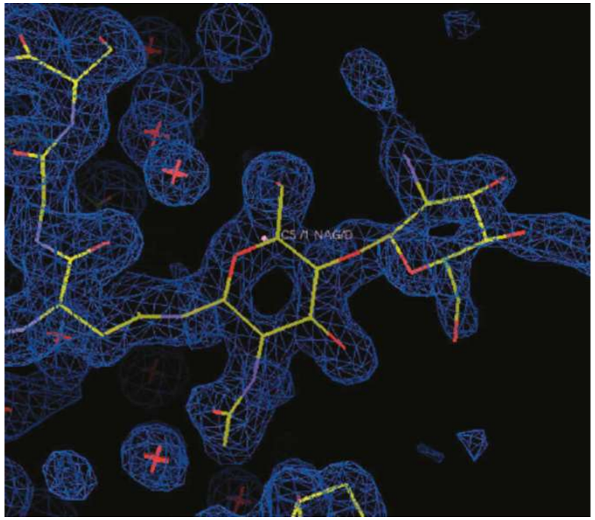
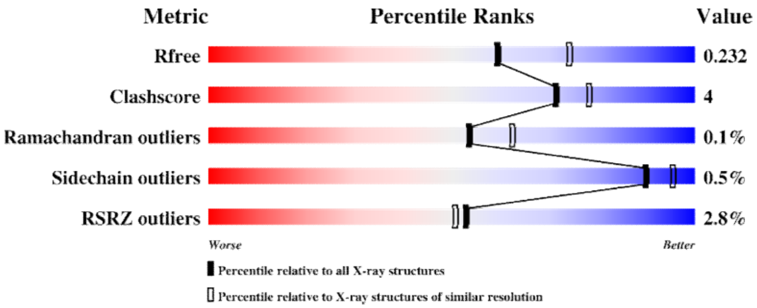
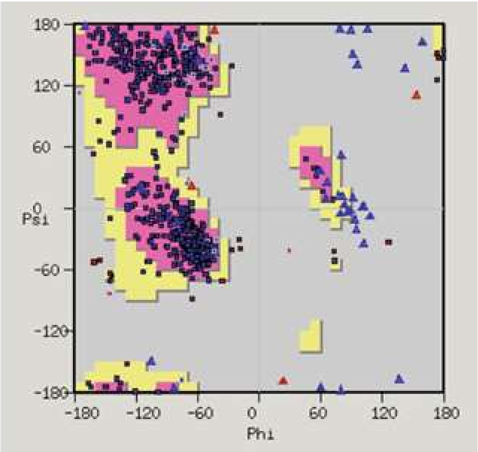
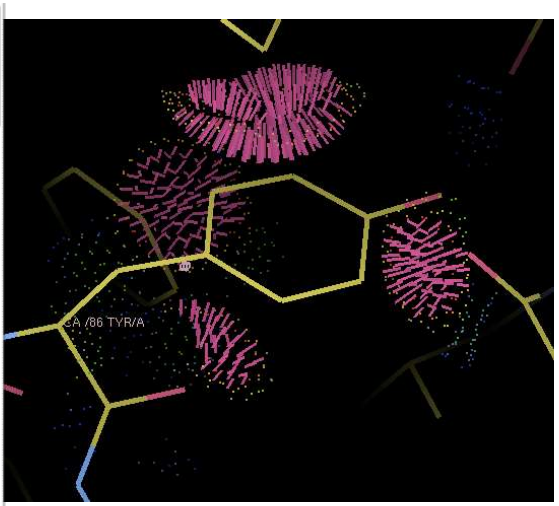
Crystal structure validation







