Model Refinement

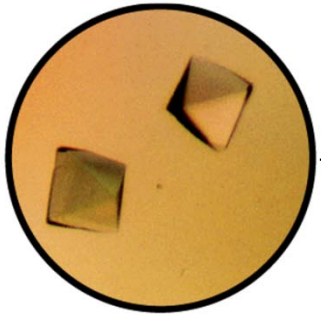
Oléron – 2018

J-L Ferrer
IBS, Grenoble

A large part of the slides from:
Pavel Afonine & **Paul adams**, Lawrence Berkeley National Lab, Berkeley, CA USA

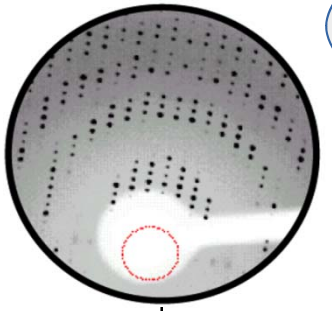
Structure determination workflow

Crystallization

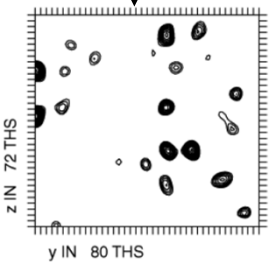


Data collection

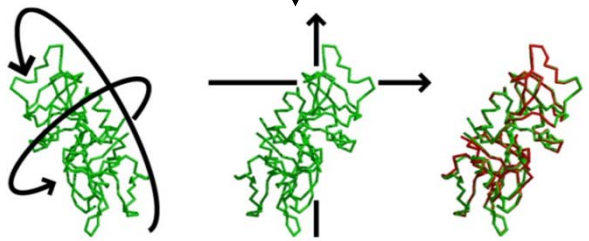
Data processing



XDS



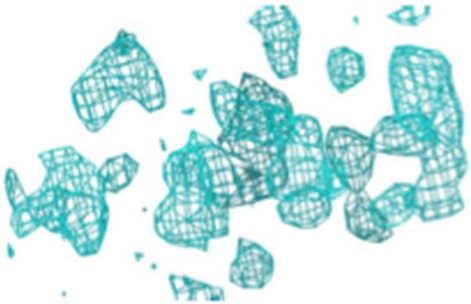
Anomalous scatterer location



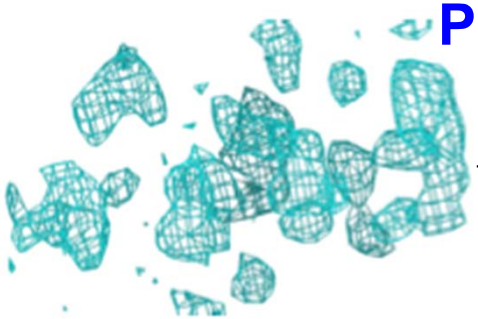
Molecular replacement

PHASER

Phase determination

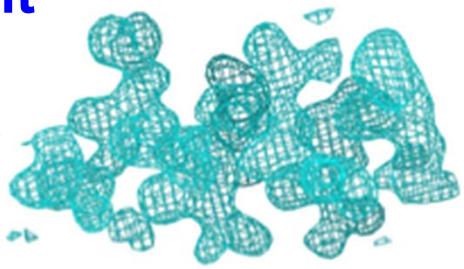


Structure refinement: iterative process of changing model parameters to optimally describe experimental data



Phase improvement

PARROT



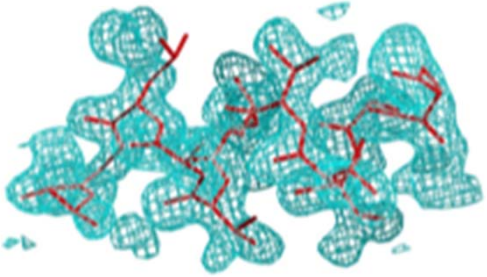
Map interpretation

BUCCANEER

Automated building

Manual reconstruction

COOT

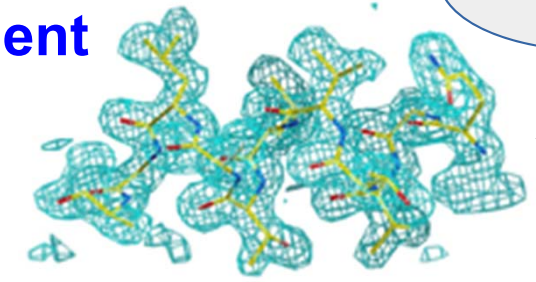


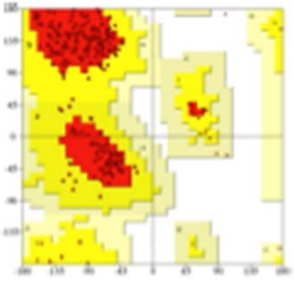
REFMAC

Manual reconstruction

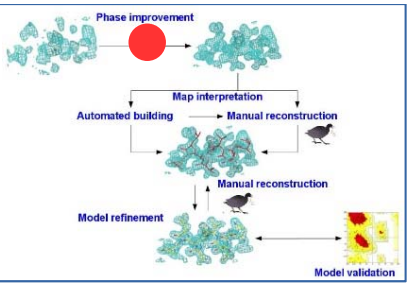
COOT

Model refinement





Model validation



Phase improvement

- Experimental phases (and those from molecular replacement) typically contain errors
- The experimental phases can be improved by the application of real space constraints
- The phases are modified to produce a map most consistent with what we know about macromolecular structures:
  - Solvent density distribution (Solvent flattening)
  - Atomicity and positivity (Sayre's equation)
  - Macromolecular density distributions (histogram matching)
  - Similarity between molecules (NCS averaging)

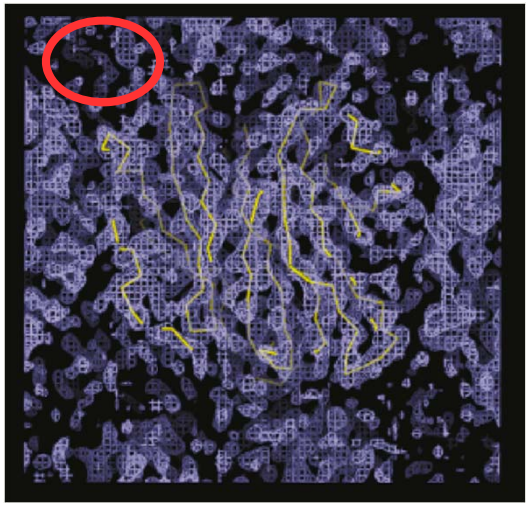
Phase improvement

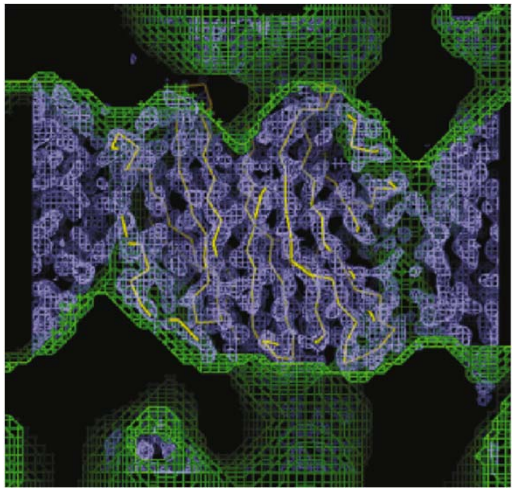
- Method to identify solvent versus macromolecular density in map
- Methods to determine relationships between different regions of the asymmetric unit
- Method to combine phase probability distributions (*e.g.* experimental phases with calculated phases)

Phase improvement: Identifying the Solvent Region

- **Experimental and MR-phased maps usually contain some information about the boundary of the macromolecule**
  - SAD and SIR maps are the combination of the correct map (made with the correct phase choice) and noise (a map made with the incorrect phase choice)
- **The envelope can be recovered by looking at the local standard deviation (the variance) of the electron density at each grid point in the map**
  - The standard deviation will be high in the macromolecular region and low in the solvent

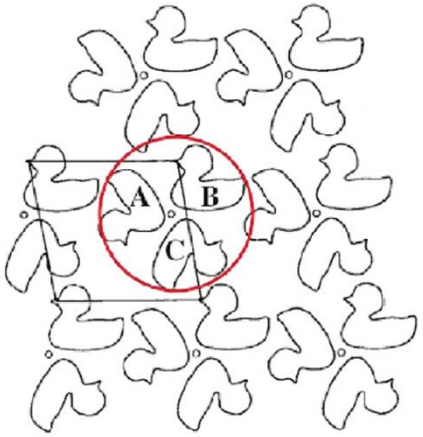
automated





Paul Adams, Macromolecular Crystallography School, Madrid, May 2017
Image from G. Taylor, Acta Cryst. D, 59, 1881-1890 (2003)

Phase improvement: Non-crystallographic Symmetry



$$\rho(\mathbf{x}_B) = \mathbf{R}_B\rho(\mathbf{x}_A) + \mathbf{t}_B$$
$$\rho(\mathbf{x}_C) = \mathbf{R}_C\rho(\mathbf{x}_A) + \mathbf{t}_C$$

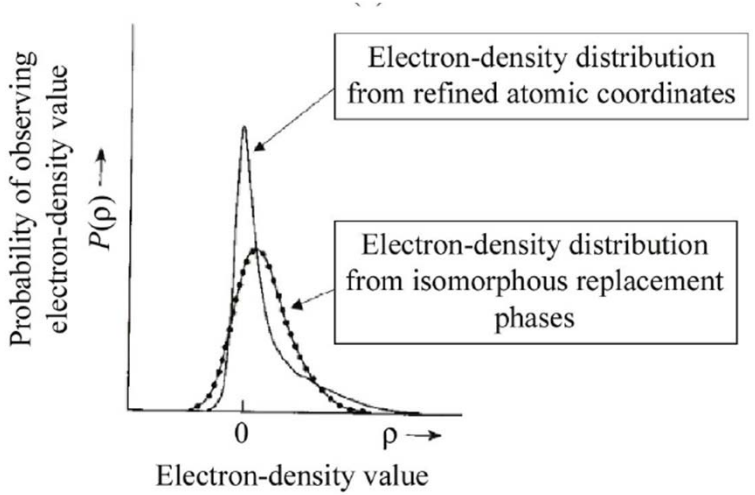
NCS symmetry operators

automated

- **The presence of multiple copies of the same molecule in the asymmetric unit provides additional information in phase improvement**
  - Electron density can be averaged to enforce the NCS relationship
  - The similarity of the related regions can be used as an indicator of the success of phase improvement
- **The relationship between molecules and the mask around them must be defined**
  - NCS is often referred to as proper (2-fold, 3-fold, 4-fold etc.) or improper (an arbitrary relationship between molecules
  - NCS is quite common

