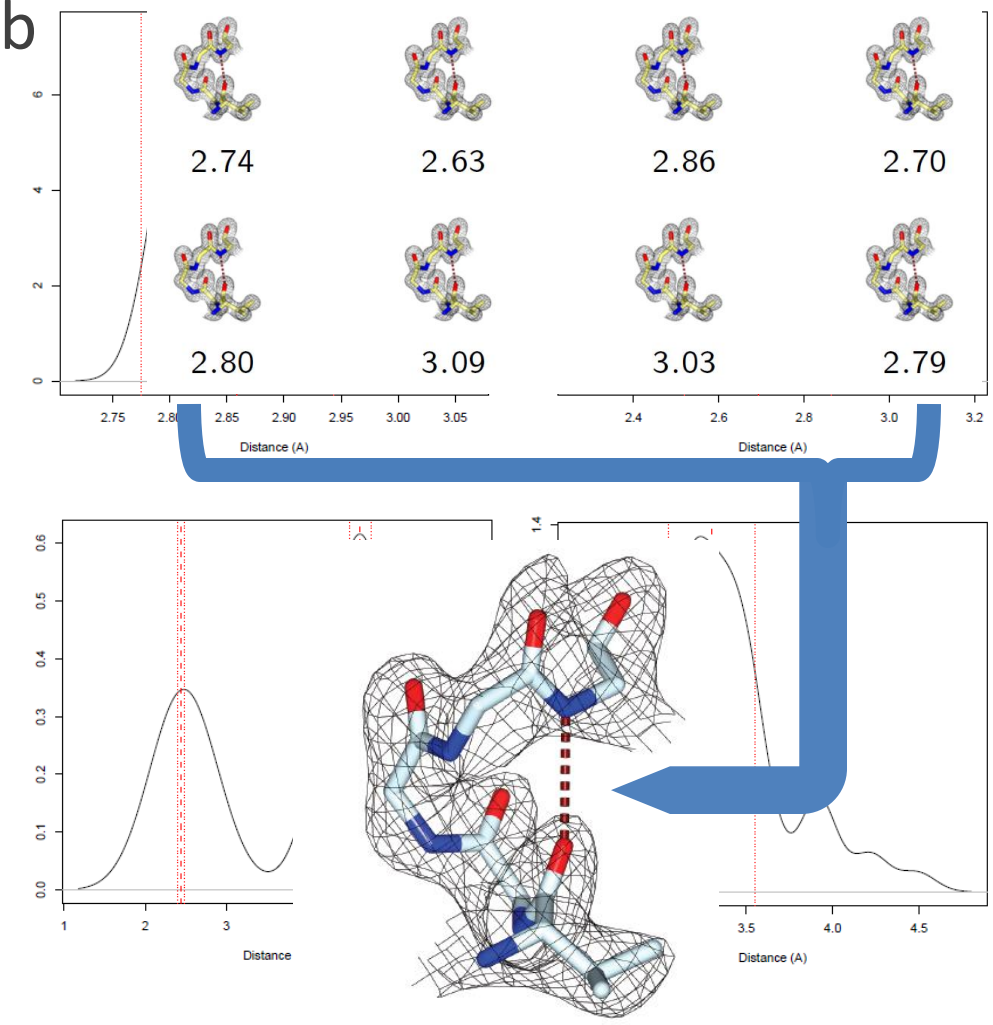






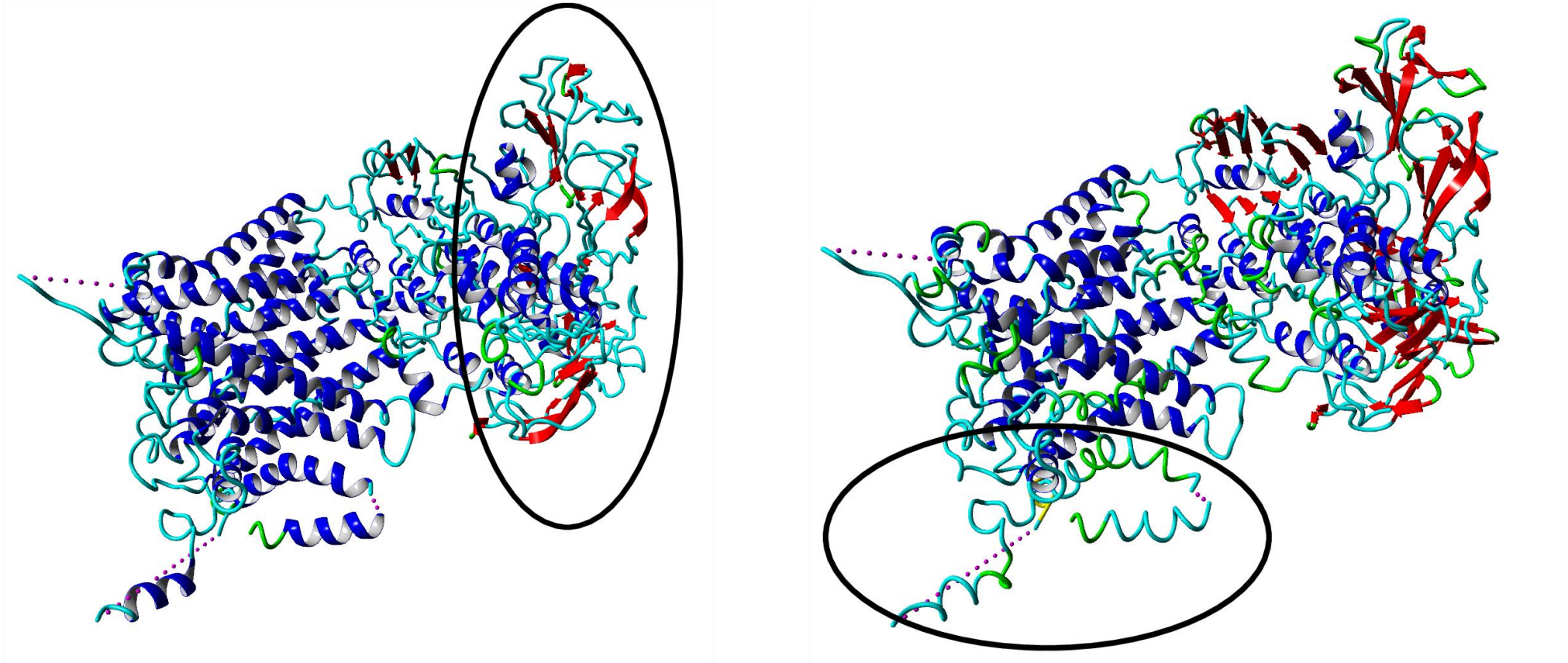
HODER: HOmology-DErivid Restraints

- BLAST model sequence against PDB-REDO db
  - Use models with sequence identity > 70% and resolution ≤ model
- Find equivalent H-bonds in homologs
  - Filter: same secondary structure
  - Filter: same side-chain rotamer
- If > 5 distances, fit target and SD
  - Higher SD, lower relative restraint weight
- Works for 82% of protein-protein H-bonds
  - Fallback to 118 types of atom and secondary structure specific H-bond restraints
  - Mined from 10k high quality structure models
    - 4 million observations



