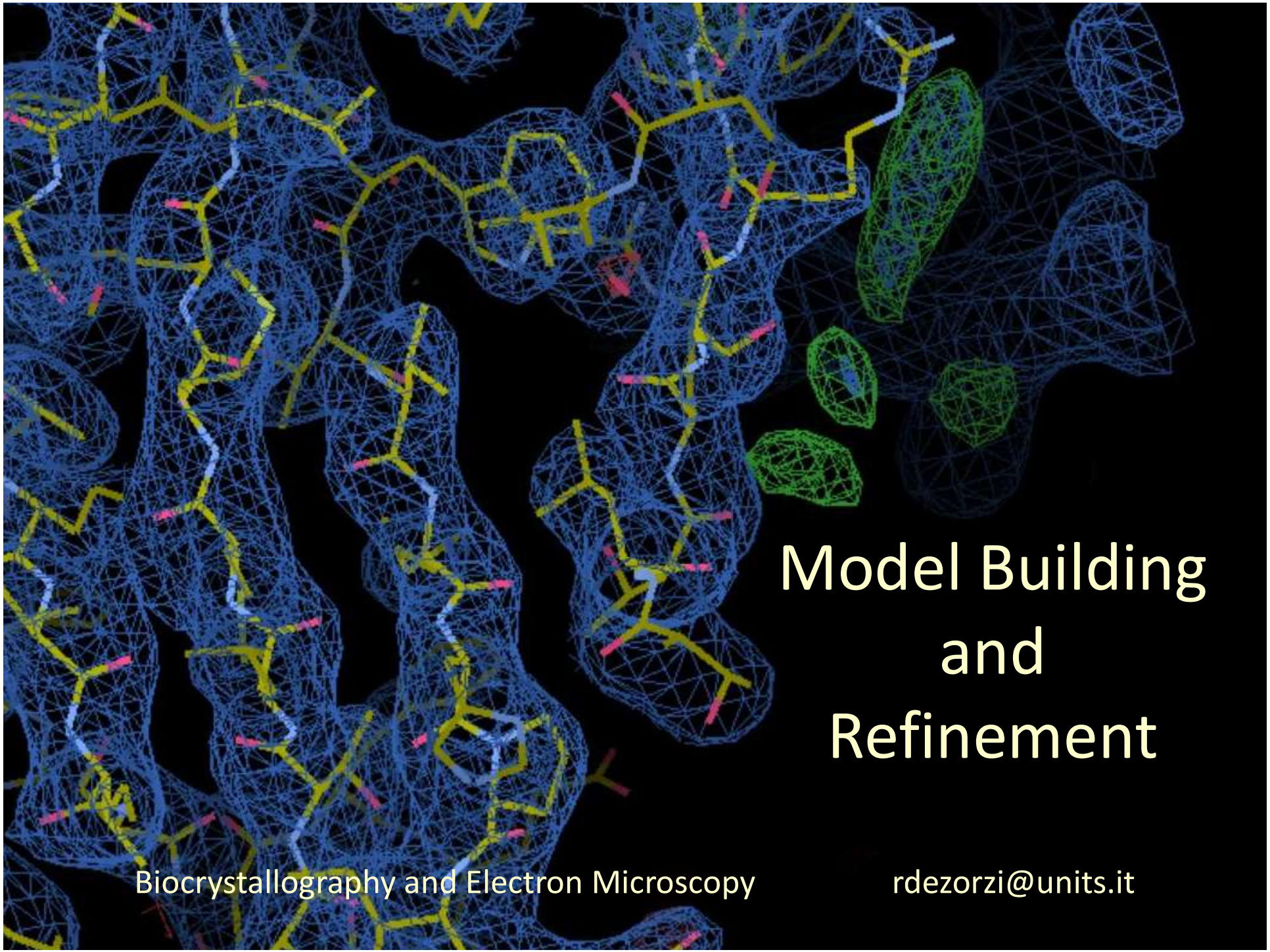
