Crystal structure validation
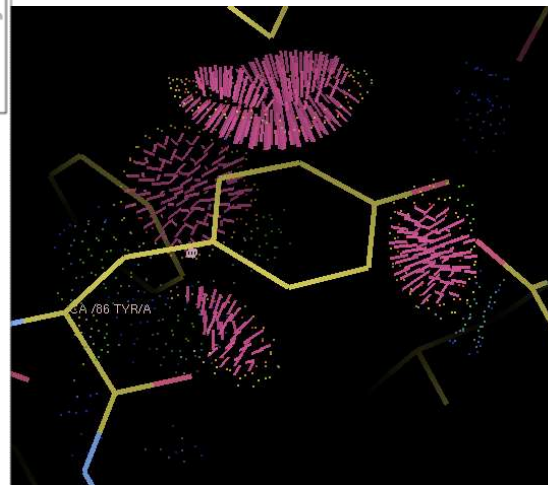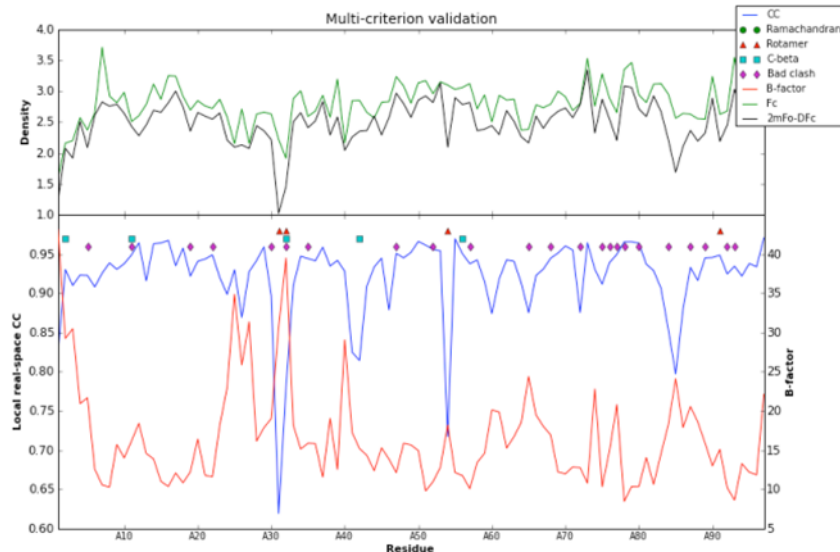
Corso di
Biocristallografia
e Microscopia
Elettronica
rdezorzi@units.it