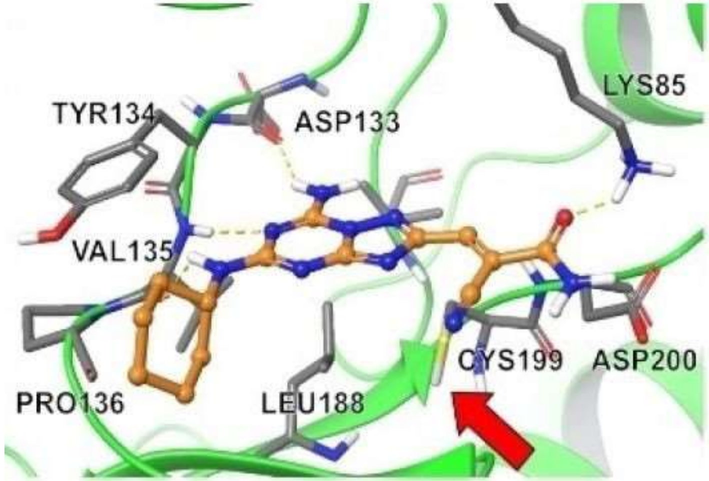
Multi-criterion validation
CC
Ramachandran
Rotamer
C-beta
Bad clash
B-factor
Fc
2mFo-DFc
Density
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Local real-space CC
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
B-factor
40
35
30
25
20
15
10
5
A10
A20
A30
A40
A50
A60
A70
A80
A90
Residue



Corso di Biocristallografia e Microscopia Elettronica

rdezorzi@units.it

Importance of validation



Once deposited on PDB, the structure can be used for multiple different studies:

- Good and correct structures are crucial for docking and structure based drug-design
- Biochemical and mutagenetic analysis are often based on structures

- *Model bias on calculated electron density*
- *Low data/parameter*
- *Overparametrization*
- *Entrapment of refinement in local minima*
- *Misinterpretation of electron density*

Analysis based on posterior probability and Bayesian statistics:

- *Chemical and physical knowledge*
- *Expectation values for geometric parameters*
- *Biochemical studies*

Improbable ≠ impossible

Bayesian statistics

$P(A|B,C,D)$ Conditional probability: probability that event *A* occurs given the previous knowledge *B*, *C*, *D*...

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}$$

Bayes' theorem (inference theorem): the conditional probability of a hypothesis *H* given the evidence *E* is the probability of obtaining the evidence E given the model, multiplied by the probability of the model (hypothesis, H) and divided by the probability of the evidence (E).

Crystallographic question:

What is the probability that the ***protein model*** *is correct, given*

(1) the crystallographic data and

(2) the chemical/physical knowledge of the system?

Bayesian approach in crystallography

In crystallography:
M model, D crystallographic data,
I previous chemical/physical information

$$P(M|D,I) = \frac{P(D|M,I)P(M|I)}{P(D|I)}$$

$P(D|M,I)$ likelihood of the data, consistency of the data with the proposed model
$P(M|I)$ consistency of the model with the previous knowledge
$P(D|I)$ probability to obtain the dataset, considering previous knowledge (it can be trated as normalization factor and neglected)

The best model is the one for which probability if maximized.

However, rather than maximizing the probability, in the maximum likelihood approach a negative logarithm is minimized:

$$L(M|D,I) = -\log[P(D|M,I)] - \log[P(M|I)]$$

The first term has a close relation with the χ^2 parameter:

$$\chi^2 = \sum_h W_h\,[F_{obs} - F_{calc}(M)]^2$$

W_h can be calculated from the previous step considering the variance evaluated (Bayesian approach).

Global and local analyses

Global criteria for structure evaluation:

- R-values and their difference:

$$R_{free} = \frac{\sum_{\boldsymbol{h}\in free}|F_{obs} - kF_{calc}|}{\sum_{\boldsymbol{h}\in free} F_{obs}}$$

$$R_{work} = \frac{\sum_{\boldsymbol{h}\notin free}|F_{obs} - kF_{calc}|}{\sum_{\boldsymbol{h}\notin free} F_{obs}}$$