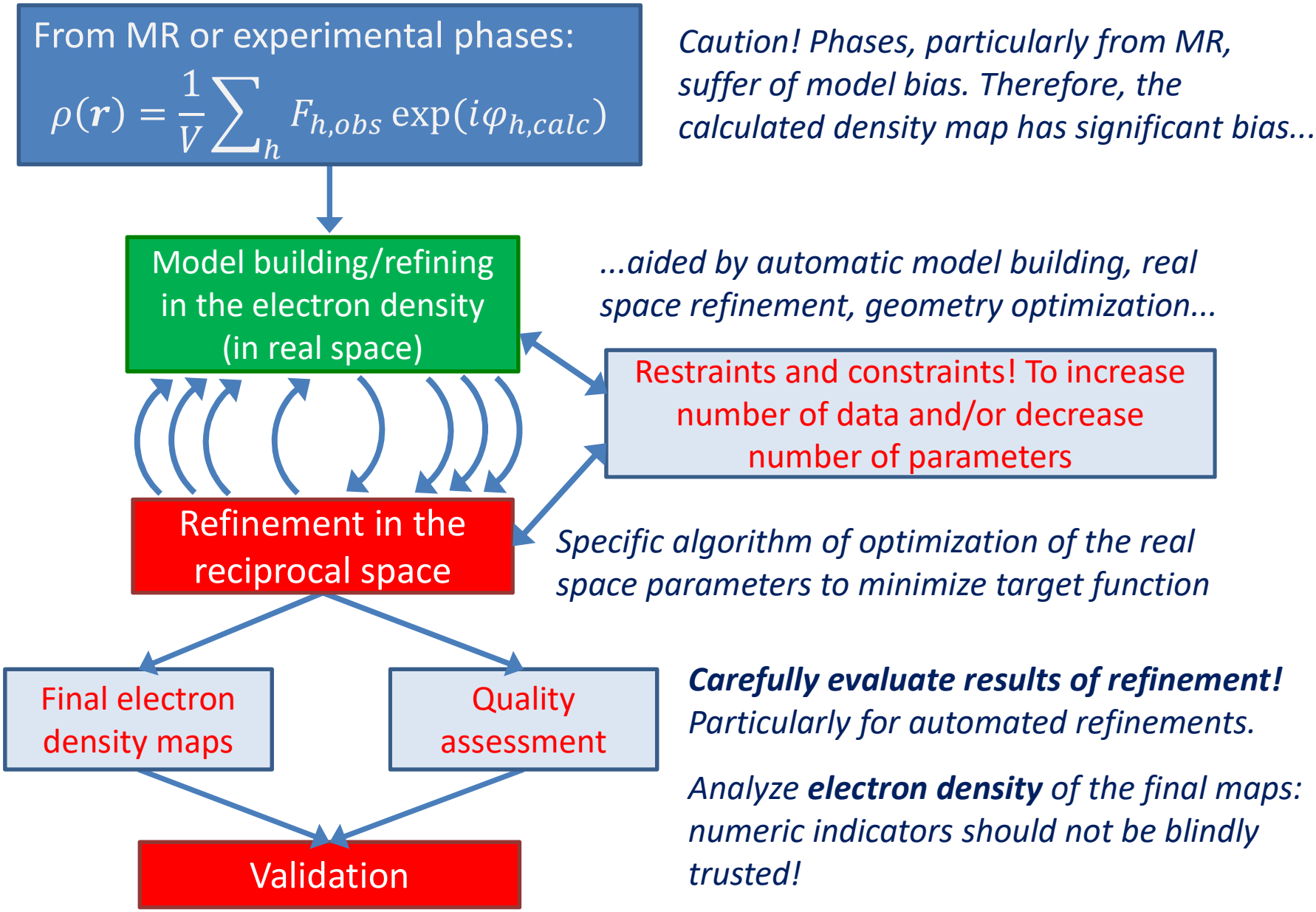