Paul Adams, Macromolecular Crystallography School, Madrid, May 2017
Image from G. Taylor, Acta Cryst. D, 59, 1881-1890 (2003)

Phase improvement: Histogram matching



automated

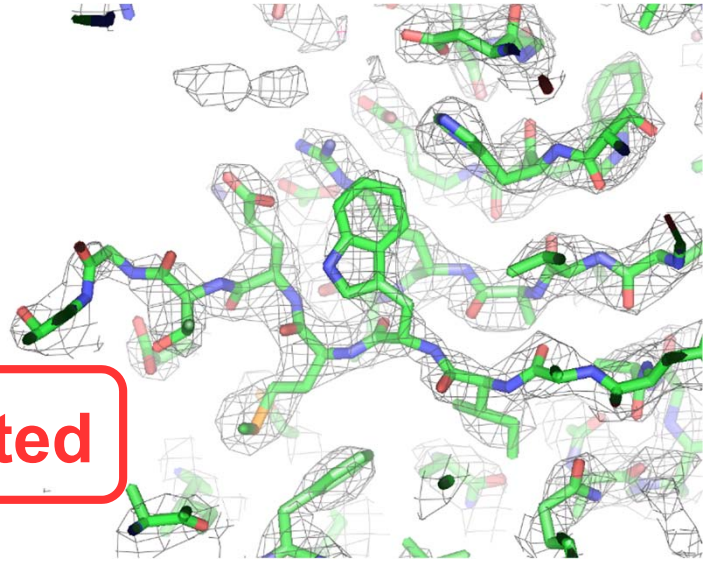
- **The electron density of macromolecules have fairly similar distributions (but are dependent on the type of molecule and the resolution)**
  - This information can be used to match the observed histogram of densities to an ideal histogram
  - This is one of the most powerful constraints on the density (and hence in phase improvement)
  - The histogram matching method is not unique to crystallography (used in many different image processing applications)

Image from G. Taylor, Acta Cryst. D, 59, 1881-1890 (2003)

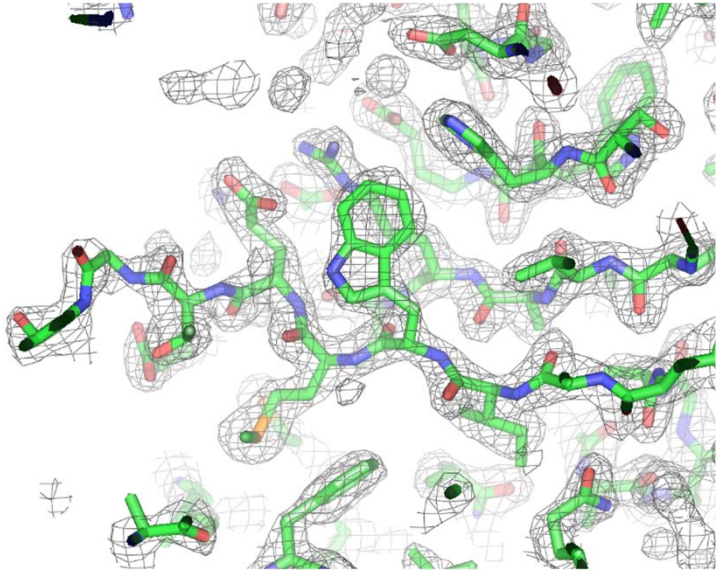
Phase improvement: Phases extension

- **Phase Extension: Sometimes high resolution native data are available in addition to the data from the phasing experiment**
  - Phase extension works because long-range relationships in the electron density (such as NCS or solvent region) lead to short range relationships in reciprocal space.
  - Determining the phases at a given resolution limit also generates some useful information about reflections at a slightly higher resolution.

automated

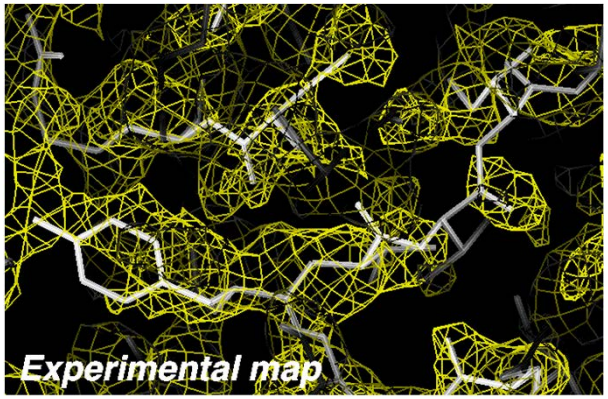


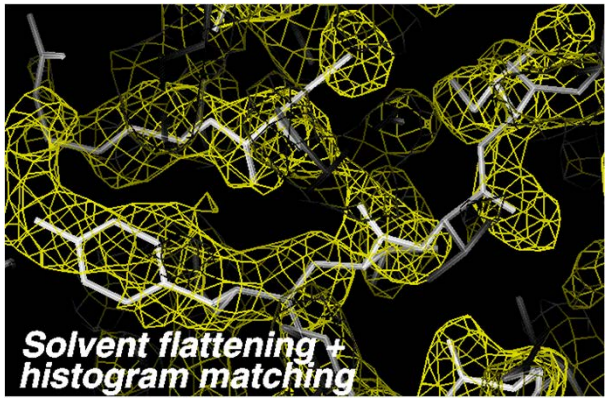
Density modified map at 3Å

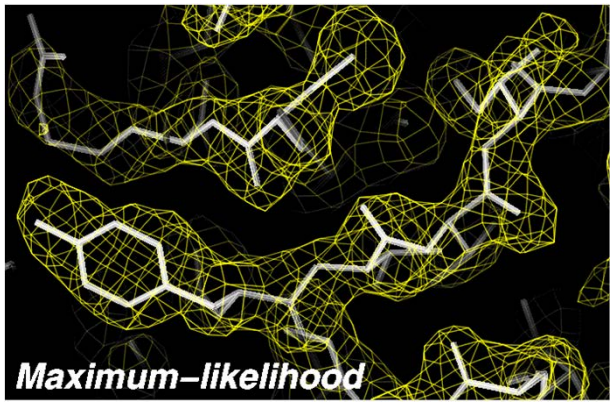


Density modified map at 2Å

Phase improvement: Stastistical phase improvement



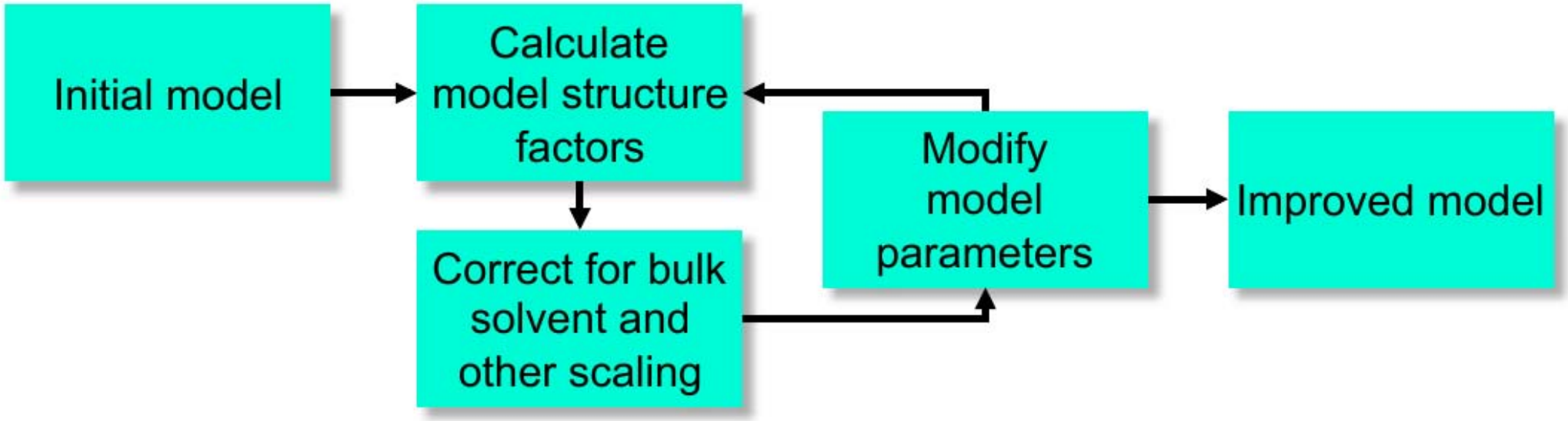




Maximum likelihood estimation is a method that determines values for the parameters of a model. The parameter (model) values are found such that they maximise the likelihood that the process described by the model produced the data that were actually observed.

automated

From experimental map to final model



Refinement:

$\rho_{model} \rightarrow \rho_{exp}$

$\{F_{model}\} \rightarrow \{F_{obs}\}$

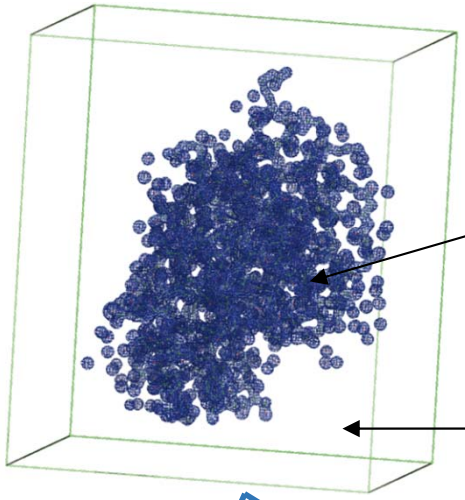
1)Calculate F_{model} from current model

2)Score how F_{model} against F_{obs}

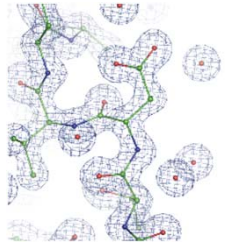
3)Change model so F_{model} are closer to F_{obs}

4)Go to step 1

From F to atomic model and back



Macromolecule



Bulk-solvent

$$\rho_{crystal} = \rho_{model} = \rho_{atoms} + \rho_{bulk\ solvent}$$

Fourier transformation

- All atoms contribute to each structure factor
- Fourier transform of a Gaussian function is a Gaussian function
- Fourier transform is a linear operation

Structure factors: $F(s)$

Structure refinement is optimization problem

Model parameters

Parameters that describe crystal and its content

- Atomic model (coordinates, B-factors, occupancies)
- Non-atomic model (bulk-solvent, anisotropy, twining)

Scoring function (refinement target, optimization goal)