Indication on overfitting or incomplete refinement

- Overall r.m.s.d. of bond distances and angles

Depending on restraints used in refinement

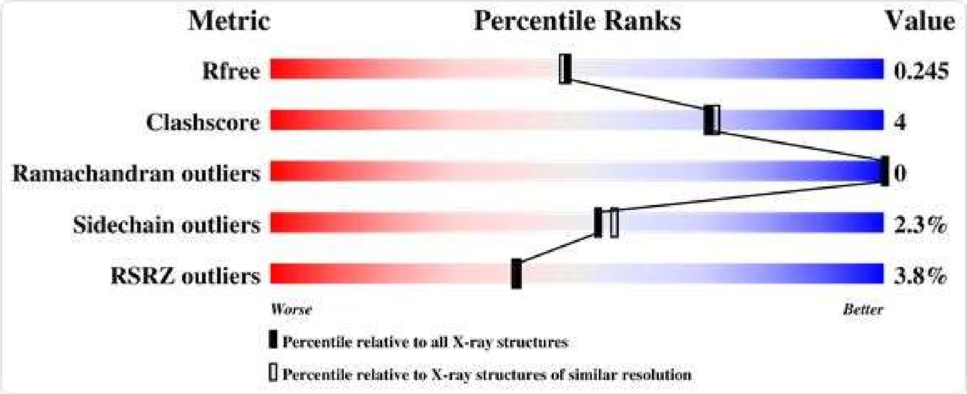
- Average B-factor

Local criteria for structure evaluation:

Residue-by-residue analysis of:

GEOMETRY ANALYSIS

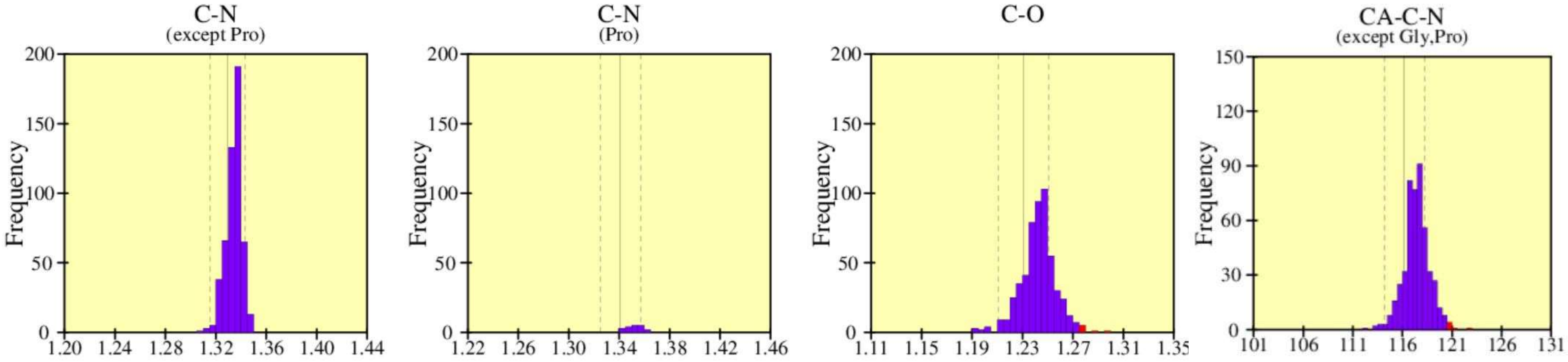
- Geometric parameters: deviation from expected values



- Real space R-value
- Fitting of the model in the electron density: Real Space Cross-correlation Coefficient
- B-factor values for each residue

Geometry analysis: bonds and angles

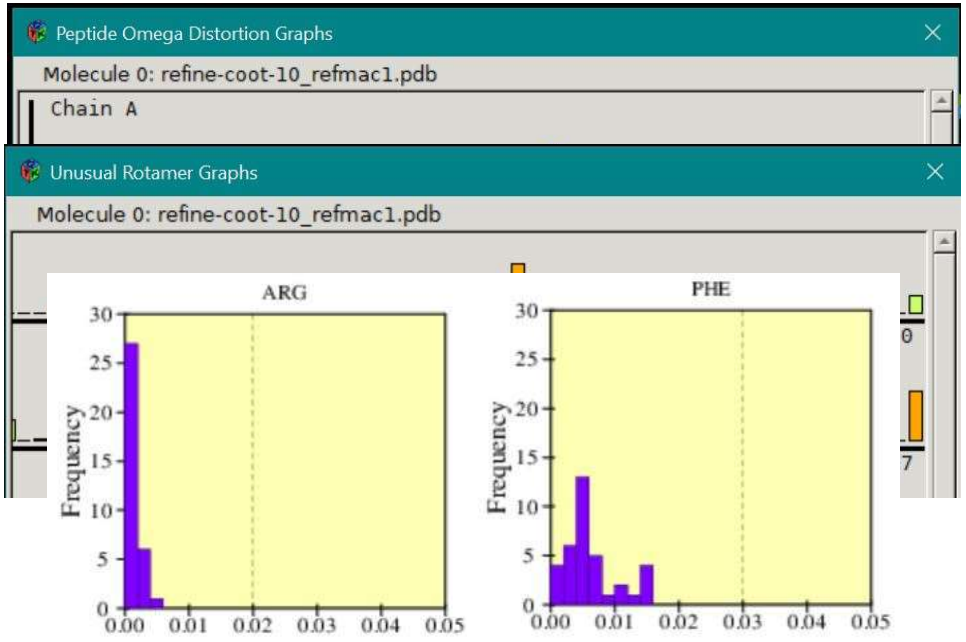
Bond lenghts and angles:



Values measured on the model compared with expectation values and their standard deviations.

In addition:

- ω angle (peptide bond dihedral angle)
- Side chain conformations
- Planarity of side chains
- ...


Molecule 0: refine-coot-10_refmac1.pdb

Chain A

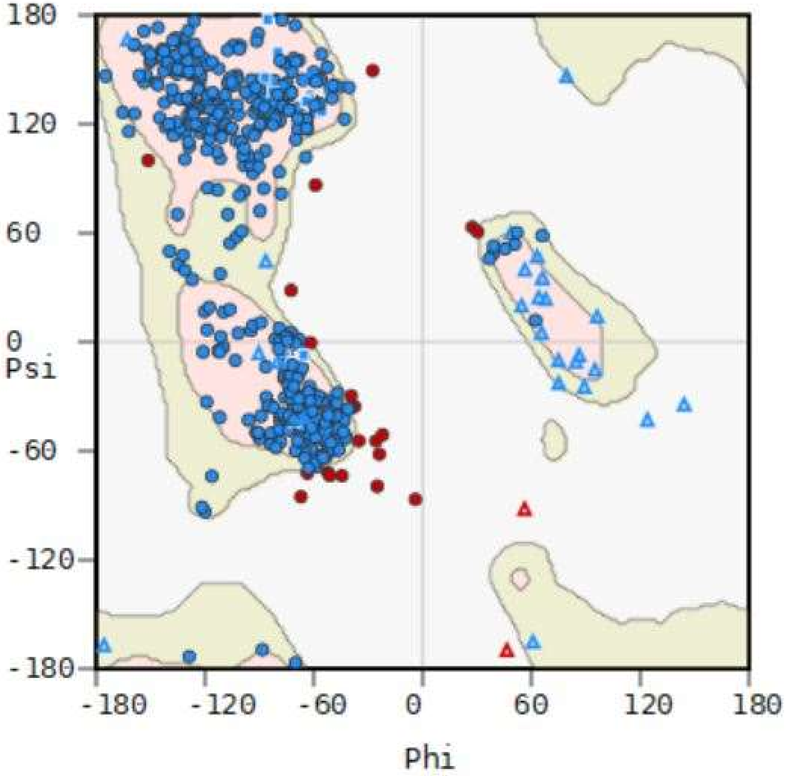
Molecule 0: refine-coot-10_refmac1.pdb

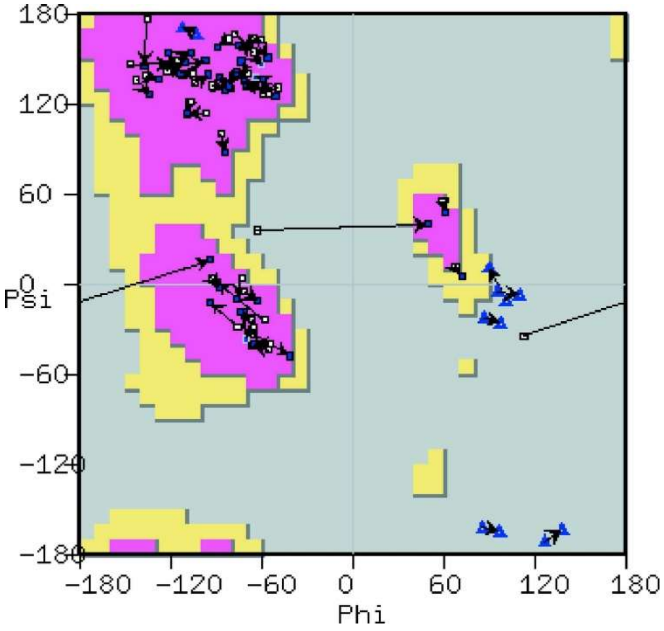
Geometry analysis: Ramachandran plots

Usually, angles ϕ and ψ are not restrained during refinement → good for cross-validation!