Added effect of homology restraints

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

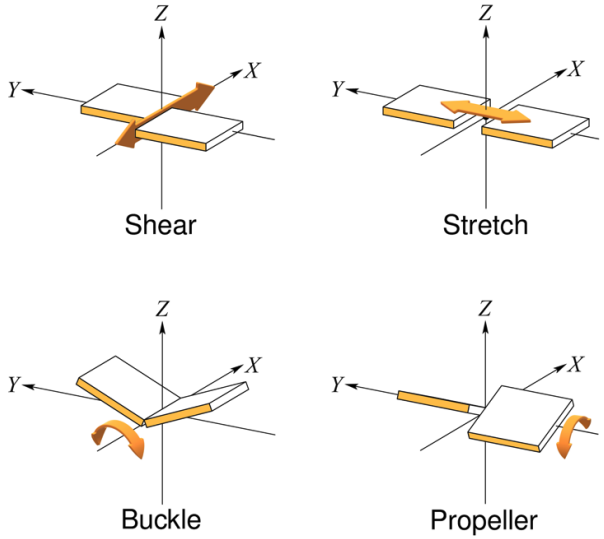


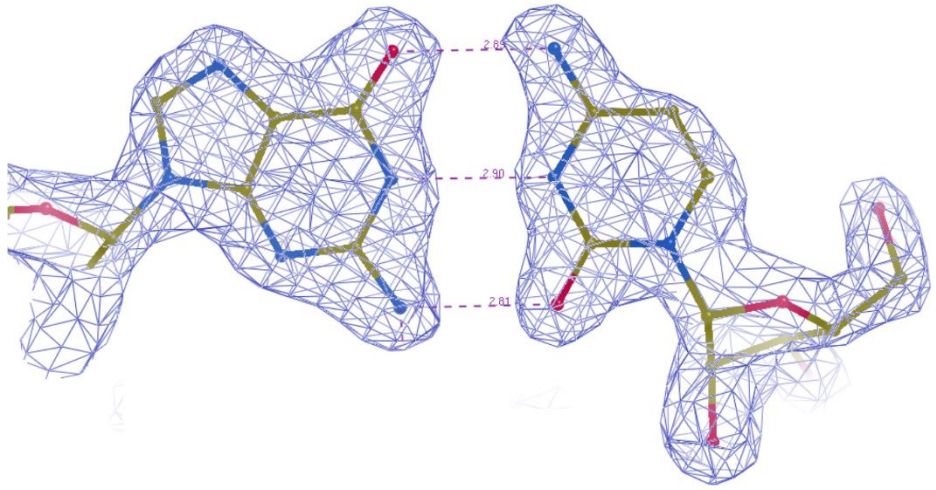
PDB

PDB-REDO new

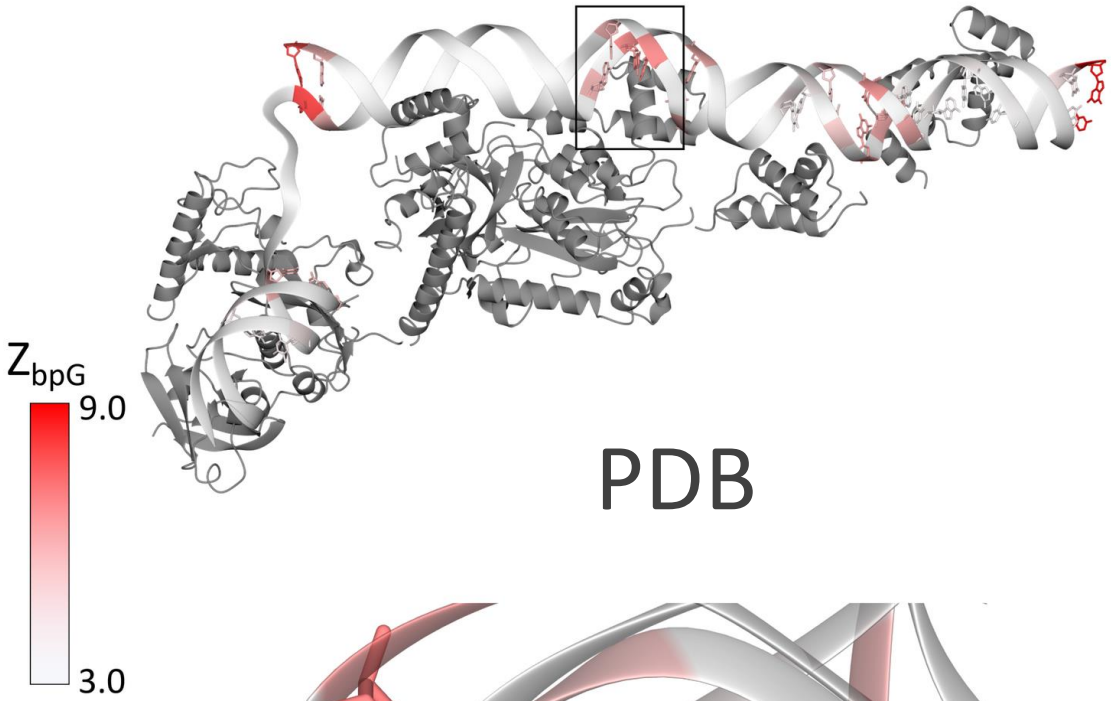
Nucleic acid features

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

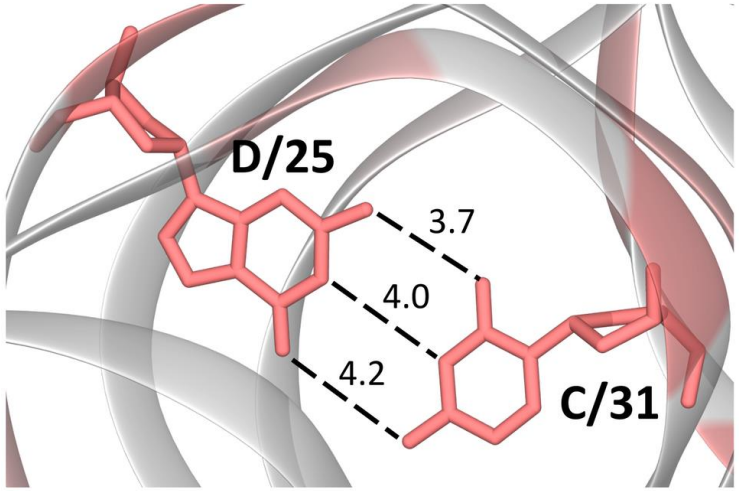


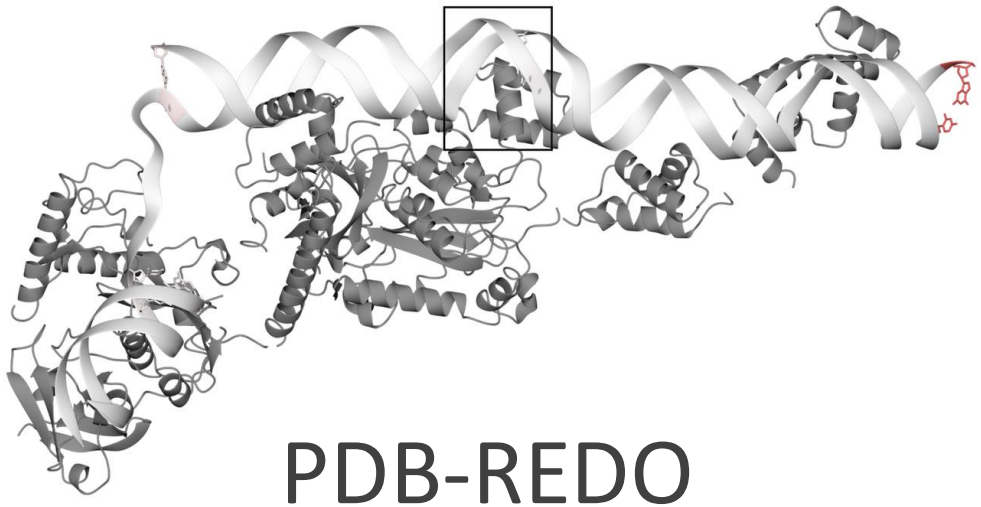