Model Building and Refinement

Biocrystallography and Electron Microscopy

rdezorzi@units.it

From phasing to refinement



Caution! Phases, particularly from MR, suffer of model bias. Therefore, the calculated density map has significant bias...

...aided by automatic model building, real space refinement, geometry optimization...

Specific algorithm of optimization of the real space parameters to minimize target function

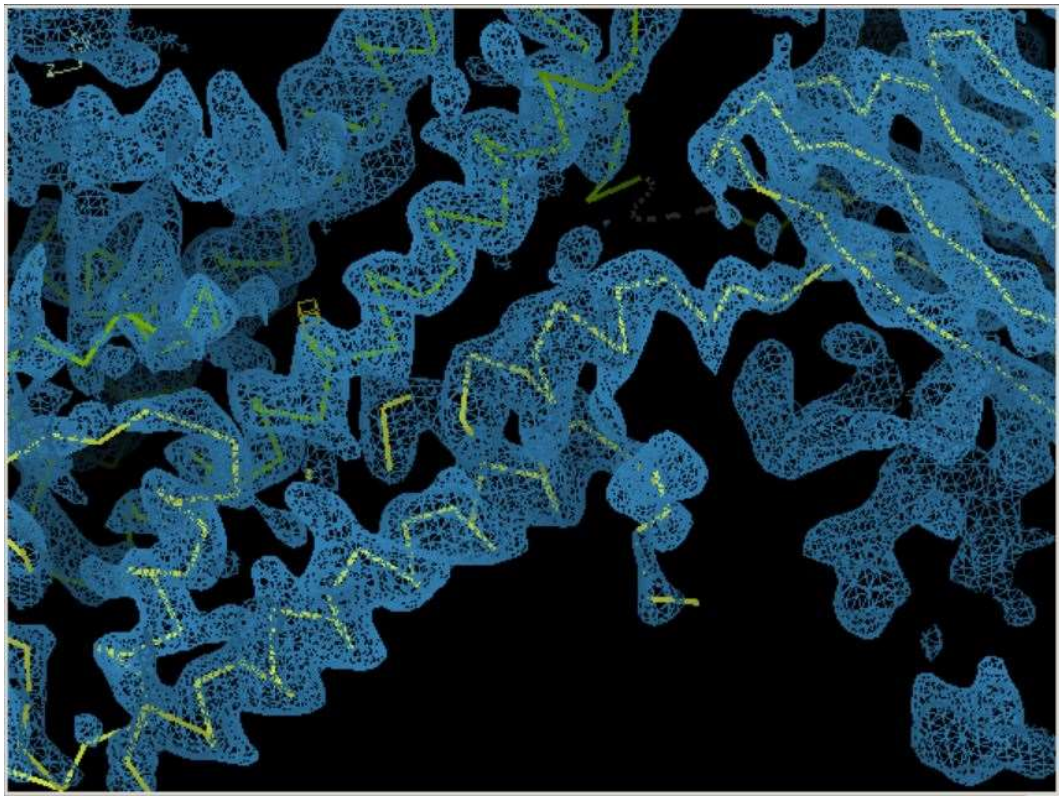
Carefully evaluate results of refinement!
Particularly for automated refinements.

*Analyze **electron density** of the final maps: numeric indicators should not be blindly trusted!*

Model building

MR phasing: model already available, except trimmed parts (loops, side chains of residues)

Experimental phasing: empty density, except for the heavy atom/anomalous scatterer



1. Trace Cα main chain
2. Identify side chains of large residues (easier) and from those identify other residues according to sequence
3. Introduce in the model cofactors and solvent molecules (when well defined)
4. Introduce ligands (<u>but it should be verified!!</u>)

Each step should be followed by Fourier space refinement.

Automatic software for model building are available.

From bad phases and/or bad data, no good model can be built!!

Refinement in reciprocal space

Optimization of N_P **parameters** against N_D **observations** to minimize (or maximize) a **target function.**

Observations: measured diffraction intensities (in reciprocal space)
+ knowledge regarding protein structure (in real space)

Parameters: atomic positions (x,y,z), atomic B-factors (isotropic or anisotropic), occupancy, scale factor, overall B-factor, bulk solvent correction, anisotropy correction... Continuously variable over the defined parameter space.

For each atom: at least 3 positional parameters and 1 B-factor (6 if anisotropy is taken into account).

Different parametrizations are possible and sometimes convenient.