Ramachandran violations are possible, but should be carefully analyzed:

1. *Evaluate electron density of the residue (often Ramachandran violations in disordered loops poorly modeled)*
2. *Analyze similar structures: strained conformations may have important biochemical roles*





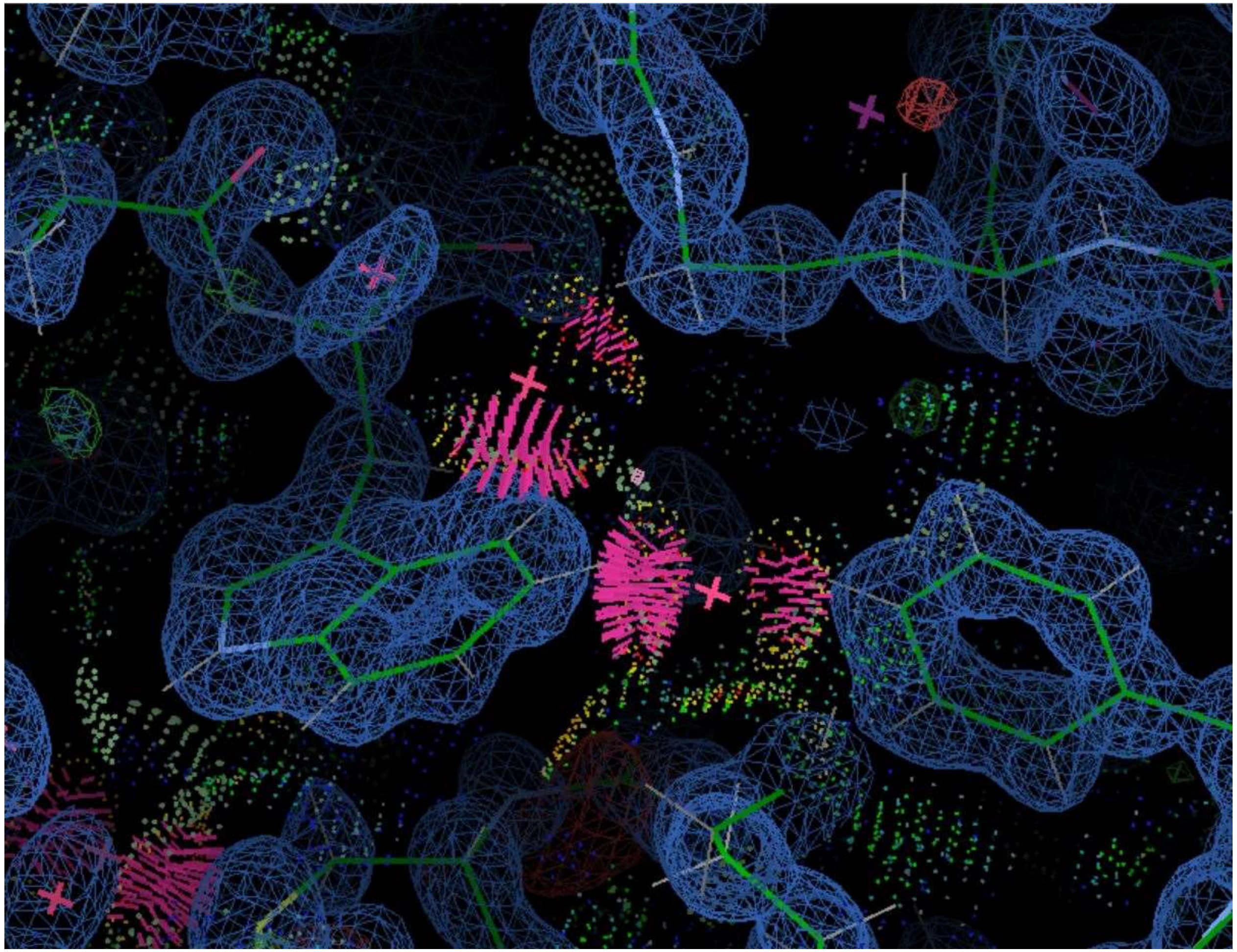
Kleywegt plots: analysis of difference in backbone dihedral angles for residues of NCS related molecules.