Importance of validation
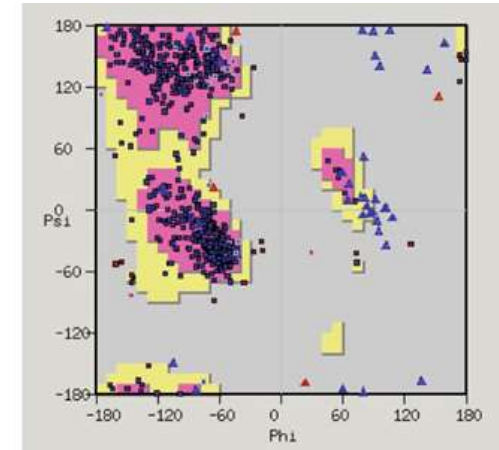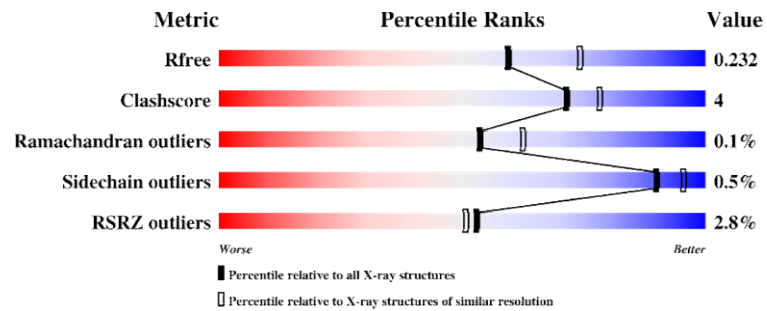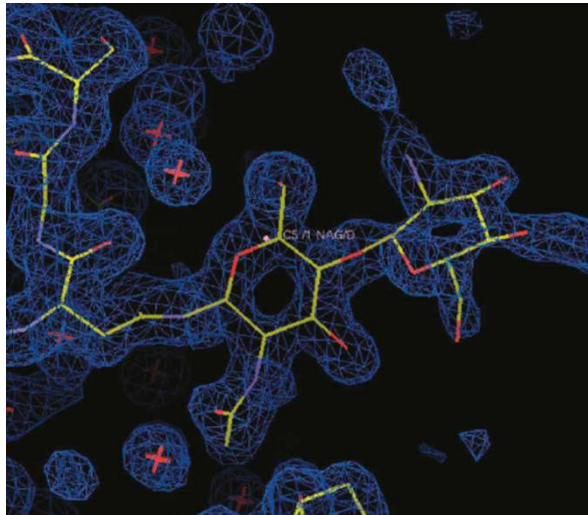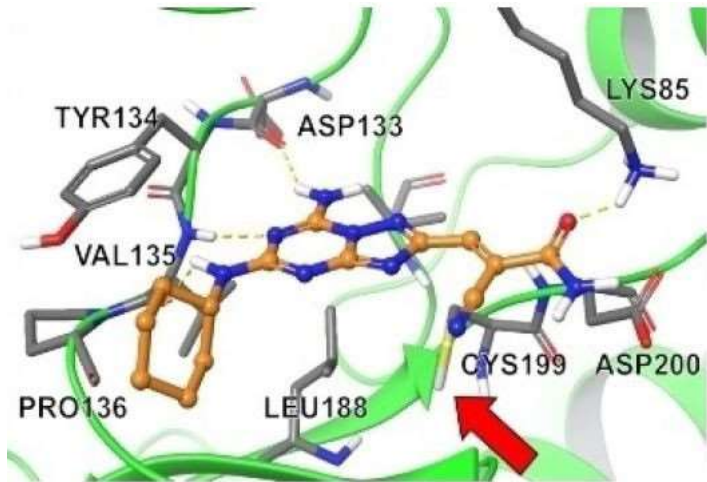

Once deposited on PDB, the structure can be used for multiple different studies:

- 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...*

*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?*

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

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).

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

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

1) Are the experimental data consisten with the proposed protein model?

2) Is the model consistent with the knowledge we have of the system?

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

Metric	Percentile Ranks	Value
Rfree		0.245
Clashscore		4
Ramachandran outliers		0
Sidechain outliers		2.3%
RSRZ outliers		3.8%

Worse   *Better*

▮ Percentile relative to all X-ray structures
▯ Percentile relative to X-ray structures of similar resolution

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

Depending on restraints used in
refinement

- Average B-factor

Global and local analyses

Global criteria for structure evaluation:

- R-values and their difference:

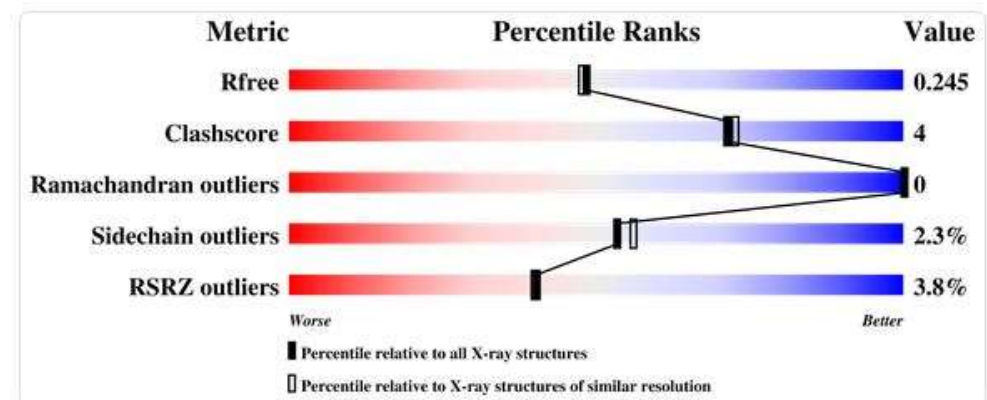$$R_{free} = \frac{\sum_{\boldsymbol{h}\in free}|F_{obs} - kF_{calc}|}{\sum_{\boldsymbol{h}\in free} F_{obs}}$$

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

> 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

- Ramachandran plot

- Non-bonding contacts closer than the van der Waals radii of the atoms (clashes)

ELECTRON DENSITY ANALYSIS

- Real space R-value

- Fitting of the model in the electron density: Real Space Cross-correlation Coefficient

- B-factor values for each residue

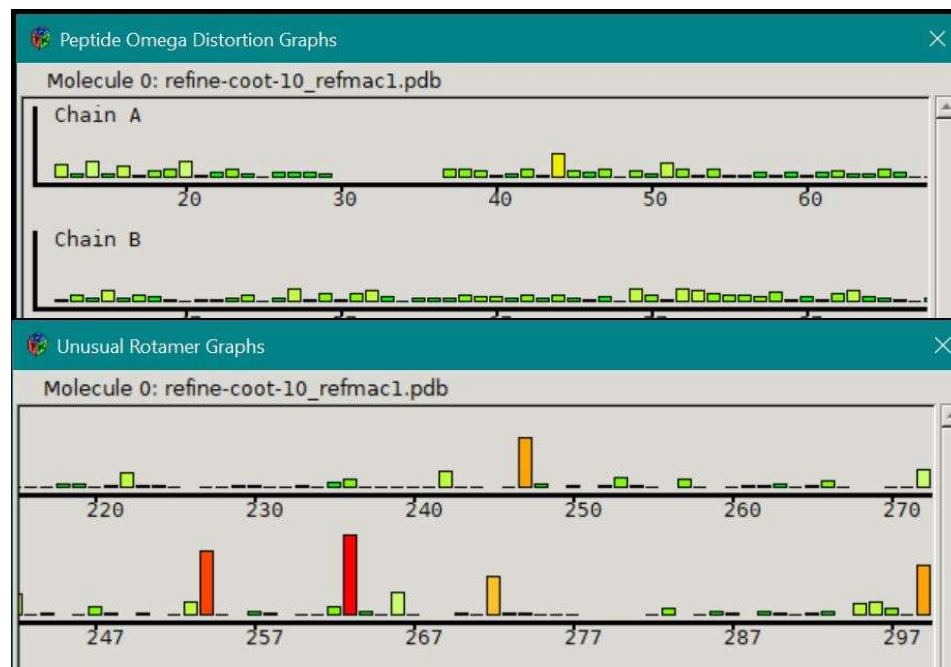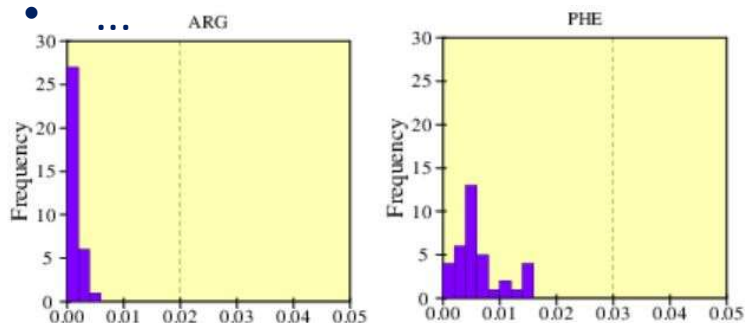# Geometry analysis: bonds and angles