Effect of nucleic acid restraints

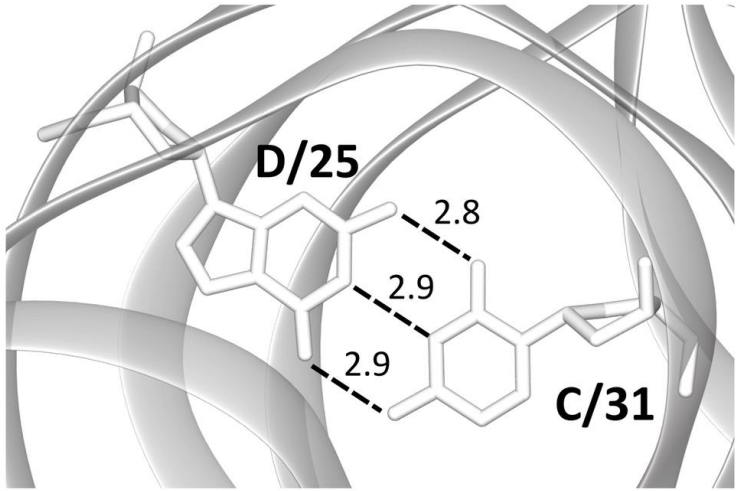


PDB



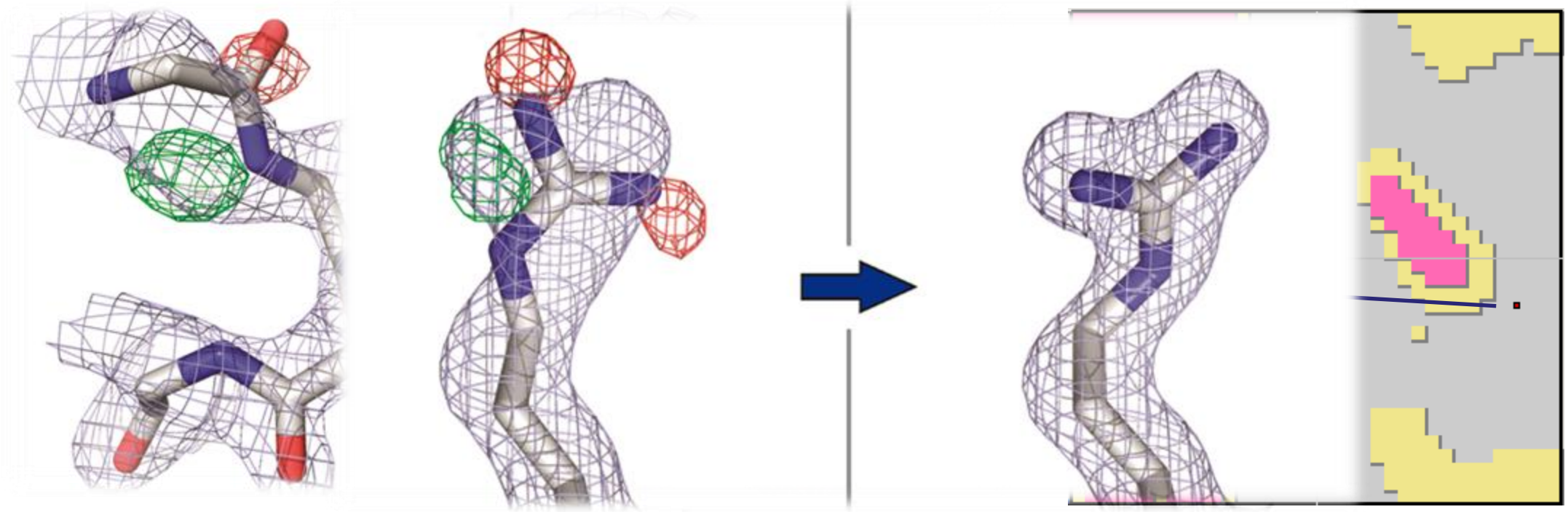


PDB-REDO



Model rebuilding

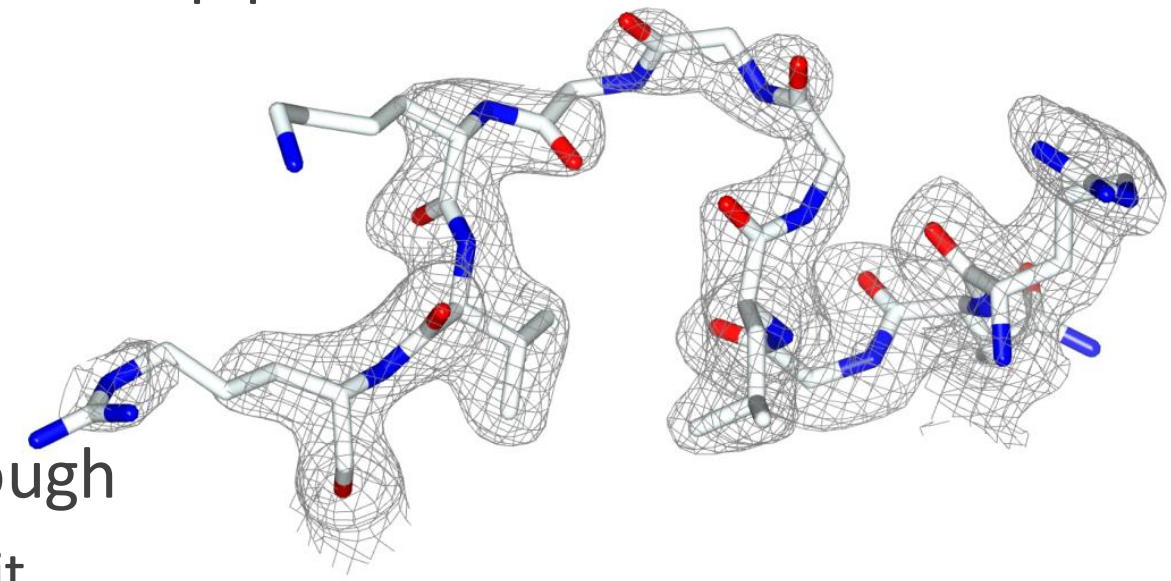
- Peptide flipping:
  - If a flip improves density fit and Ramachandran plot: accept flip
- Side-chain rebuilding and completion:
  - Add missing side-chains, rebuild existing ones, flip His/Asn/Gln if needed