Geometry analysis: clashes

Non-bonding contacts shorter than van der Waals radii of atoms.

Software can recognize hydrogen bonding networks and identify hydrophobic close contacts.

Particularly dangerous contacts with symmetry related molecules.



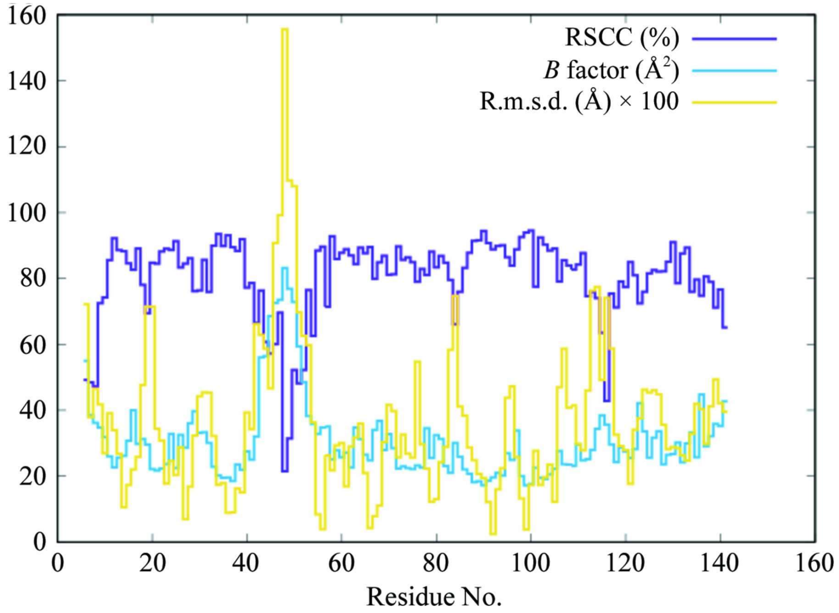
Electron density validation

Real Space R-value (RSR):

$$RSR = \frac{\sum_{\boldsymbol{r}} |\rho_{obs}(\boldsymbol{r}) - \rho_{calc}(\boldsymbol{r})|}{\sum_{\boldsymbol{r}} |\rho_{obs}(\boldsymbol{r}) + \rho_{calc}(\boldsymbol{r})|}$$

with $\rho_{obs} = \sum_{\boldsymbol{r}} (2F_{obs} - F_{calc}) \cdot \exp(-i\varphi_{calc})$ and $\rho_{calc} = \sum_{\boldsymbol{r}} F_{calc} \cdot \exp(-i\varphi_{calc})$ for each voxel of the real space (***r***)

Values are plotted residue by residue, often together with the B-factor value.



Real Space Cross-correlation Coefficient (RSCC):

$$RSCC = \frac{\sum_{\boldsymbol{r}} \left[\rho_{obs}(\boldsymbol{r}) - \overline{\rho_{obs}(\boldsymbol{r})}\right] \cdot \left[\rho_{calc}(\boldsymbol{r}) - \overline{\rho_{calc}(\boldsymbol{r})}\right]}{\left(\sum_{\boldsymbol{r}} \left[\rho_{obs}(\boldsymbol{r}) - \overline{\rho_{obs}(\boldsymbol{r})}\right]^2 \cdot \sum_{\boldsymbol{r}} \left[\rho_{calc}(\boldsymbol{r}) - \overline{\rho_{calc}(\boldsymbol{r})}\right]^2\right)^{1/2}}$$

Visual analysis of maps: bias-minimized maps, omit maps (using calculated phases and amplitudes from model where a specific portion was removed).