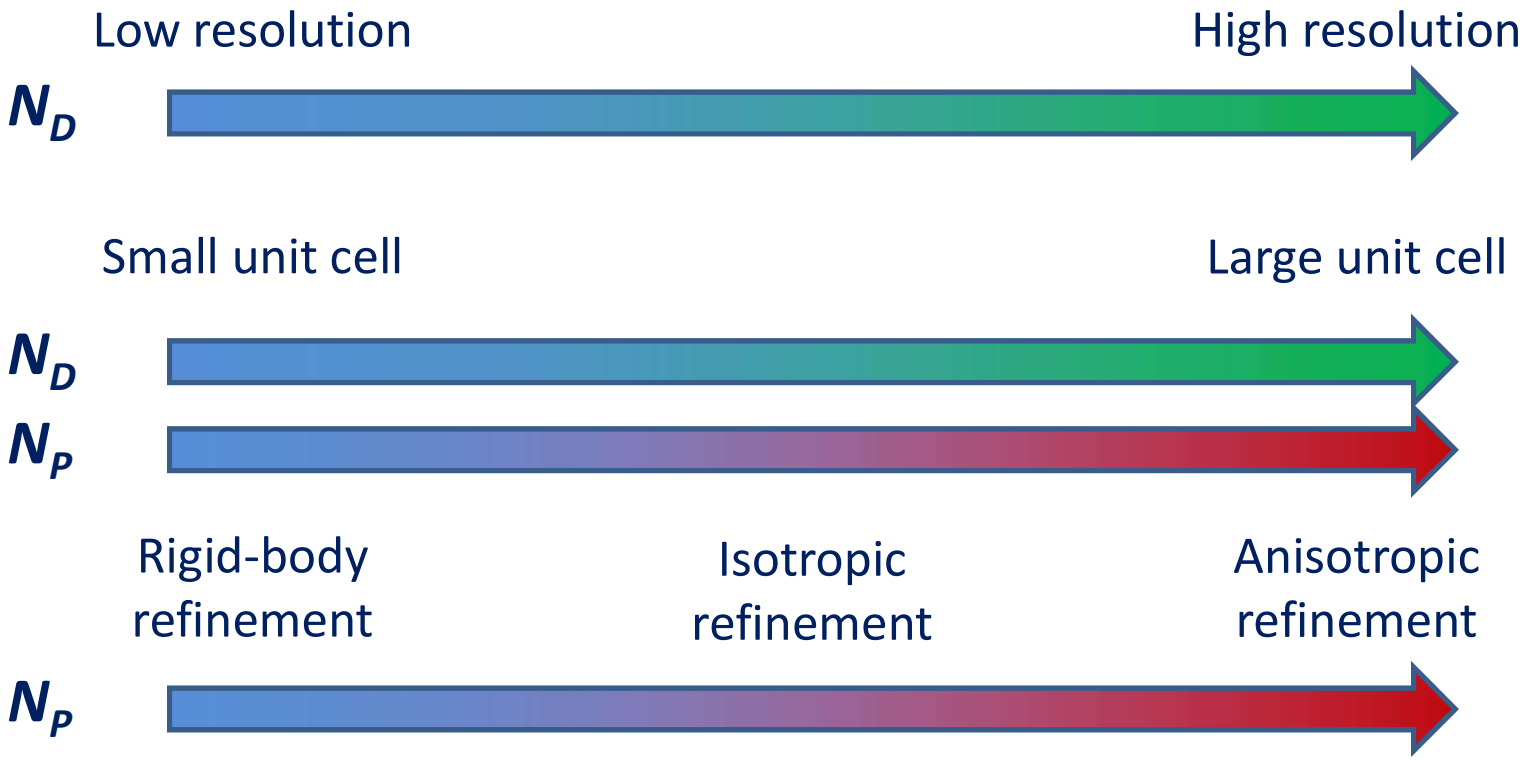
Target function: Changes in parameters don't have a simple effect on data fitting, as errors on a single atom affect many structure factors and their relation is not linear...

Problem: trapping in local minima.