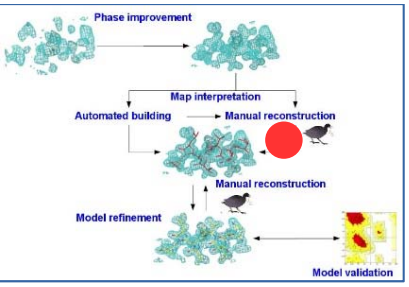
A function that relates model parameters and experimental data

- Least-squares (LS), Maximum-Likelihood (ML), R-factor, …

Optimization method

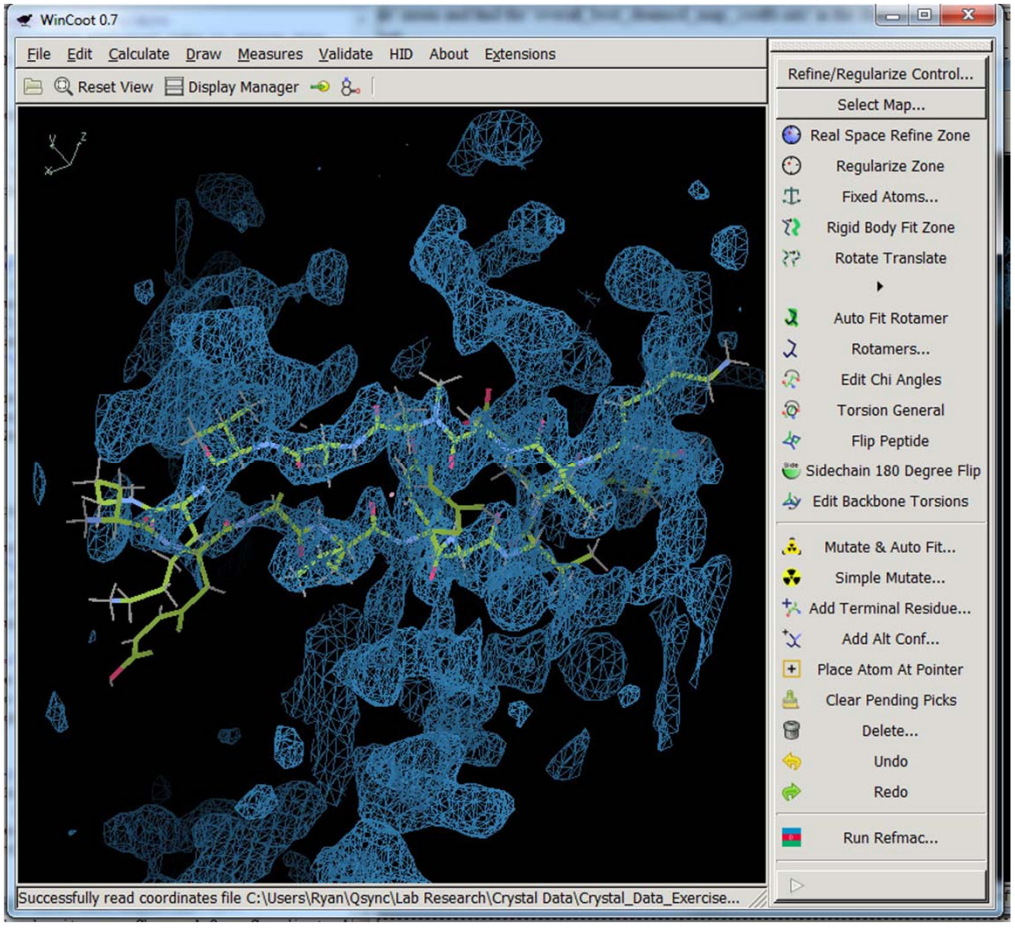
A tool that varies model parameters to optimize refinement target

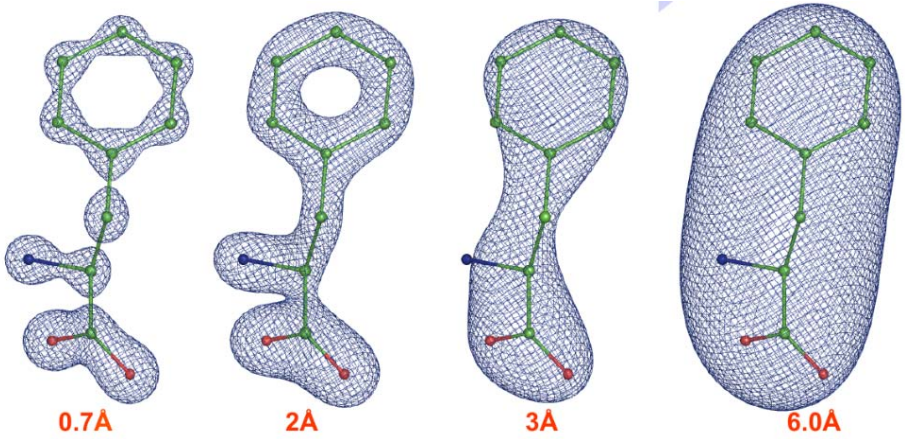
- Gradient-based minimization
- Simulated Annealing (SA)
- Systematic (grid) searches
- Changing parameters by hand (using graphics, like Coot)



Map interpretation

Resolution





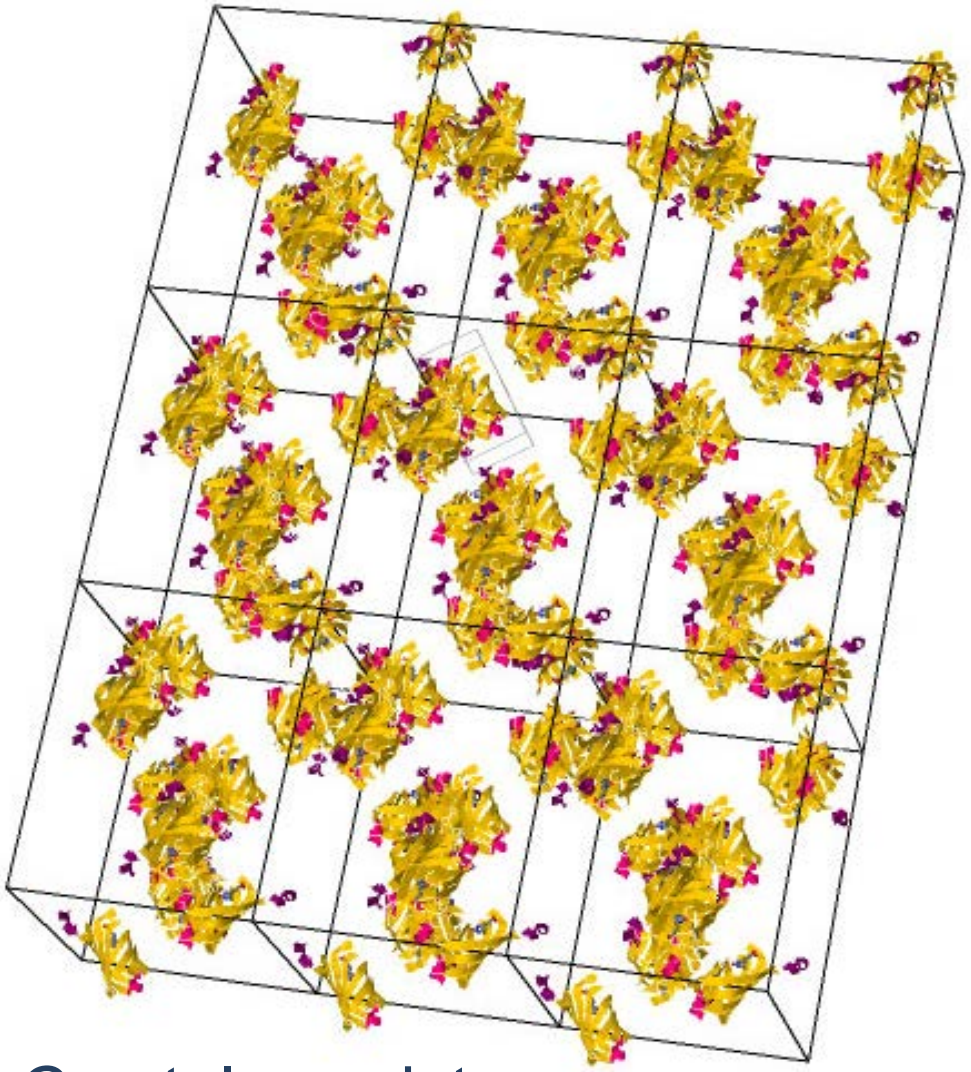
High res., Å	Size of details	Features of images
~0.7	half of a X-X bond	deformation density
~0.9	X-H bond	some deformation density
~1.0		H atoms
~1.2	shortest covalent bond	individual atoms
~1.5	C_α-C	
~2.0		most of ordered solvent
~2.5	distance N-C_β, C-C_β, N-O, C_α-O	clear side chains
~3.5	inter-C_a-distance	side chains may be guessed
~4.5	distance between chains	main chain
~6		α-helices
~12-15		
~20	small domains	molecular envelopes

manual

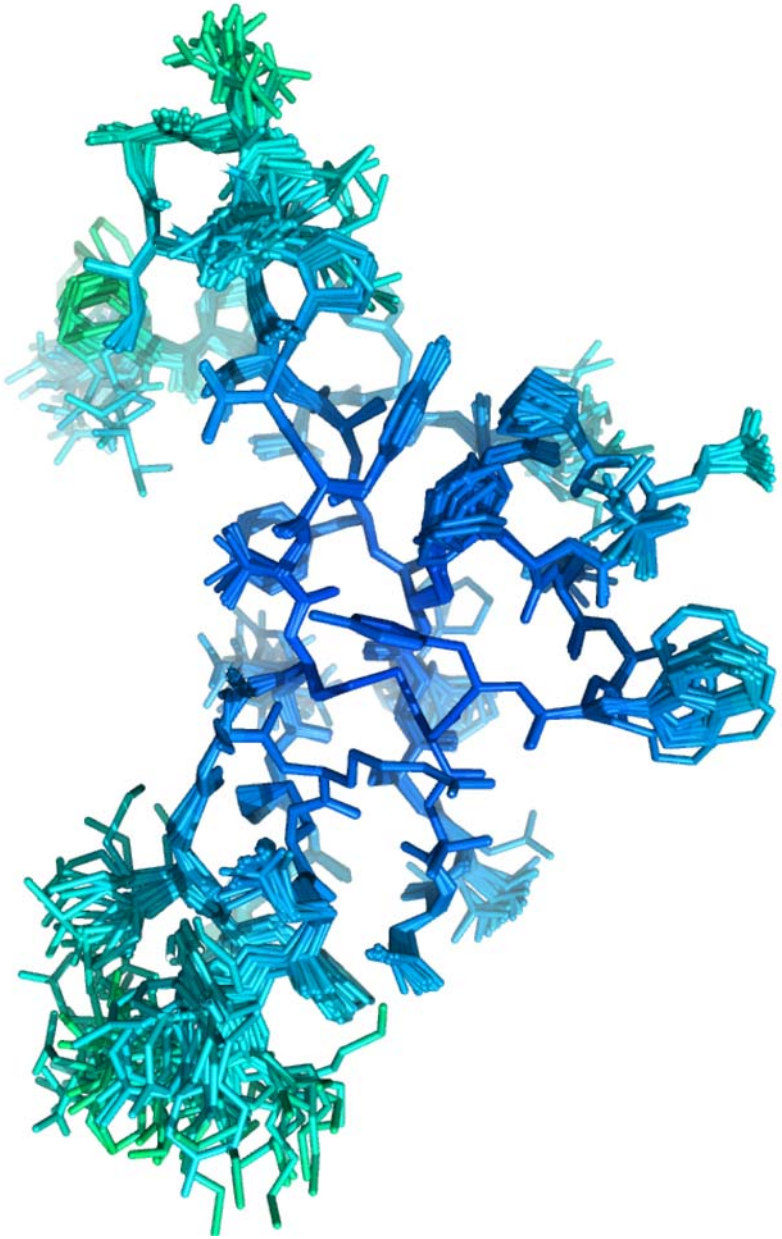
Completeness

+ lack of completness: higher degradation of the map if uncomplete low resolution