Bond lenghts and angles:
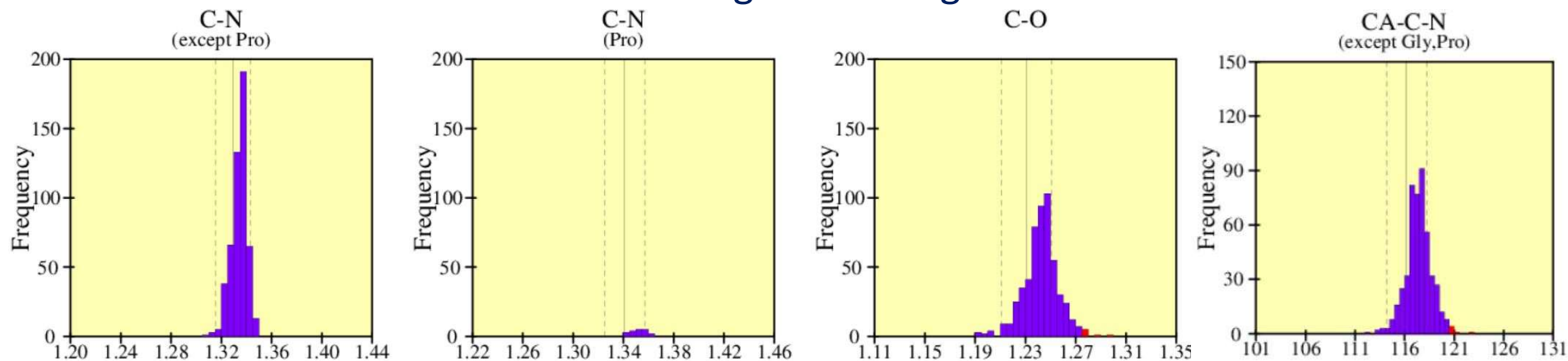
Values measured on the model compared with expectation values and their standard deviations.

In addition:

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

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*
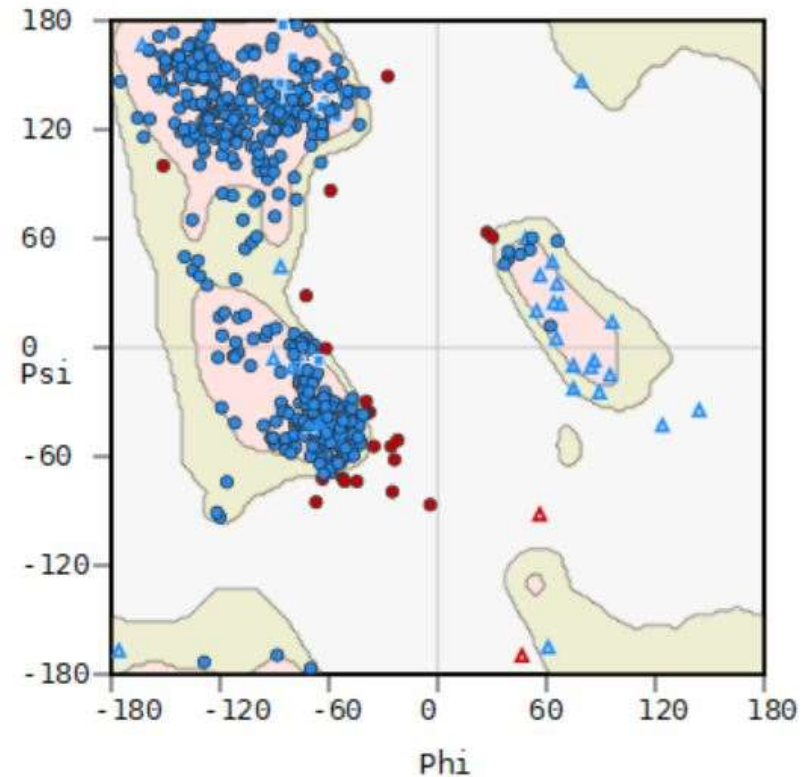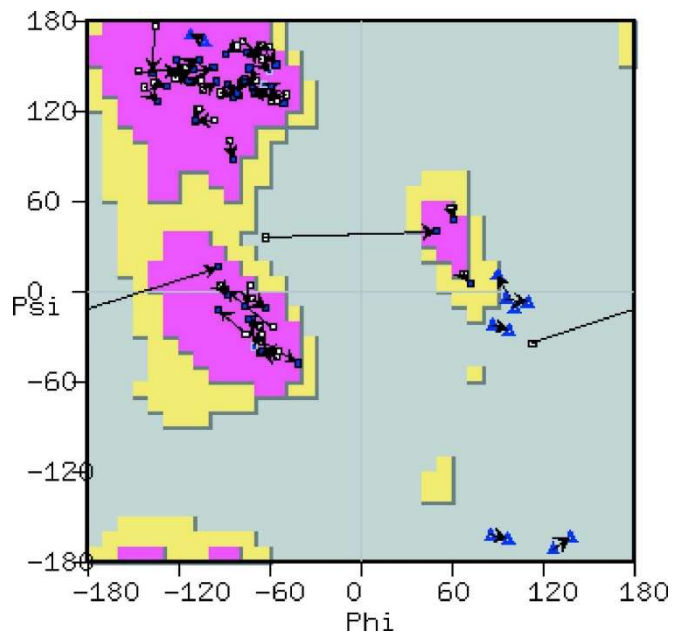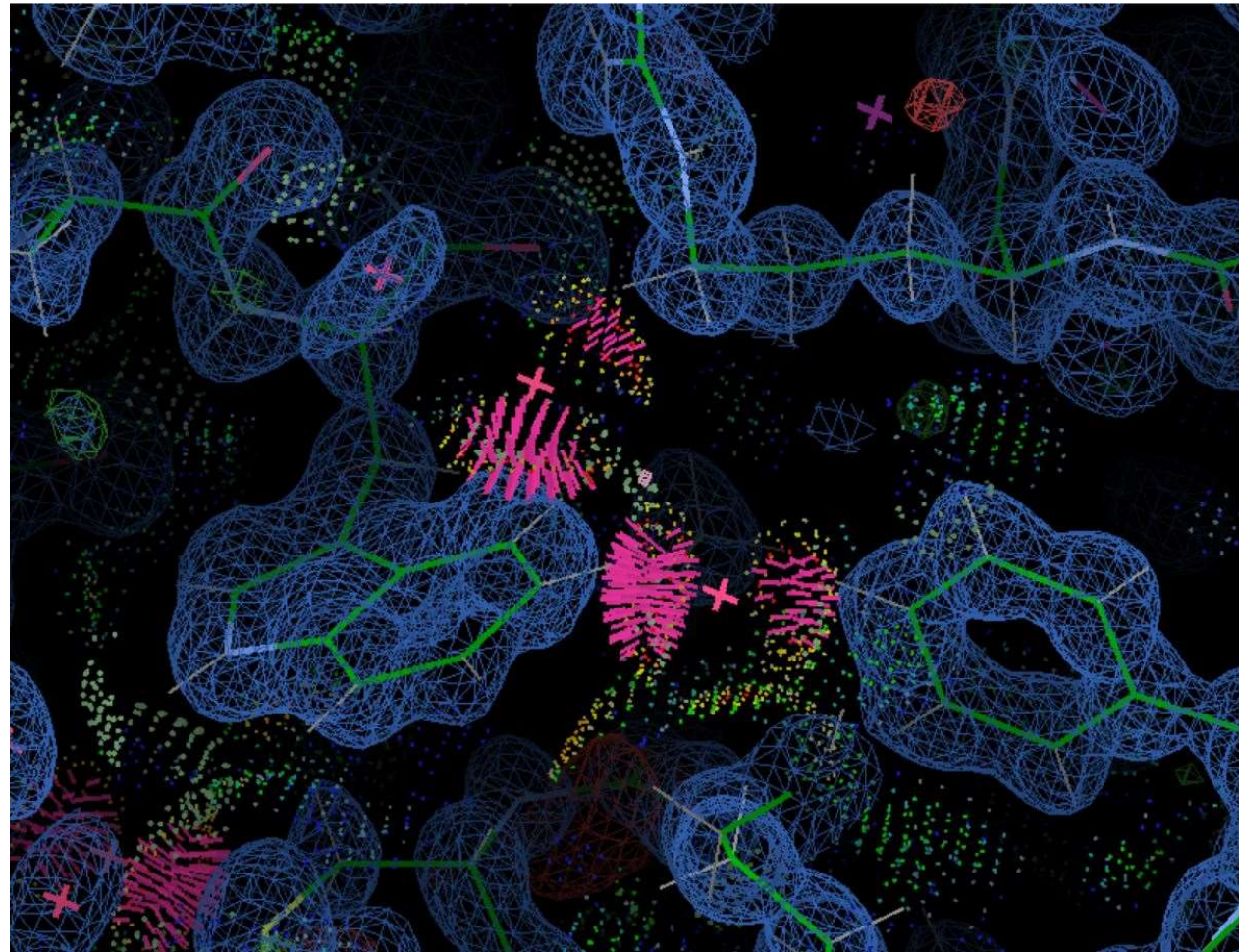