To monitor refinement, **linear residual value *R***

$$R = \frac{\sum_{\boldsymbol{h}} |F_{obs} - F_{calc}|}{\sum_{\boldsymbol{h}} F_{obs}}$$

Data and parameters

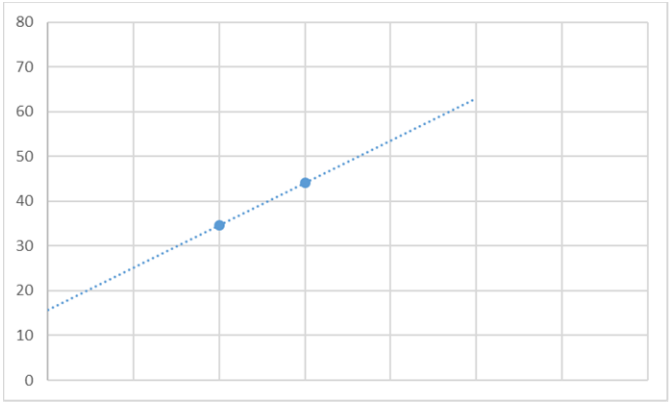


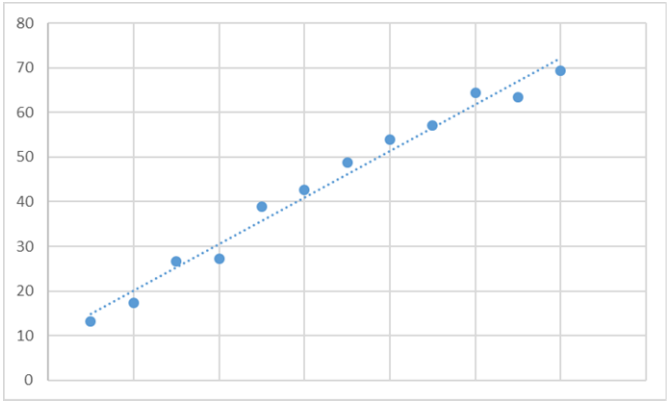
N_P: For a 129 residue protein, each residue containing an average of 8 non-hydrogen* atoms (+ solvent, ions, ligands...), each atom with 4 refinement parameters (in the isotropic case... otherwise 9!) = 4128 parameters

*Hydrogen atoms are usually not refined with positional and atomic displacement parameters: their scattering contribution is not sufficient to justify the refinement.
(However, hydrogen atoms are usually included in refinement in calculated postitions, because their contribution is important, particularly at high resolution.)

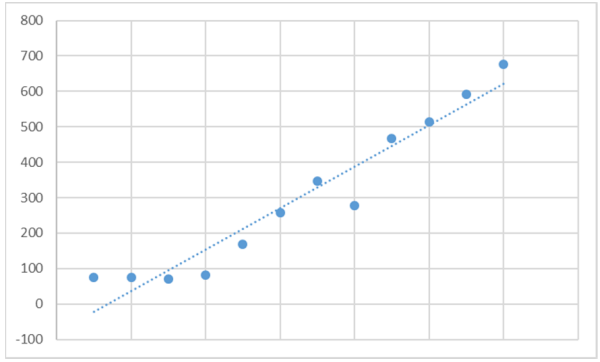
Data-to-parameter ratio (N_D/N_P)

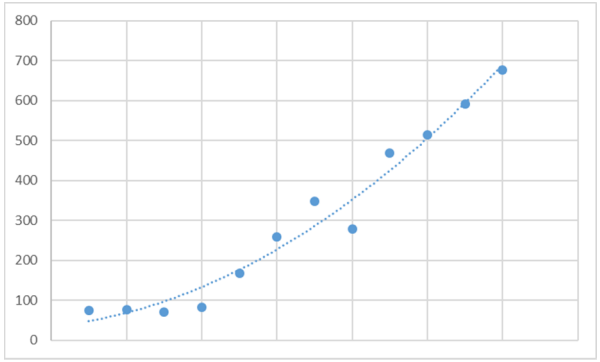
1. $N_D >> N_P$

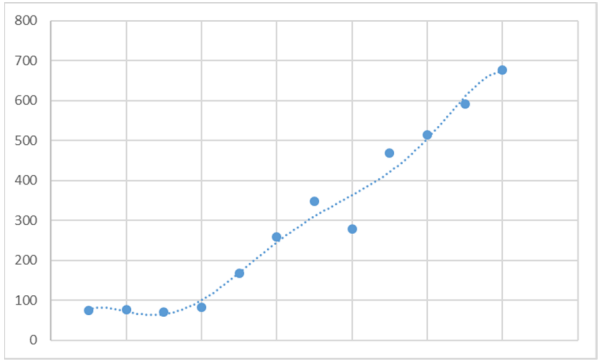




2. *The number of parameters should not be increased (even when the number of observations allows it) arbitrarly: the introduction of additional parameters should be justified by chemical reasons and by inspecting the electron density map.*







Restraints and constraints

Constraints: Relations between parameters that make some dependent from others

E.g. In rigid-body refinement, protein model is maintained rigid: atomic positions are not refined, reducing the number of parameters in the refinement.
E.g. For the phenyl group, the planarity of the aromatic ring can be expected from chemical considerations, the C-C distance is 1.39Å, and the CĈC angle is 120°.