Atomic model parameters: disorder

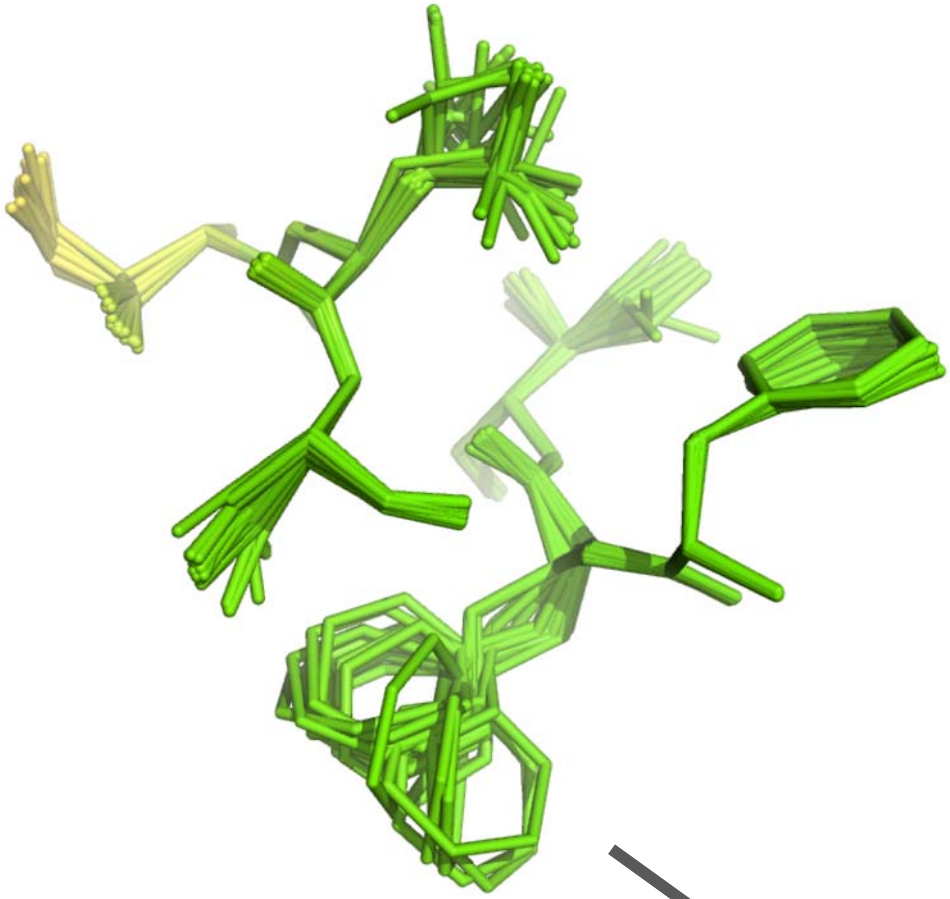


Crystal consists
of may unit cells



Superpose all structures
from each unit cell

Atomic model parameters



automated

Small disorder

Accounted for using B-factor

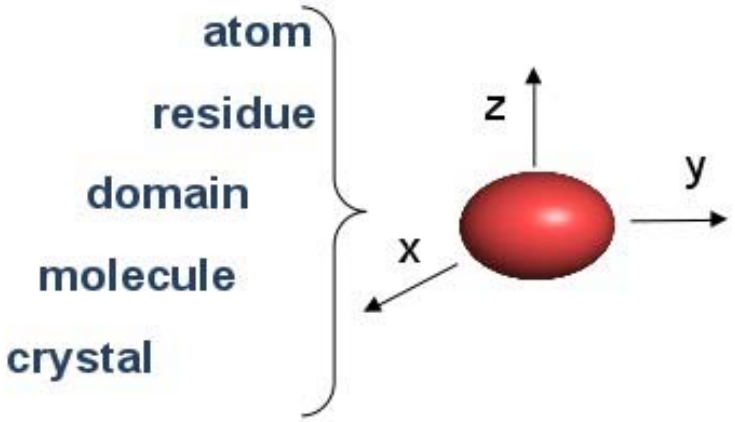
Modeled with occupancy

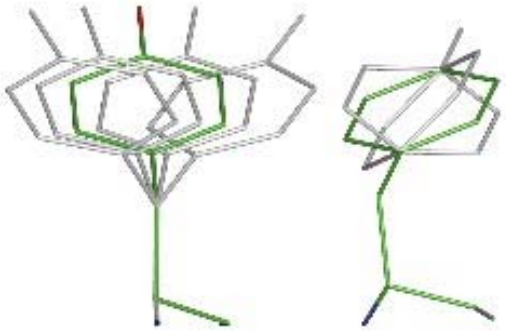
Large disorder

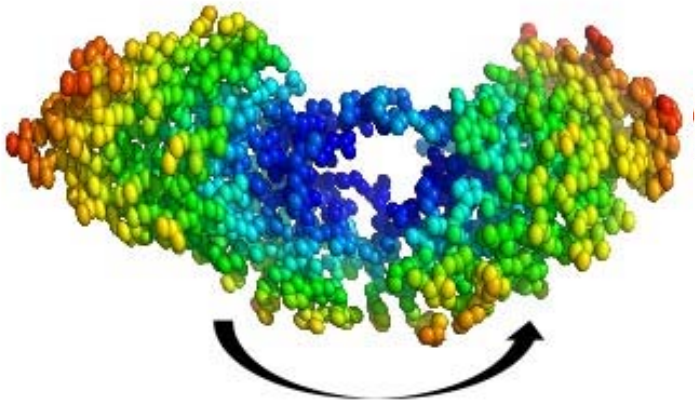
manual

```
ATOM      25  CA  PRO A   4      31.309  29.489  26.044  1.00 57.79           C
ANISOU    25  CA  PRO A   4     8443   7405   6110   2093    -24    -80       C
```

Atomic Displacement Parameters (ADP, B-factors)