Homology-based loop building

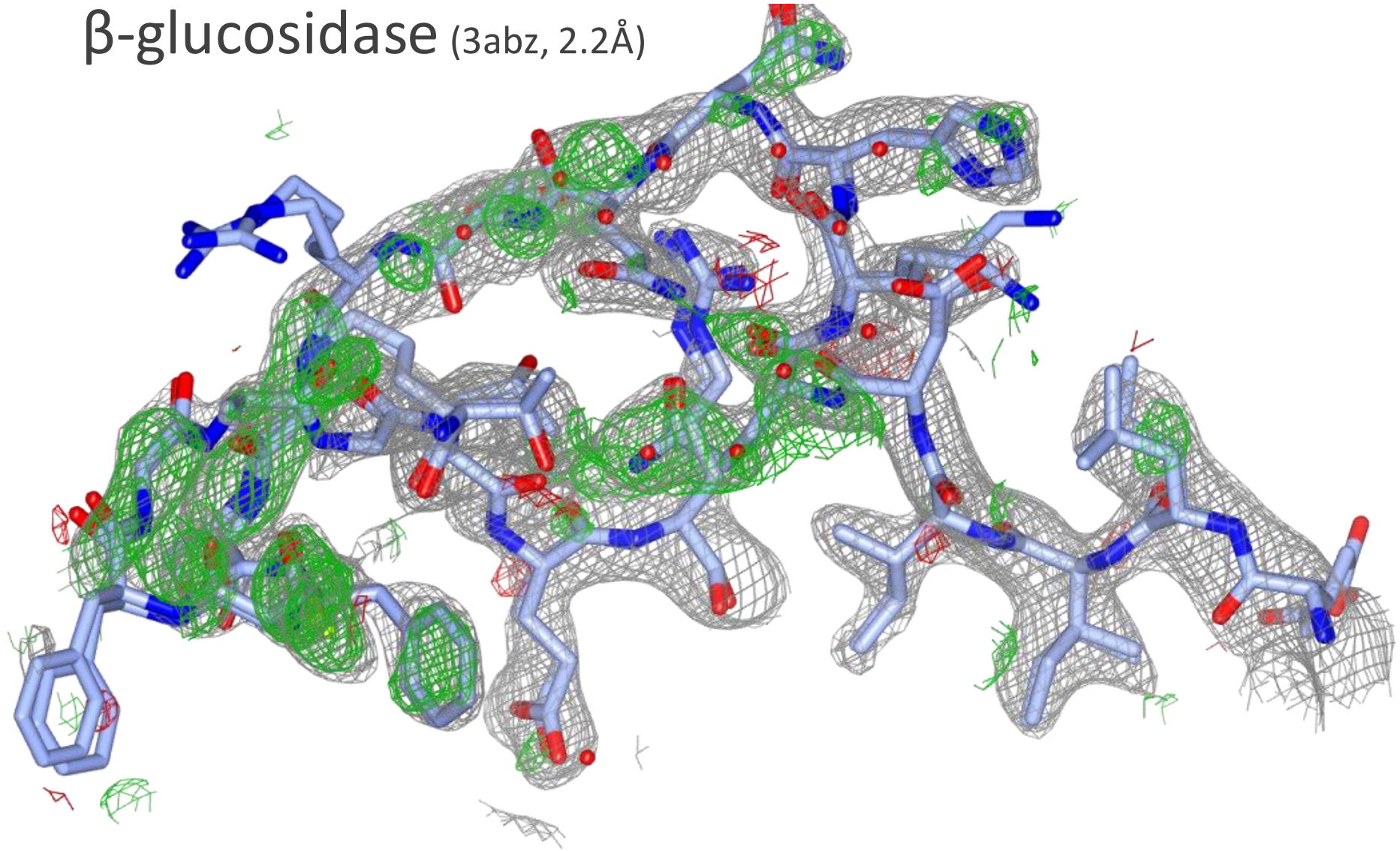
Loopwhole adds missing loops by homology

1. Find missing loop based on sequence
2. Find homologous structures with the loop present
3. Prepare for loop transfer
4. For all homologous loops:
   - Transplant the loop to target model
   - Refine with coot-mini-rsr
5. Keep the best loop if it is good enough
   - Filter on structural quality and map fit



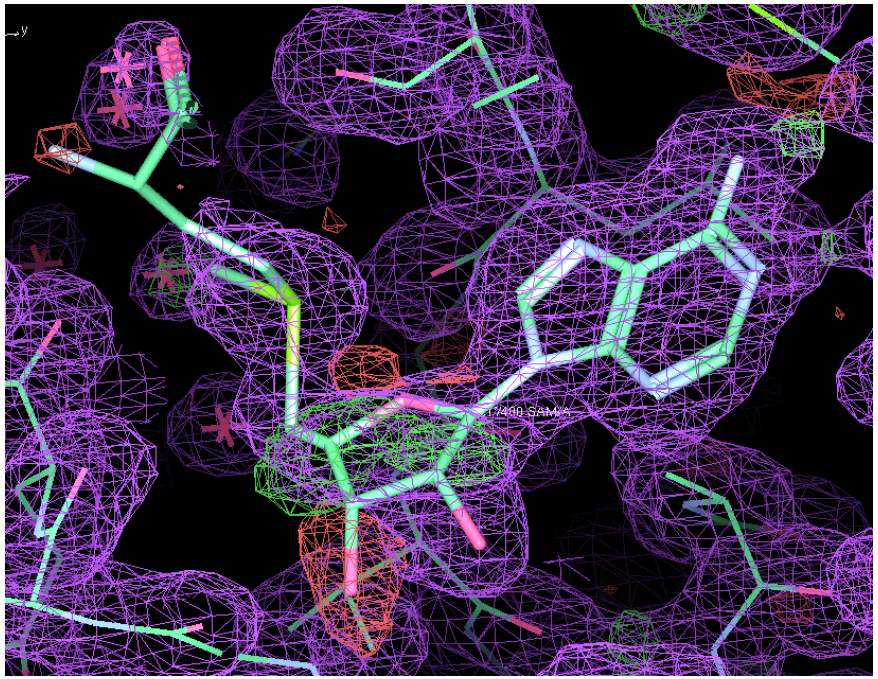
Real-life example

β-glucosidase (3abz, 2.2Å)