Constraints reduce the number of parameters to refine!

Restraints: Relations between parameters based on statistical analysis that yield expected values with a defined uncertainty

E.g. For the carboxylic group of aspartate and glutamate residues, planarity is expected from chemical considerations. Atoms are not forced on the same plane, but deviations from the planarity are considered less probable occurrences. C-O distances are allowed within 0.1 Å from the expected value of 1.26 Å.

Restraints increase the number of observations!

Restraints

Geometric restraints: bond distances, bond angles, planarity, chirality... Usually dihedral angles are not restrained but analyzed as diagnostic parameters, i.e. Ramachandran plot.
Geometric restraints are usually tighter on main chain, looser on side chains.

Antibumbing restraints: distances between non-bonding atoms must be larger than van der Waals radii.

Non-Crystallographic Symmetry (NCS) restraints: core residues of NCS-related protein chains are likely to have similar conformations, while surface residues have more variable conformations.

B-factor restraints: atoms of the same group are restrained to have similar B-factors (particularly in case anisotropy is accounted for).

Libraries of restraints are available in the main refinement software and are automatically applied to each specific residue according to the residue type (which is written in the .pdb coordinate file).

Libraries of restraints for ligands or unusual cofactors must be prepared.

Target function

Least squares:

$$Q_{LS} = \sum_{i=1}^{n} \frac{[X_{obs}(i) - X_{calc}(i, \boldsymbol{p})]^2}{\sigma_{obs}^2(i)}$$

with $X_{obs}(i)$ observations, including diffraction data and additional knowledge and $\boldsymbol{p}$ parameters vector, i.e. a vector containing all parameters sequentially.

Good for high resolution data and complete and correct structural models. Considers Gaussian distribution for all errors on parameters.

Maximum likelihood:

$$Q_{ML} = \sum_{i=1}^{n} \frac{[X_{obs}(i) - \langle X_{calc}(i, \boldsymbol{p})\rangle]^2}{\sigma_{obs}^2(i) + \sigma_{calc}^2(i, \boldsymbol{p})}$$

with $\langle X_{calc}(i, \boldsymbol{p})\rangle$ expectation value of a Bayesian probability distribution, and including $\sigma_{calc}^2(i, \boldsymbol{p})$ to estimate non-random errors for the proposed model.

Better suited when the model is incomplete and/or partially incorrect: this function takes into account the conditional probability of model against data.

Refinement protocol

Details of the refinement:

- Parametrization: xyz? TLS?

TLS parametrization: uses translation-libration-screw parameters. Reduces number of parameters.

- Restrained or unrestrained?

Unrestrained only for high resolution datasets!

- Isotropic or anisotropic B-factors?

Increase of number of parameters with anisotropic B-factors: check N_D/N_P!