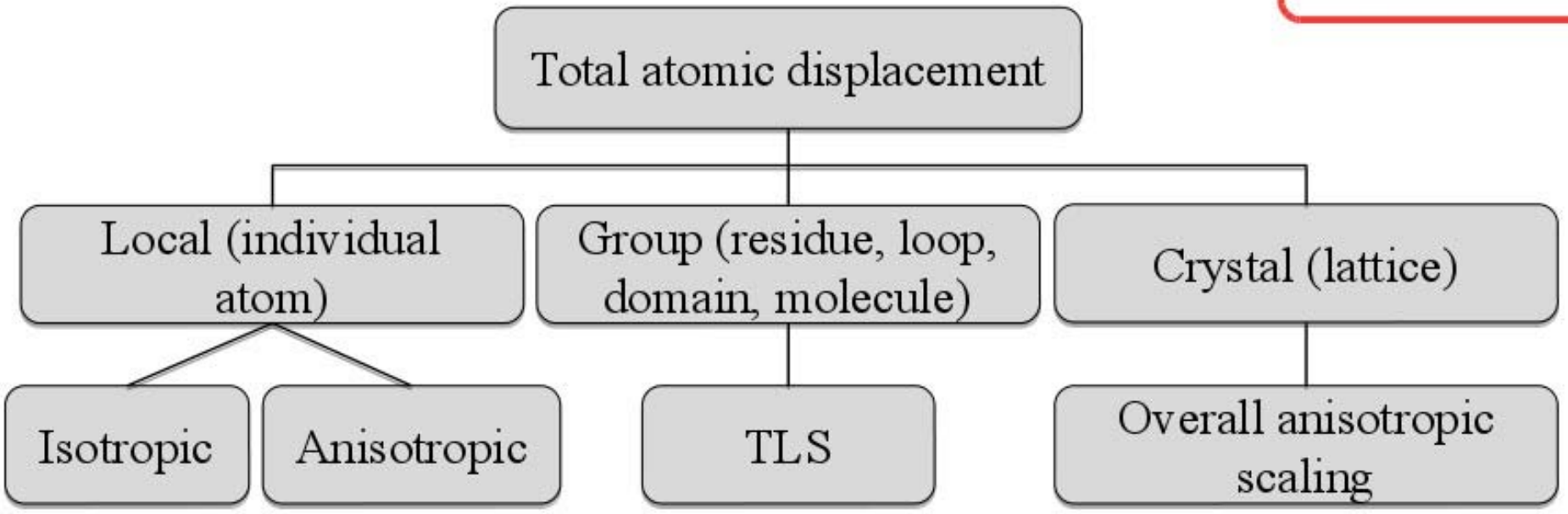






- **Parameterization for refinement**

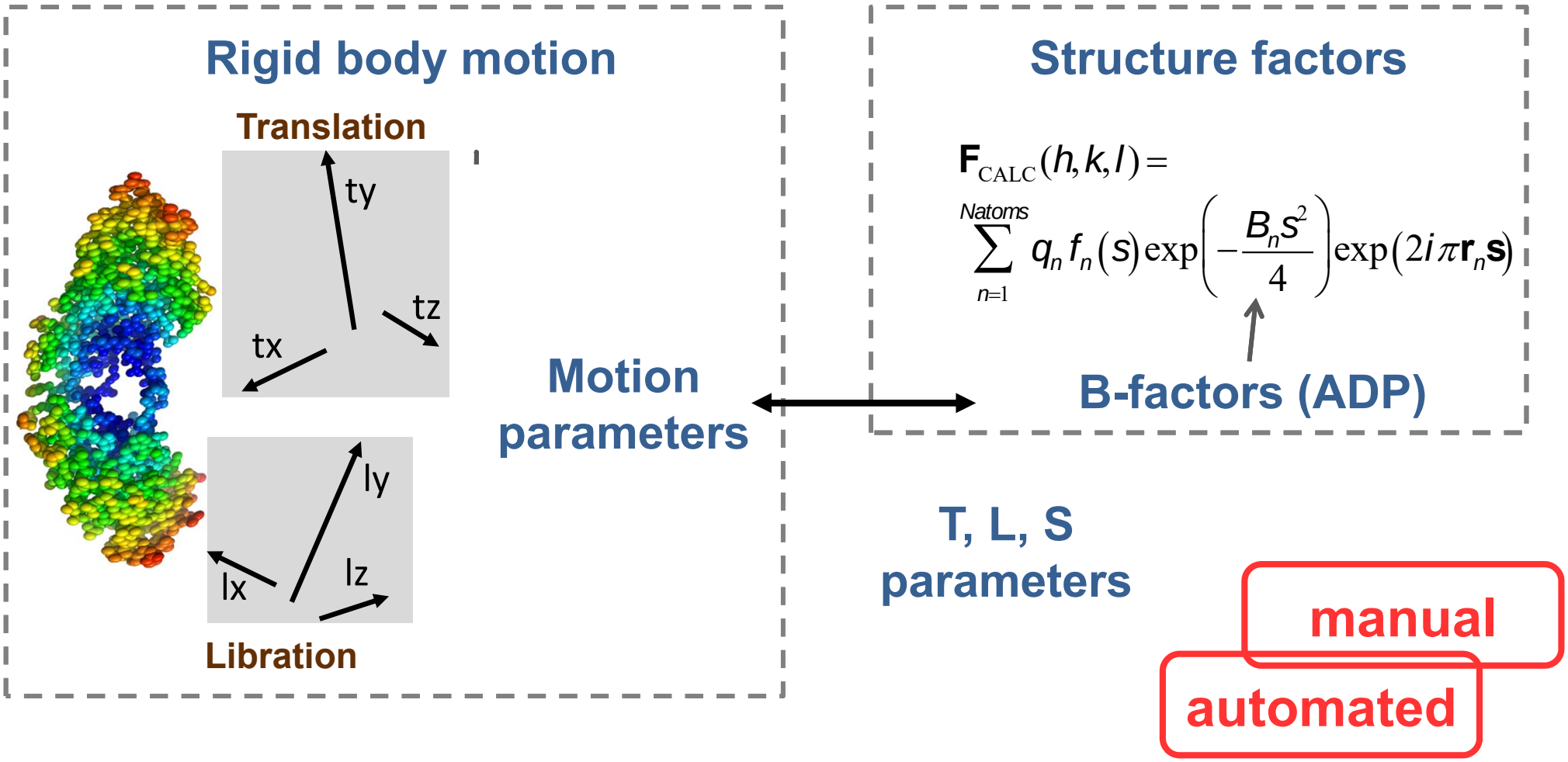
automated



$$U_{TOTAL} = U_{CRYST} + U_{GROUP} + U_{LOCAL}$$

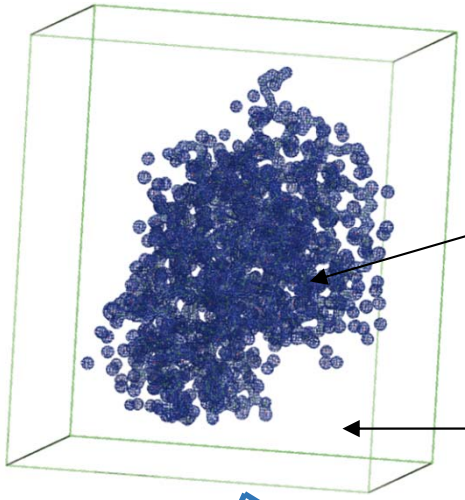
TLS: Translation-Libration-Screw motion model

- TLS parameters is a way to pack descriptors of rigid-body motion into a form suitable to calculate structure factors

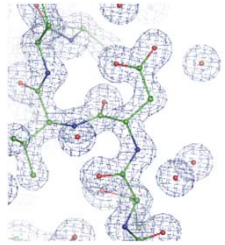


- **TLS do not directly provide parameters of rigid body motion**
  Descriptions of motions need to be extracted from TLS matrices

From F to atomic model and back



Macromolecule



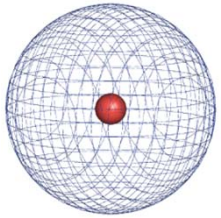
Bulk-solvent

$$\rho_{crystal} = \rho_{model} = \rho_{atoms} + \rho_{bulk\ solvent}$$

Electron density distribution: basic model

- Gaussian function – a good approximation of an atom
  - Convenient computationally (FT of a Gaussian is a Gaussian)

- *Isotropic* distribution of electron density at point r of isolated atom:



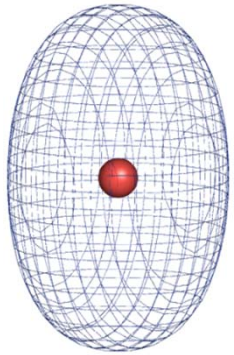
$$\rho_{atom}(\mathbf{r},\mathbf{r}_0,B,q) = q\sum_{k=1}^{5} a_k \left(\frac{4\pi}{b_k + B}\right)^{3/2} \exp\left(-\frac{4\pi^2 |\mathbf{r}-\mathbf{r}_0|^2}{b_k + B}\right)$$

Number of terms depends on how accurately we want to model an atom

a_k and b_k are atom-specific tabulated values

Electron density distribution: anisotropic model

- More accurate approximation assumes atoms moving anisotropically



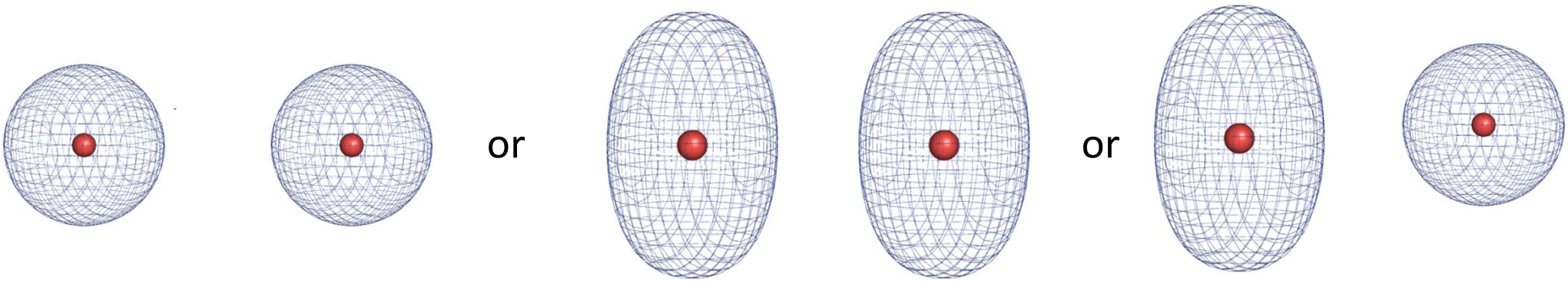
$$\rho_{atom}(\mathbf{r},\mathbf{U},q) = q\sum_{j=1}^{5} \frac{q\,a_j\,(4\pi)^{3/2}}{\left|8\pi^2\mathbf{U}_{\text{cart}} + b_j\mathbf{I}\right|^{1/2}} \exp\left(-4\pi^2(\mathbf{r}-\mathbf{r}_0)^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\left[8\pi^2\mathbf{U}_{\text{cart}} + b_j\mathbf{I}\right]^{-1}\mathbf{A}(\mathbf{r}-\mathbf{r}_0)\right)$$

$\mathbf{U}_{\text{cart}}$ – anisotropic atomic displacement parameters (3*3 symmetric matrix)

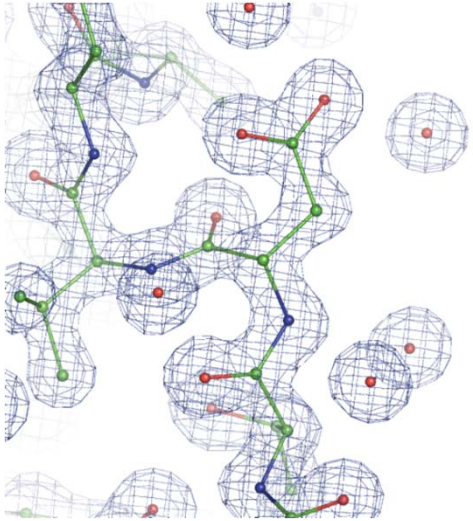
$\mathbf{U}_{\text{cart}}$ is what is in ANISOU records of PDB files

Electron density distribution: mixed model

- **Electron density of whole molecule is a sum of electron densities of individual atoms (isotropic or anisotropic)**



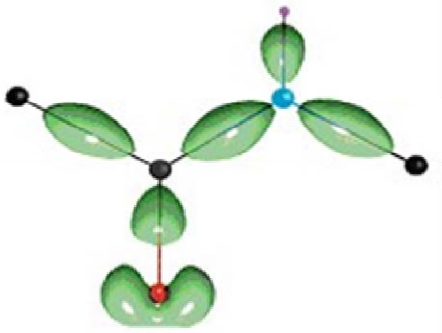
- **Bonding effects are ignored (atoms isolated):**



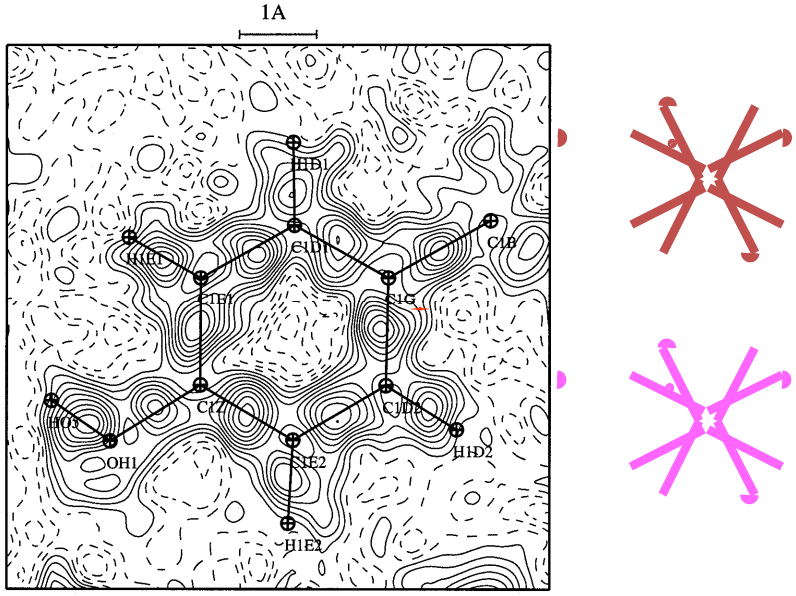
$$\rho_{crystal}(\mathbf{r}) = \sum_{i=1}^{Natoms} \rho_{atoms}(\mathbf{r})$$