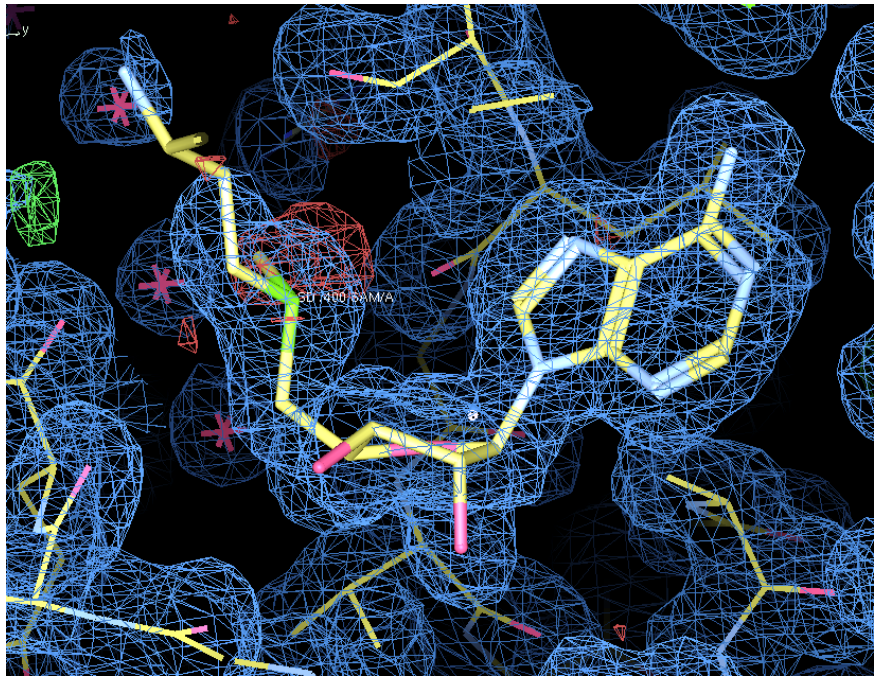


Validation

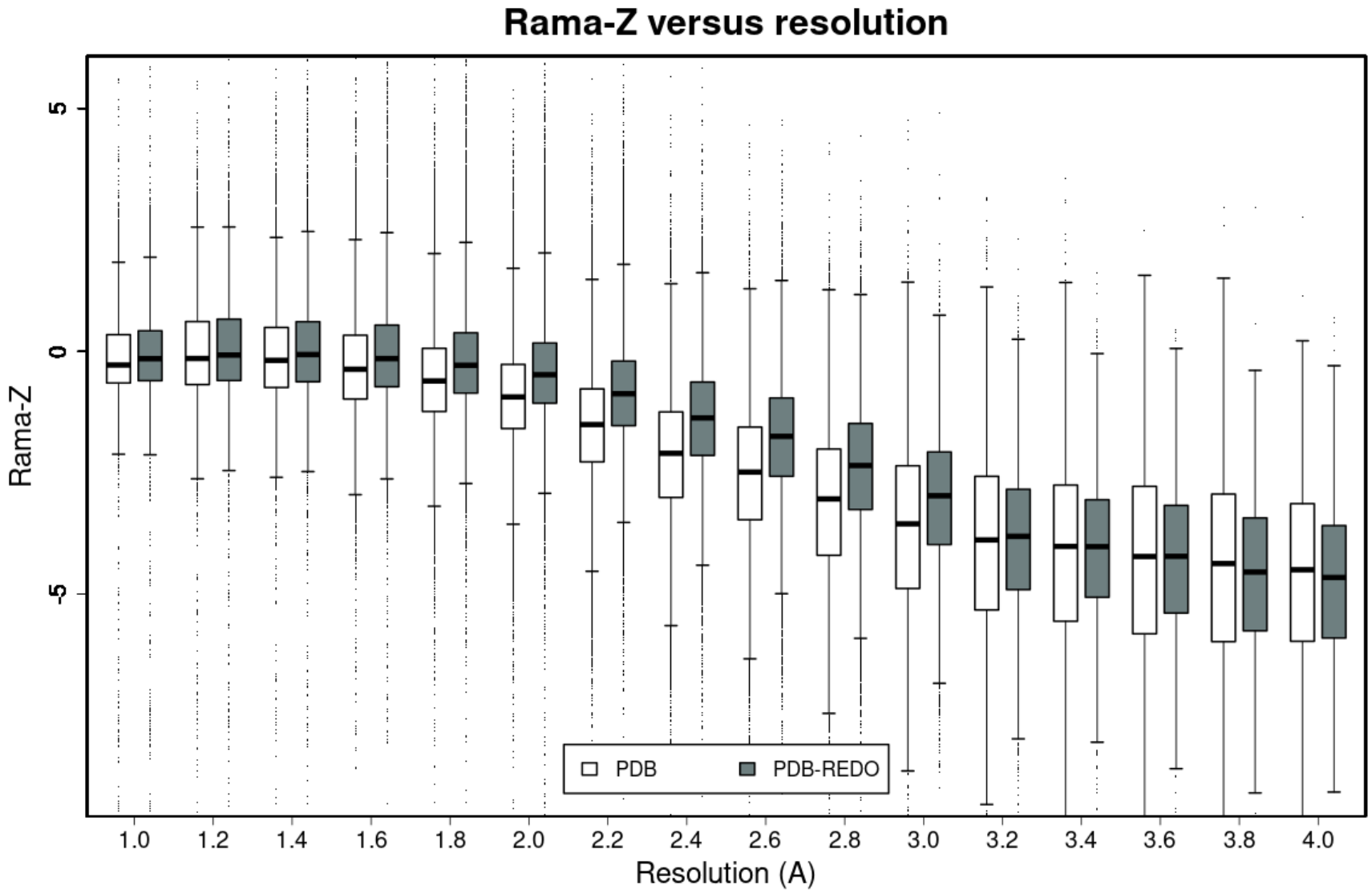
- Before-and-after scores for:
  - Geometry: packing, rotamers, Ramachandran plot, H-bonds, bumps
  - Density fit: per-residue RSR, RSCC, EDIAm and OPIA
  - Ligands: density fit, interactions and Heat-of-Formation



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



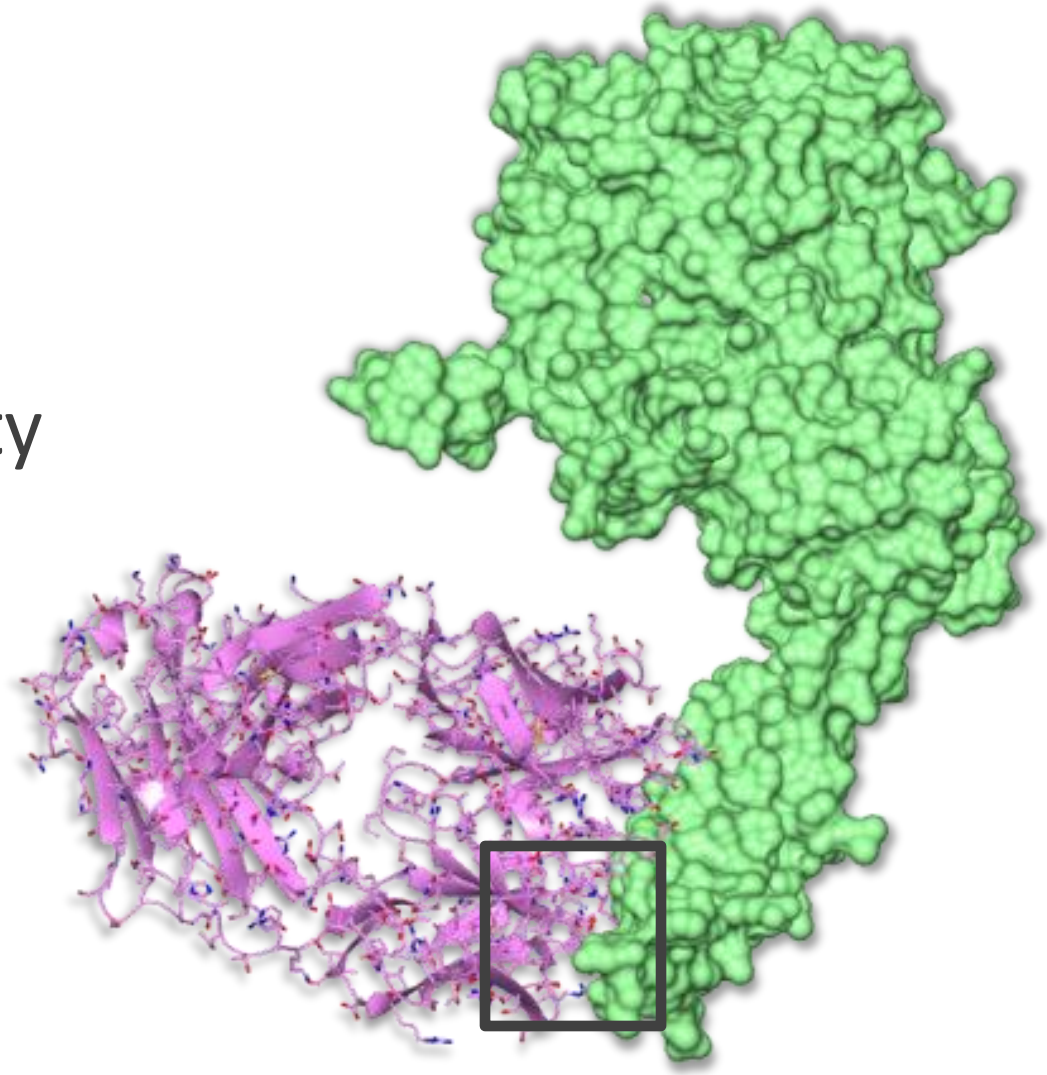
PDB-REDO model quality (n = 166k)