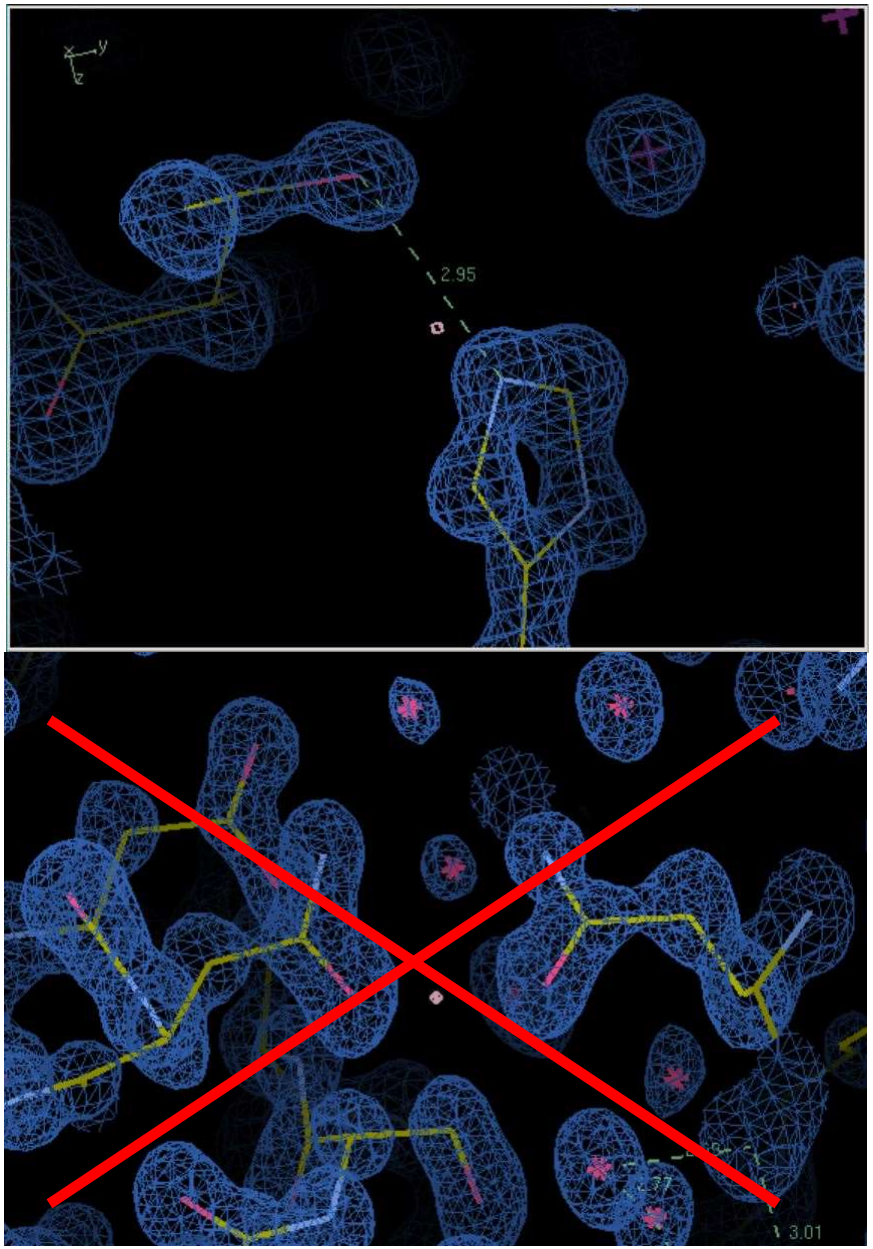
- Minimization function: Least squares (LS)? Maximum Likelihood (ML)? Energy minimization (based on Molecular Dynamics methods, e.g. simulated annealing)?

LS: high resolution structures and reliable models.
Energy refinement is alternative to restrained refinement: minimum of energy function.

- Optimization algorithm:
  - Gradient descent methods (full-matrix, sparse matrix, steepest descent...)
  - Stochastic algorithms (often with energy minimization)

Compromize between rate of convergence (speed of calculation) and radius of convergence (dimension of parameter space covered)

Side chain conformations



Side chain conformations should be adjusted in the electron density when visible.

Pay particular attention to orientation of His, Gln and Asn side chains: orientations are particularly impotant when structures are used for docking/drug design.

To distinguish side chain orientation, analyze hydrogen bonding network.

In some cases it is necessary to distinguish two different conformations for some residues (part. on the surface).

The sum of occupancy of the two conformations should be equal to 1 (or less).

Multiple conformations should not be introduced in the model if not clearly identified in the maps!