Electron density distribution: high resolution model

- **Even more accurate approximation assumes atoms are bonded: multipolar model**



$$\rho_{\text{atom}}(\mathbf{r}) = \rho_{core}(\mathbf{r}) + P_{val}\kappa^3\rho_{val}(\kappa\mathbf{r}) + \sum_{l=0}^{l_{\max}}\kappa'^3 R_l(\kappa'\mathbf{r})\cdot\sum_{m=-l}^{l} P_{lm}y_{lm}(\theta,\varphi)$$



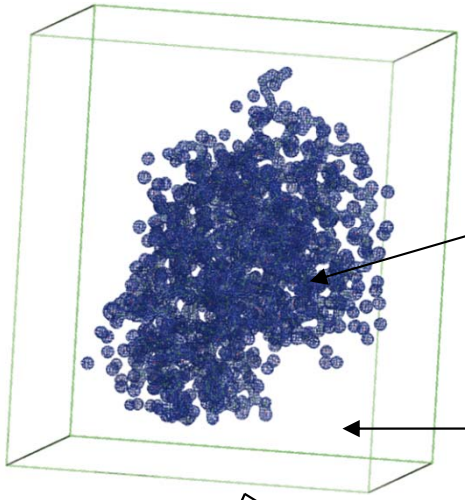
ρ_{ATOM} = core electrons

+ valence electrons

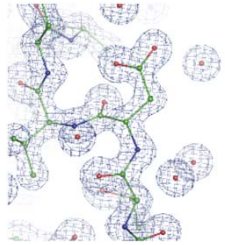
+ non-spherical part of the valence electron distribution

- **Used at ultra-high resolution (better than 1Å)**

From F to atomic model and back

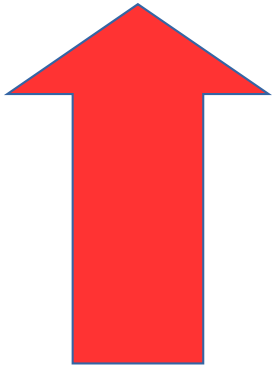


Macromolecule



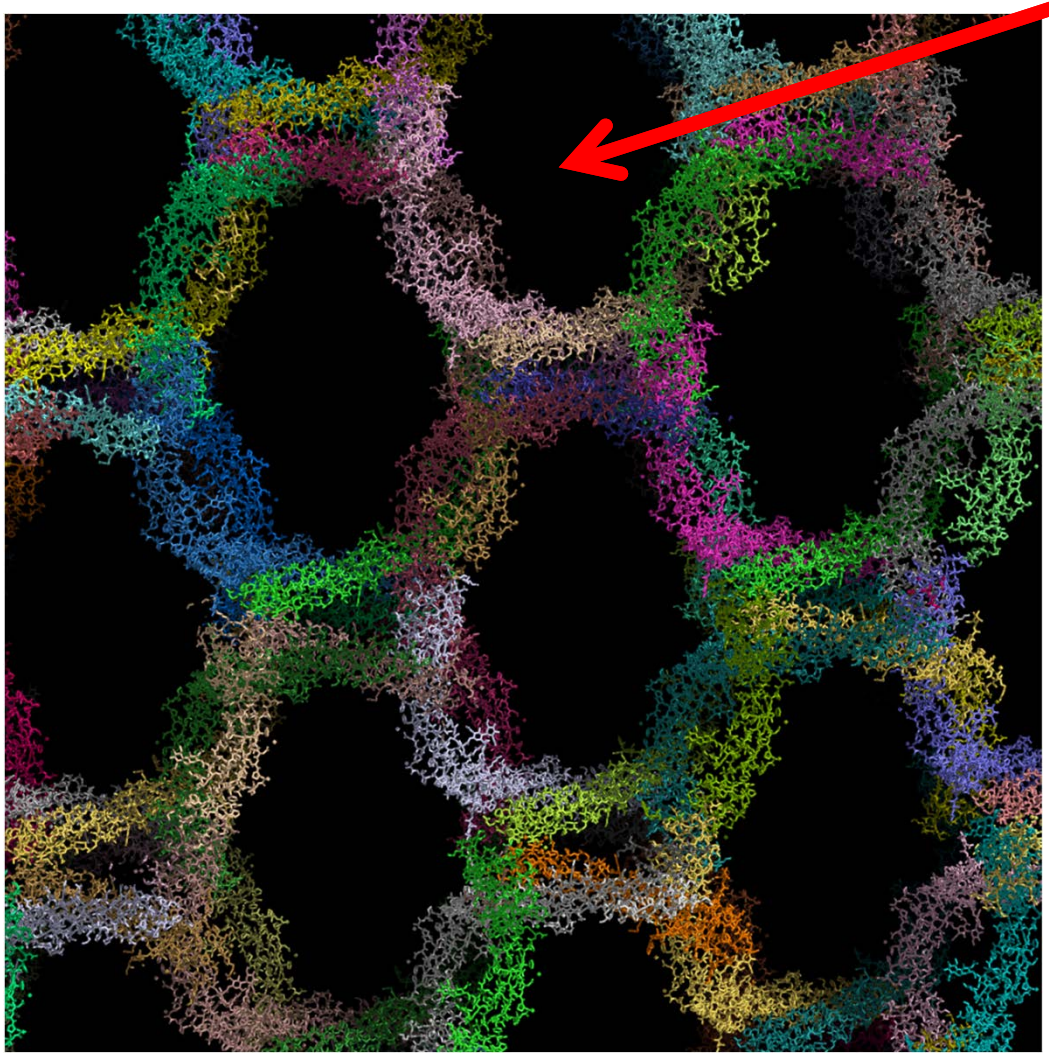
Bulk-solvent

$$\rho_{crystal} = \rho_{model} = \rho_{atoms} + \rho_{bulk\ solvent}$$

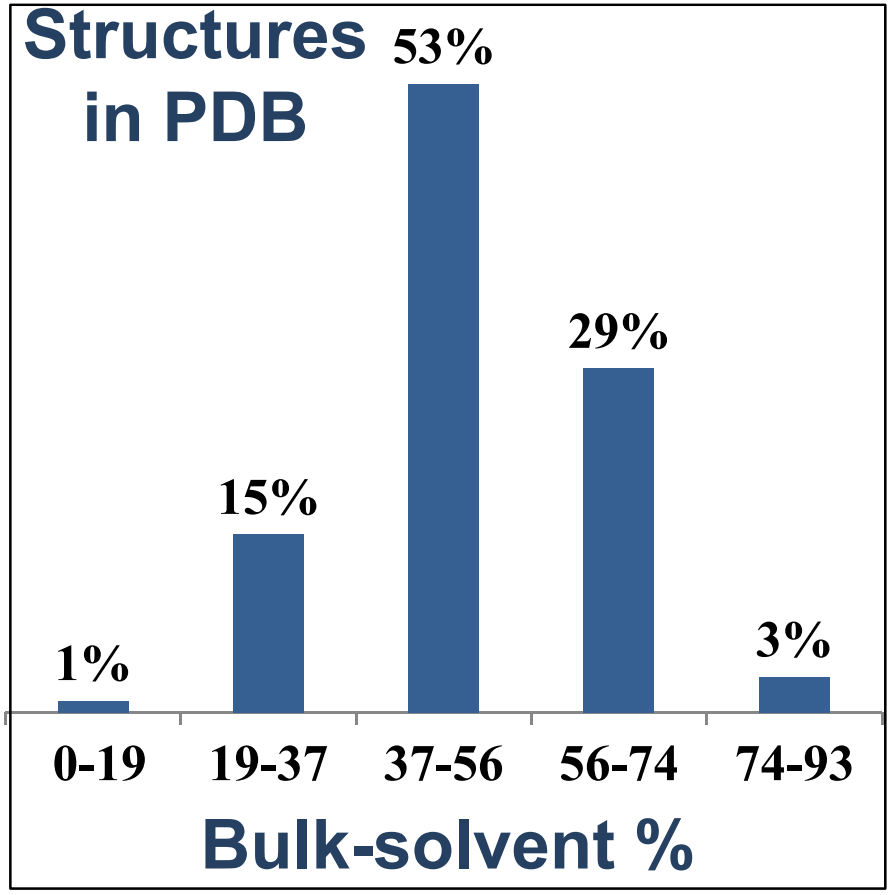


Crystal structure

PDB code: 1QUB

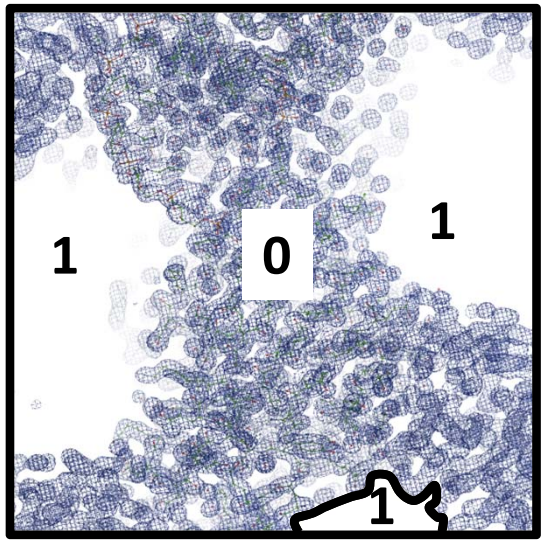


Bulk-solvent



Crystal structure: $\rho_{crystal} = \rho_{atoms} + \rho_{bulk\ solvent}$

Flat bulk solvent model



automated

Steps to account for bulk-solvent:

1. Compute solvent mask, M:

   0 – inside protein, 1 – outside

2. Structure factors from M:

   F_{MASK}= FT(M)

2. Define solvent contribution F_{BULK}:

   $F_{BULK} = k_{MASK} * F_{MASK}$

4. Combine with $F_{CALC(ATOMS)}$

   Refine k_{MASK} by fitting $|F_{MODEL}|$ to F_{obs}

$$F_{MODEL} = k_{OVERALL}\left(F_{CALC\,(ATOMS)} + F_{BULK}\right)$$

- Regions not interpreted by atoms are filled with constant density (bulk solvent)

Structure refinement is optimization problem

Model parameters

Parameters that describe crystal and its content

- Atomic model (coordinates, B-factors, occupancies)
- Non-atomic model (bulk-solvent, anisotropy, twining)

Scoring function (refinement target, optimization goal)