Ligands

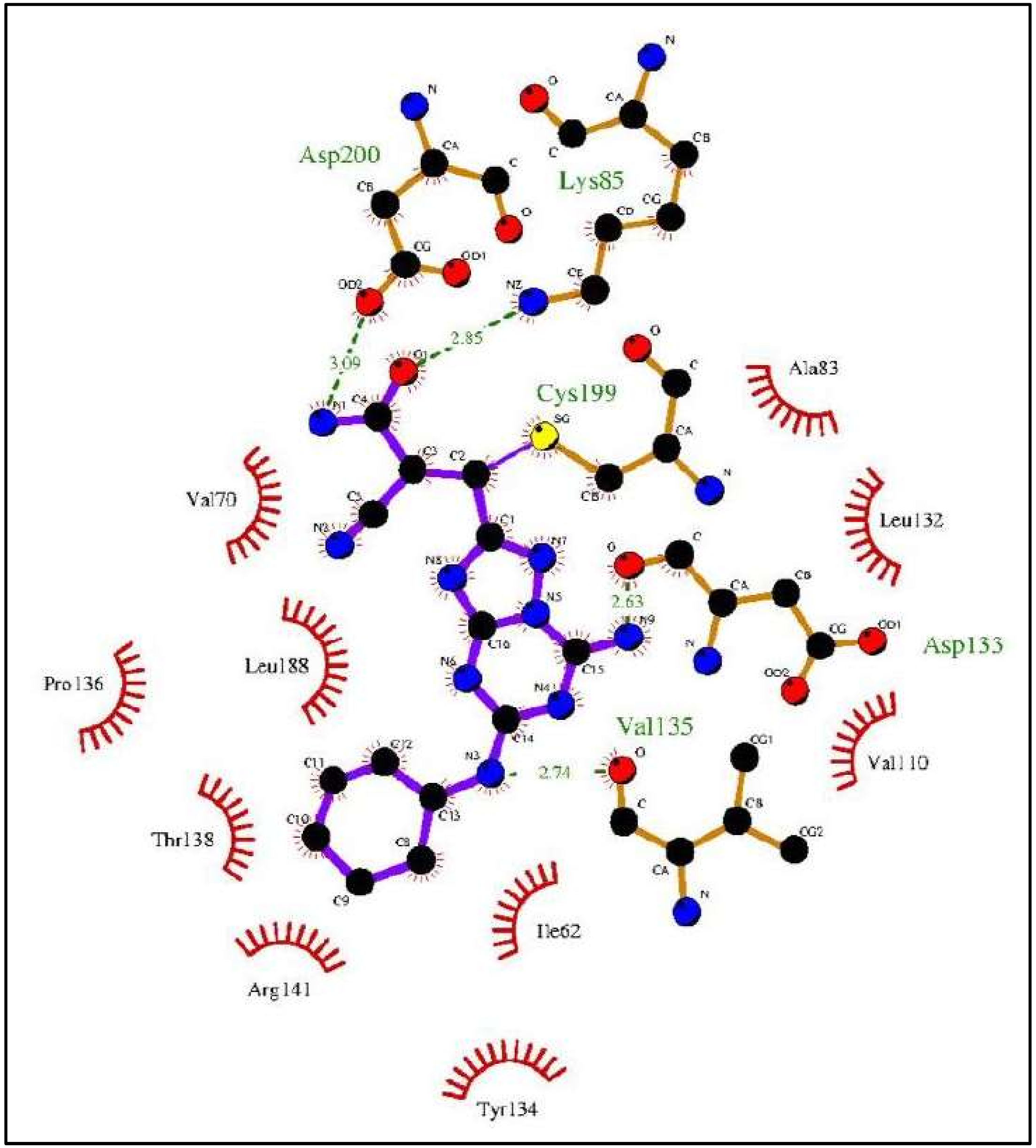
For known ligands:
Geometric restraint file are available for the main ligands

For new ligands:
Geometric restraints should be generated considering chemical knowledge (e.g. ibridization state, bond leght and angles...)

Omit maps and difference maps are extremely useful tools to confirm the presence of ligands.

Often, ligands occupy the binding site only in a percentage of proteins. **Partial occupancy** should be considered together with **B-factor values**.

Careful analysis of ligand contacts (**LIGPLOT**):


Asp200
Lys85
N
CA
C
O
CB
CG
OD1
OD2
CD
CE
NZ
3.09
2.85
O1
C4
N1
C3
C2
SG
Cys199
Ala83
Val70
C5
N2
C1
N7
N8
N5
N9
2.63
Leu132
Asp133
Leu188
Pro136
C16
C15
N6
N4
C14
N3
2.74
Val135
Val110
C12
C11
C10
C13
C8
C9
CG1
CG2
Thr138
Ile62
Arg141
Tyr134