Kleywegt plots: analysis of difference in backbone dihedral angles for residues of NCS related molecules.

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.

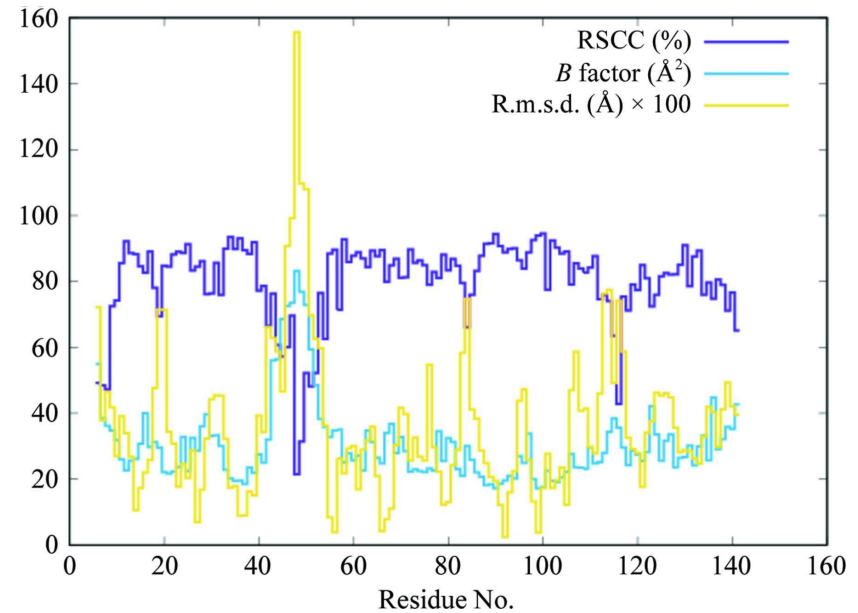Particularly dangerous contacts with symmetry related molecules.

Electron density validation

Real Space R-value (RSR):

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

with $\rho_{obs} = \sum_r (2F_{obs} - F_{calc}) \cdot \exp(-i\varphi_{calc})$
and $\rho_{calc} = \sum_r F_{calc} \cdot \exp(-i\varphi_{calc})$ 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):

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

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**):