A function that relates model parameters and experimental data

- Least-squares (LS), Maximum-Likelihood (ML), R-factor, …

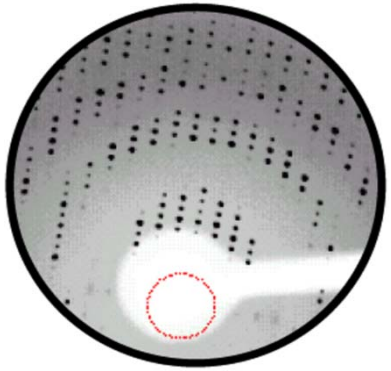
Optimization method

A tool that varies model parameters to optimize refinement target

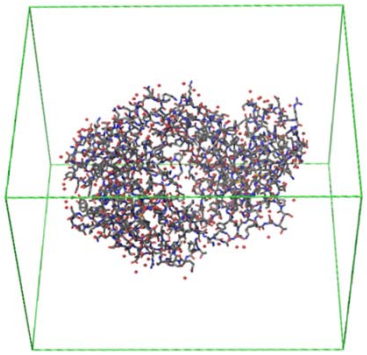
- Gradient-based minimization
- Simulated Annealing (SA)
- Systematic (grid) searches
- Changing parameters by hand (using graphics, like Coot)

Refinement target: reciprocal space

Experimental data



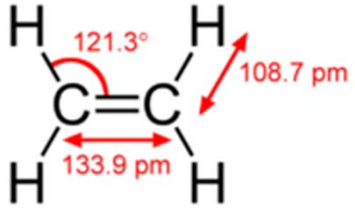
Model



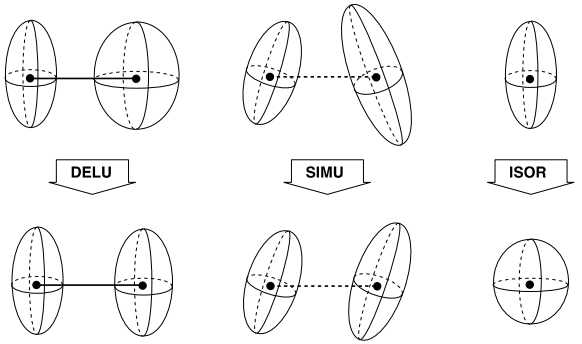
Model parameters:
Coordinates, B-factors,
Bulk-solvent…

A *priori* knowledge about model

Covalent geometry:



Similarity of B-factors



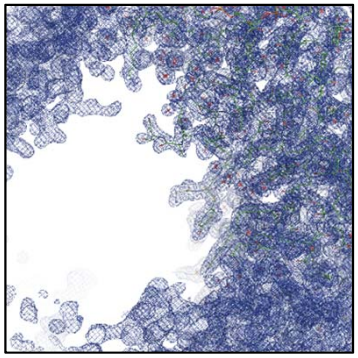
$\mathbf{F}_{\mathrm{OBS}}$

$\mathbf{F}_{\mathrm{MODEL}}$

$$T = T_{\mathrm{DATA}}(F_{\mathrm{OBS}}, F_{\mathrm{MODEL}}) + wT_{\mathrm{RESTRAINTS}}$$

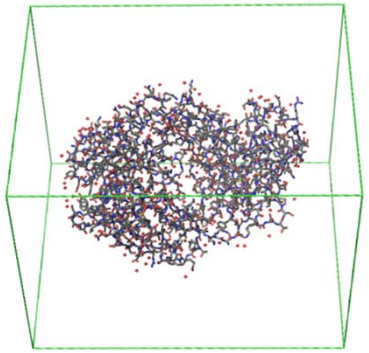
Refinement target: real space

Experimental data



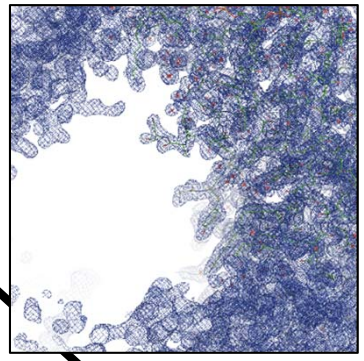
ρ_{OBS}

Model



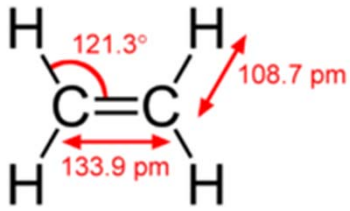
Model parameters: XYZ, B-factors, Bulk-solvent…

ρ_{MODEL}

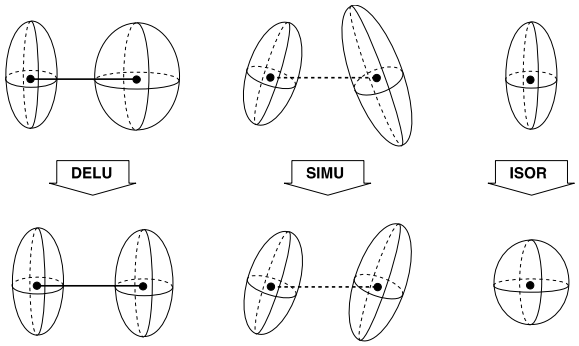


A *priori* knowledge about model

Covalent geometry:



Similarity of B-factors



$$T = T_{DATA}(\rho_{OBS}, \rho_{MODEL}) + wT_{RESTRAINTS}$$

Refinement target: reciprocal space

$$T_{\mathrm{DATA}}(F_{\mathrm{OBS}}, F_{\mathrm{MODEL}})$$

- **Least-Squares**

$$T_{\mathrm{DATA}} = \sum_{\mathbf{s}} \mathbf{w}\left(F_{\mathrm{OBS}} - F_{\mathrm{MODEL}}\right)^2$$

  - Used in small molecule crystallography
  - Used in macromolecular crystallography in the past

- **Maximum-Likelihood**

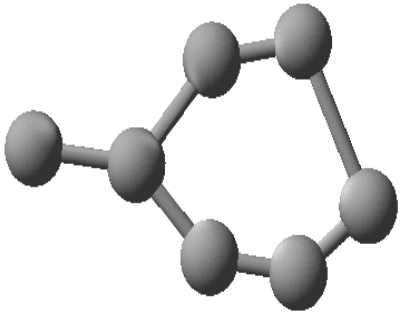
$$T_{\mathrm{DATA}} = \sum_{\mathrm{s}} (1-K_{\mathbf{s}}^{CS})\left(-\frac{\alpha_{\mathbf{s}}^2\left(F_{\mathbf{s}}^{MODEL}\right)^2}{\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}} + ln\left(I_0\left(\frac{2\alpha_{\mathbf{s}}F_{\mathbf{s}}^{MODEL}F_{\mathbf{s}}^{OBS}}{\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}}\right)\right)\right) + K_{\mathbf{s}}^{CS}\left(-\frac{\alpha_{\mathbf{s}}^2\left(F_{\mathbf{s}}^{MODEL}\right)^2}{2\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}} + ln\left(\cosh\left(\frac{\alpha_{\mathbf{s}}F_{\mathbf{s}}^{MODEL}F_{\mathbf{s}}^{OBS}}{\varepsilon_{\mathbf{s}}\beta_{\mathbf{s}}}\right)\right)\right)$$

  - Option of choice for macromolecules

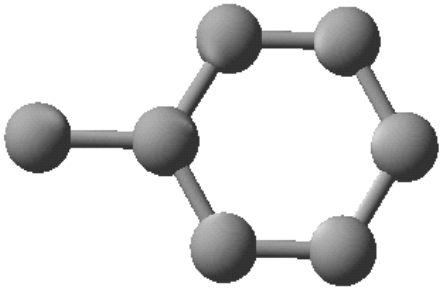
Least-Squares vs Maximum-Likelihood refinement