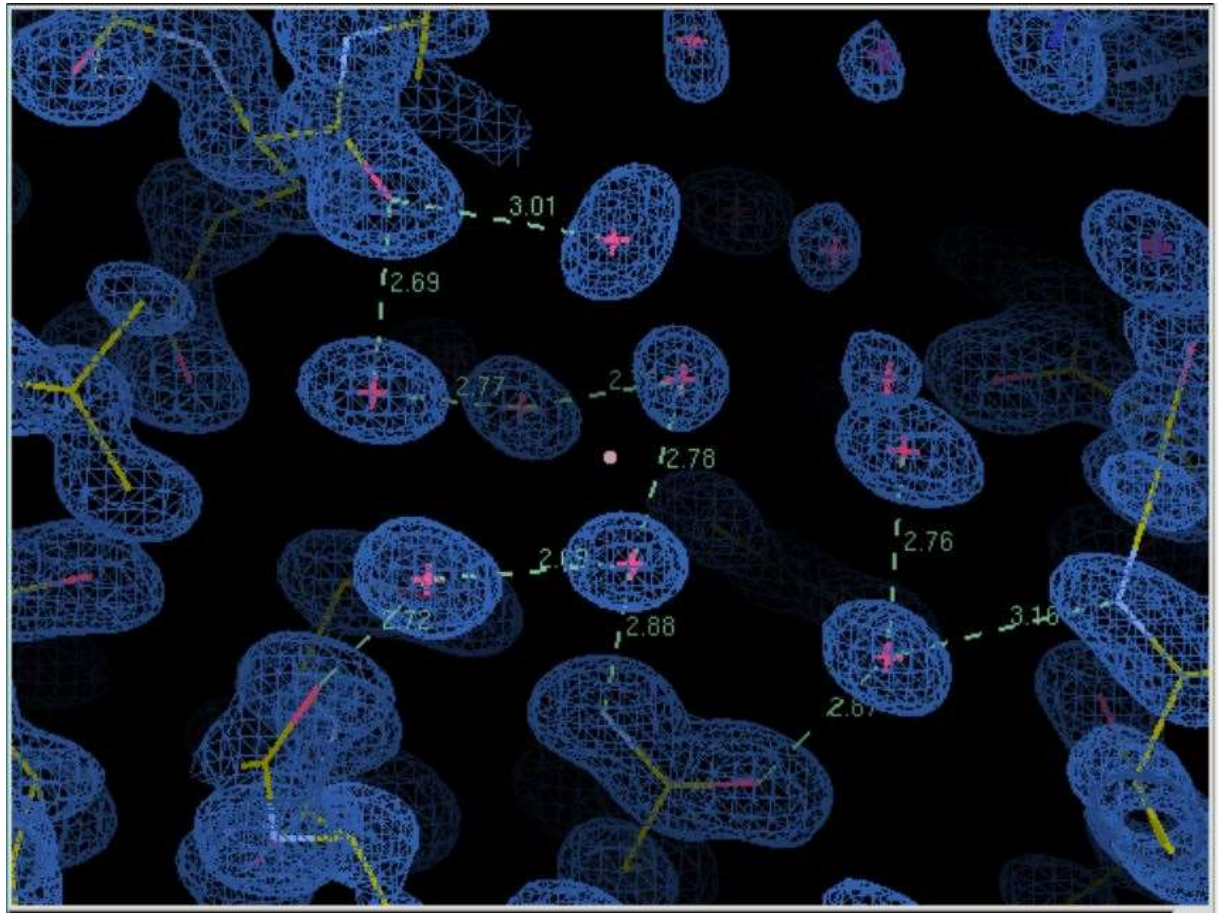
Solvent molecules and ligands

Solvent molecules: they should be introduced only if clear electron density is visible in the map.

Analyze contacts: water molecules should be at hydrogen bonding distance from protein residues.

Ions: check charge of residues in contact with the ions; check distances and compare them with reported distances.

Difference in electron density between, e.g., K^+ and Cl^- is not clear even at high resolution.



Ligands: The presence and conformation of ligands should be analyzed with particular care!! In this case, omit maps can be calculated to highlight unaccounted electron density.

R_{free} and R_{work}

Cross-validation: a subset of the reflections is set aside during refinement and used to validate the model obtained.

The small percentage of removed reflections (usually ≈ 5% of all data) does not affect the density maps.

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

This value is compared to the R-value obtained on data used during refinement:

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

A high value of R_{free} compared to R_{work} is indicative of **overparametrization** of the refinement

If R_{free} and R_{work} are too close, the refined model can be further improved. The expected difference between R_{work} and R_{free} depends on resolution.

Not all errors are highligthed by R-values: a careful inspection of the maps is still required!!