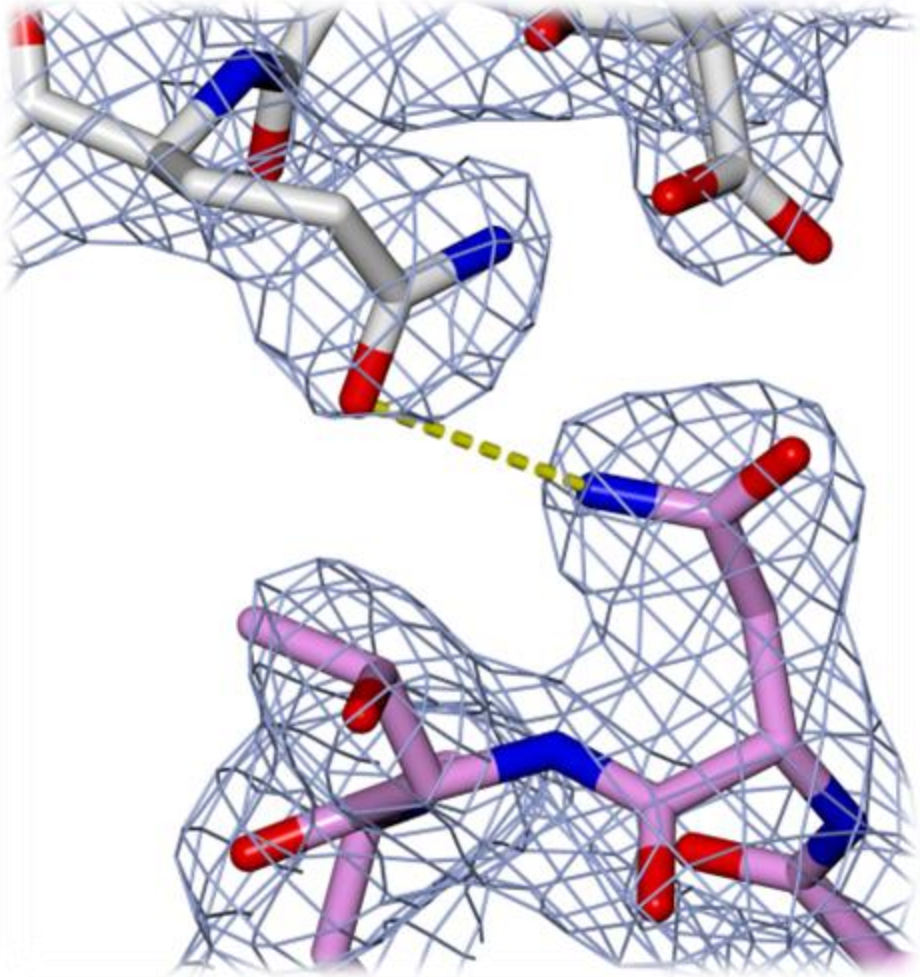
Herceptin – HER2 interface

After PDB-REDO:

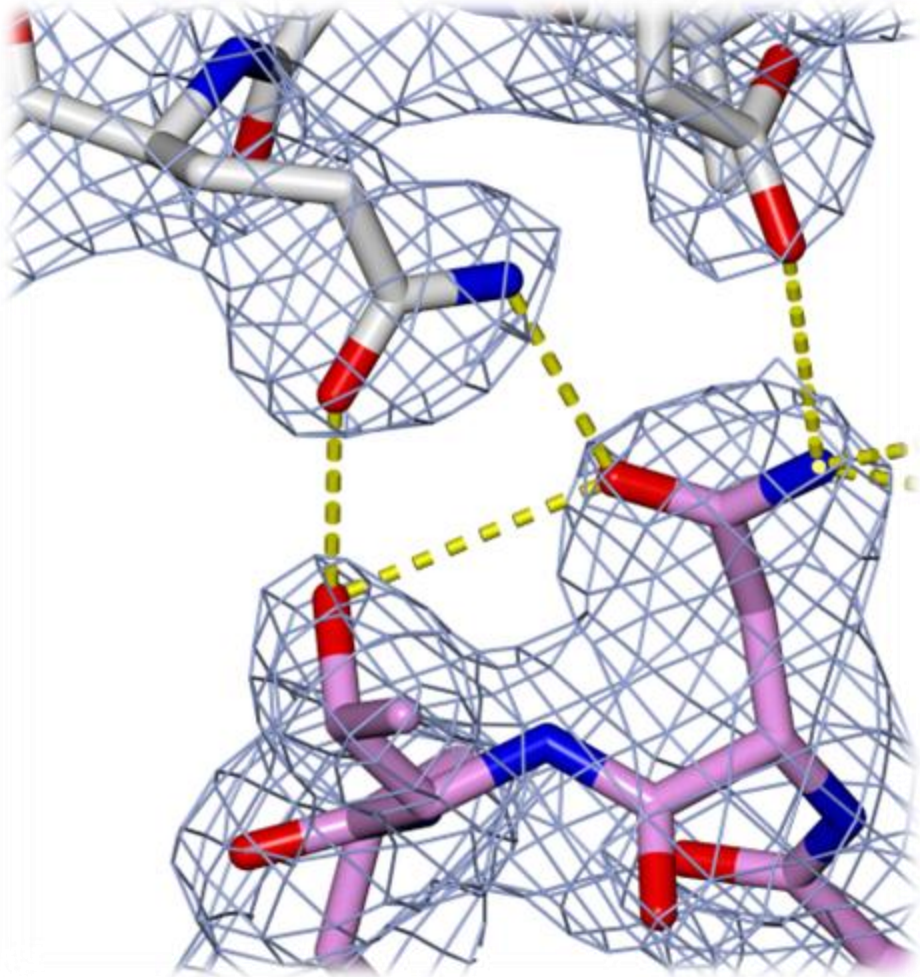
- R-free from 31.6% to 26.7%
  - 7σ improvement
- Moved from 34th to the 99th quality percentile in MolProbity



Herceptin – HER2 interface



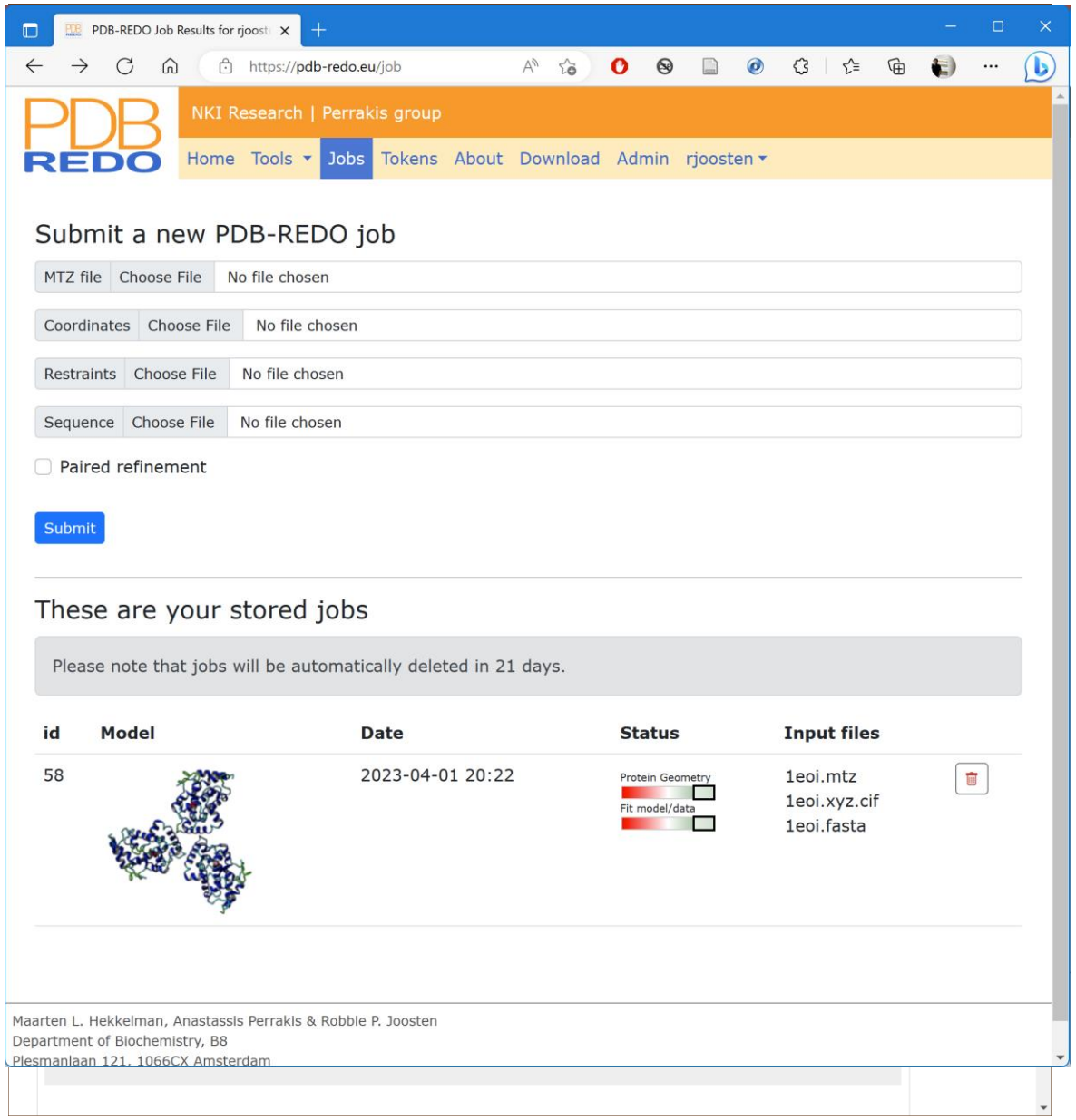
PDB



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
  - >10k jobs submitted in 2023 so far
  - Jobs take ~1 hour



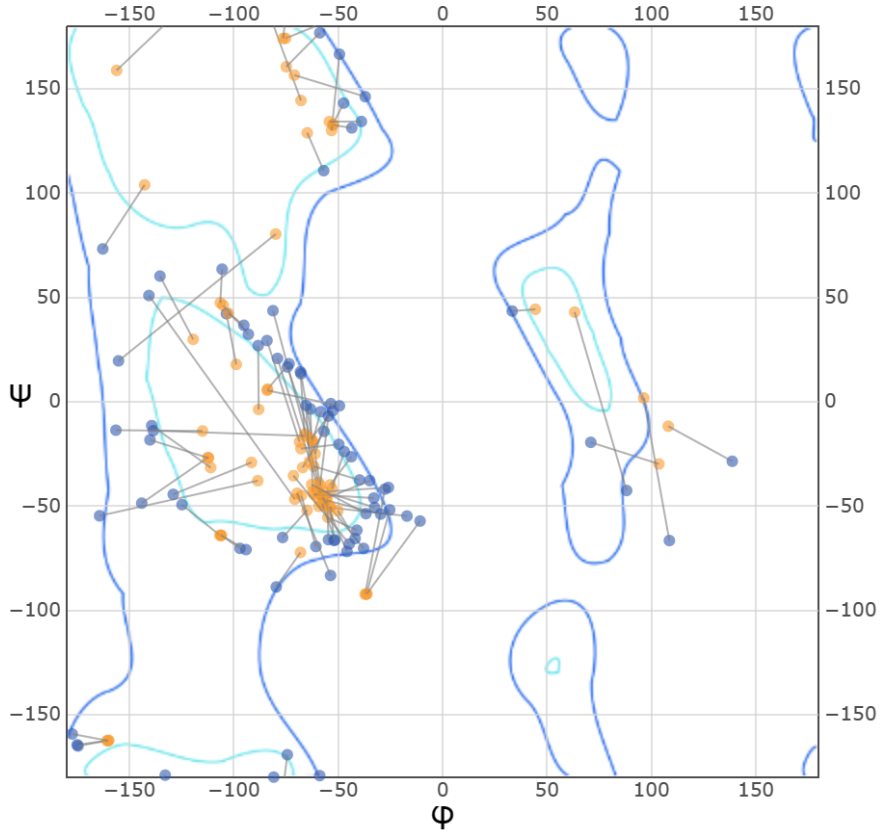
PDB-REDO output

- New model + new map coefficients
- Tools to continue working on the structure model
  - Optimised settings for refinement in REFMACAT
  - Ready-made extra restraints
- Description of model changes
  - At the local and the global level
  - Visually oriented: colour coding, plots, visualisation script for COOT

Validation metrics from PDB-REDO

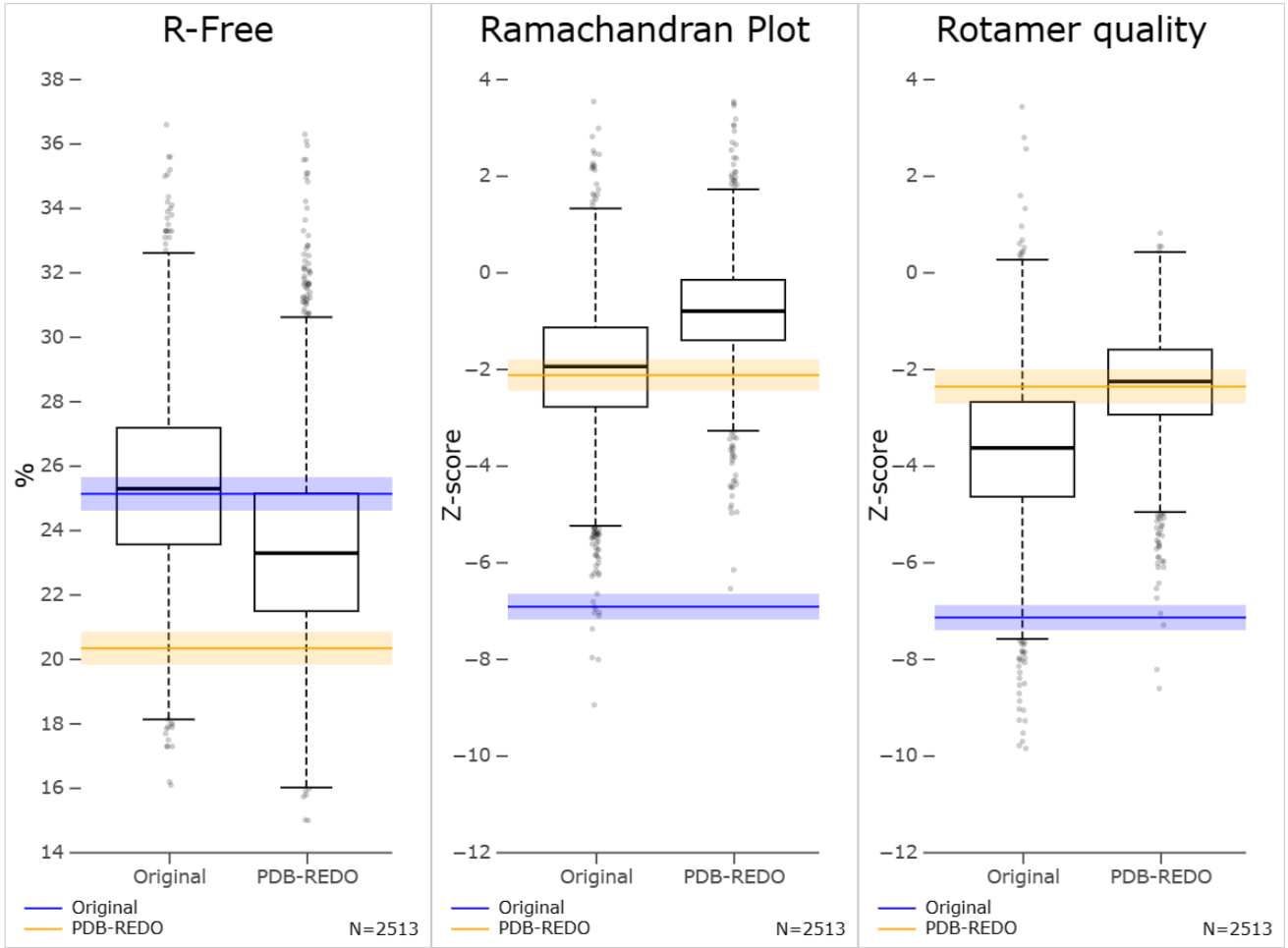
	PDB	PDB-REDO
Crystallographic refinement		
R	0.2094	0.1660
R-free ⚠	0.2512	0.1988
Bond length RMS Z-score	0.526	0.423
Bond angle RMS Z-score	0.748	0.668
Model quality raw scores | percentiles		
Ramachandran plot normality	1	35
Rotamer normality	4	52
Coarse packing	94	99
Fine packing	20	82
Bump severity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

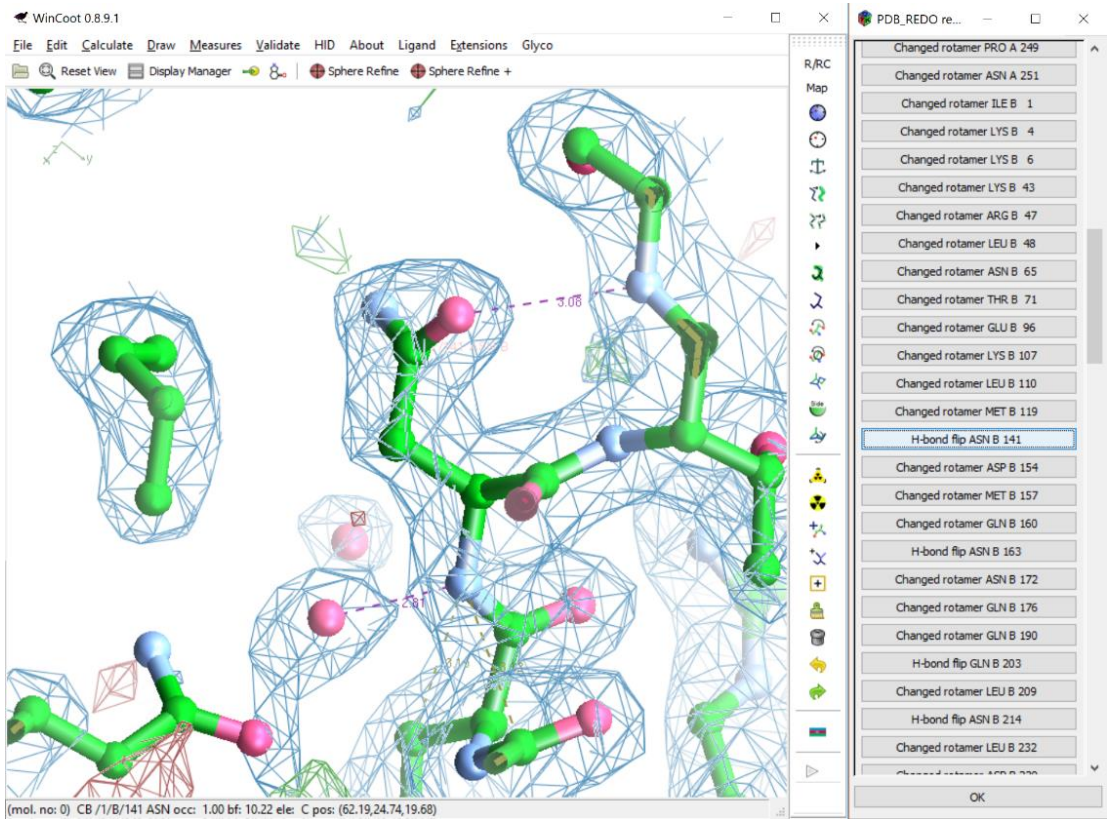
Kleywegt-like plot



Description	PDB ●	PDB-REDO ●
Ramachandran Z-score	-6.909	-1.713
☐ Preferred regions	581	647
☑ Allowed regions	55	10
☑ Outliers	24	3

Model quality compared to resolution neighbours





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

Acknowledgements

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

- Gert Vriend

- Jon Agirre

- CCP4 developers
- PDB annotators

- Garib Murshudov
- Paul Emsley
- Rob Nichols

- Xiang-Lun Ju