- **Complete model with errors**

Before refinement

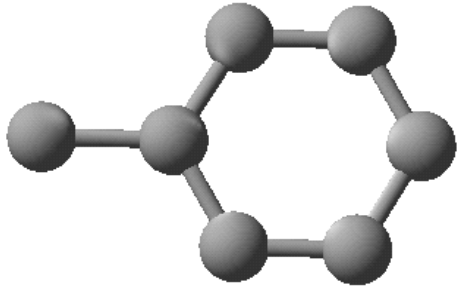


Least-Squares

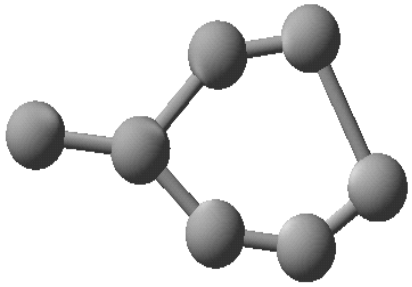


After refinement

- **Partial model, no errors**

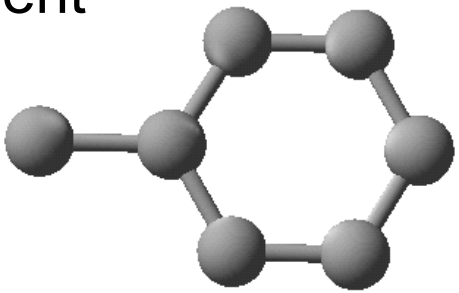


Least-Squares

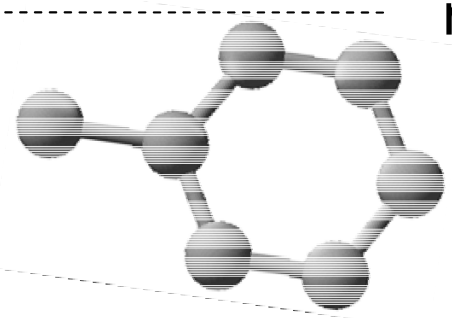


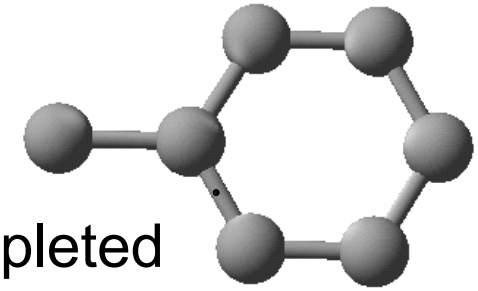
Before refinement

After refinement



Maximum-Likelihood

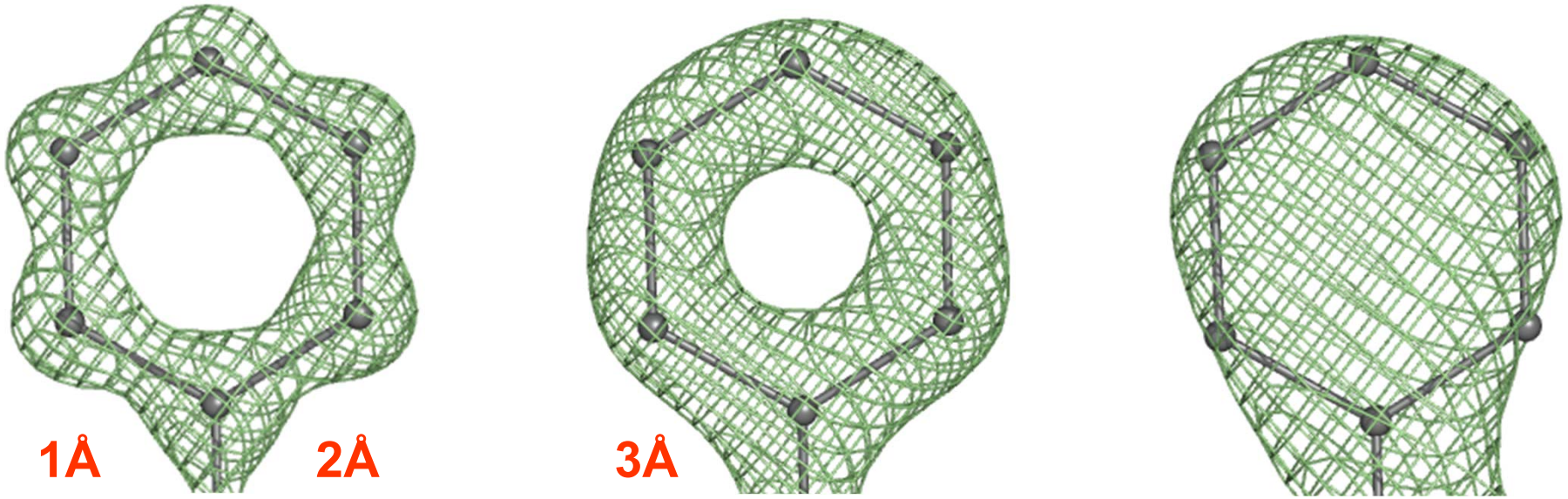




Model is completed statistically (implicitly)

Final model is less affected by missing atoms

Restraints for coordinate refinement



- Lower the resolution, less detailed the map
- Need extra information to keep correct geometry during refinement

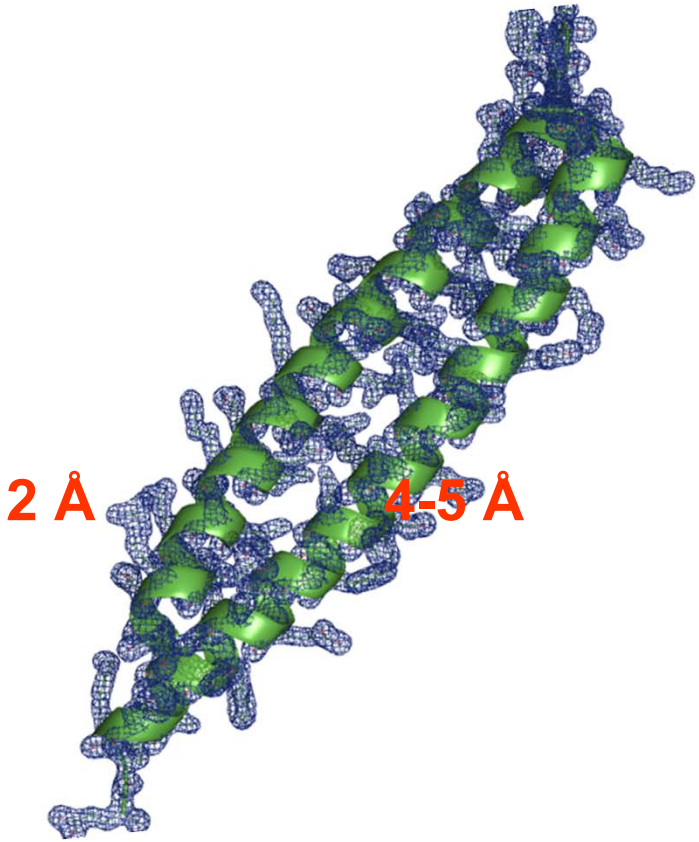
$$T = T_{\mathrm{DATA}}(F_{\mathrm{OBS}}, F_{\mathrm{MODEL}}) + wT_{\mathrm{RESTRAINTS}}$$

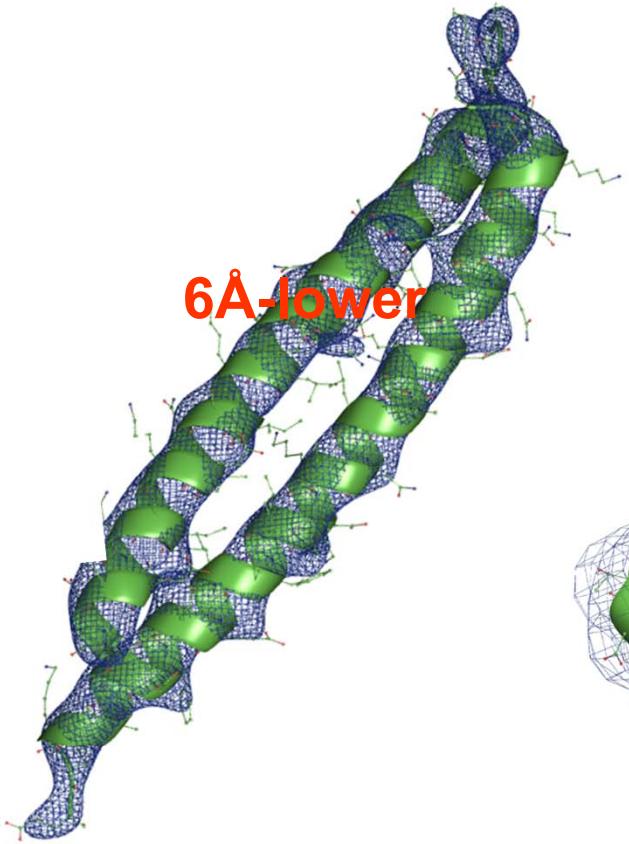
$$T_{RESTRAINTS} = T_{BOND} + T_{ANGLE} + T_{DIHEDRAL} + T_{PLANARITY} + T_{NONBONDED} + T_{CHIRALITY}$$

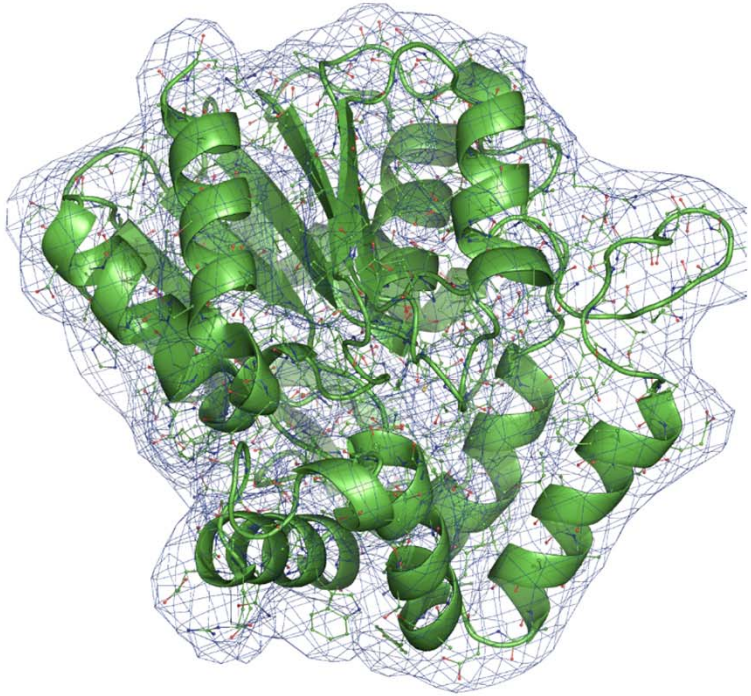
$$T_{BOND} = \Sigma_{\text{all bonded pairs}} w(d_{ideal} - d_{model})^2$$

Restraints for low-resolution refinement

- Low resolution map is not sufficient to maintain secondary

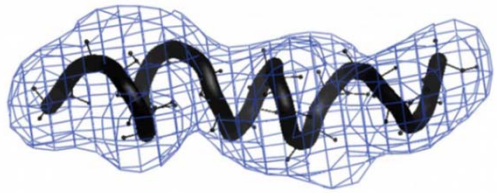






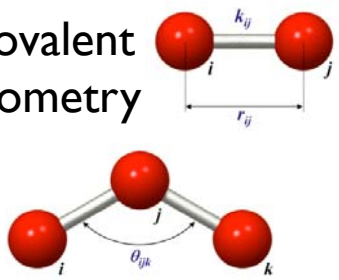
Special extra restraints for low-resolution refinement

- Example: refinement of a perfect α-helix into low-res map
  - Using standard restraints on covalent geometry is insufficient
    - Model geometry deteriorates as result of refinement

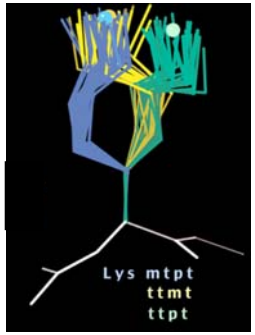


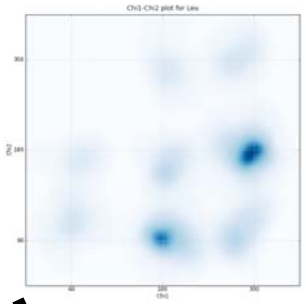
Sources of extra information to use as restraints

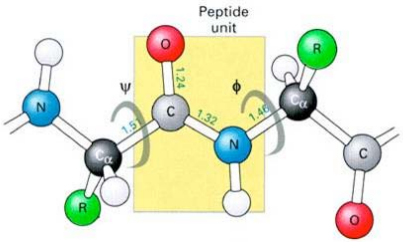
Covalent geometry



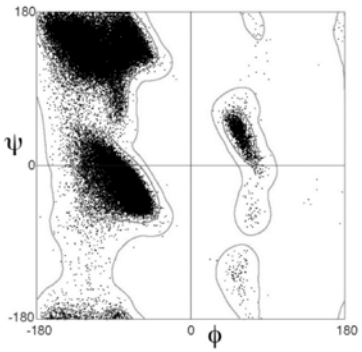
Sidechain distributions

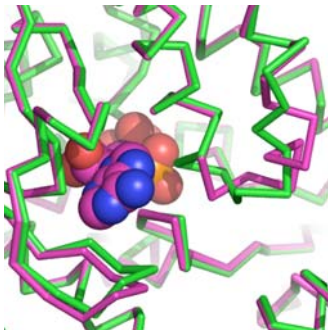




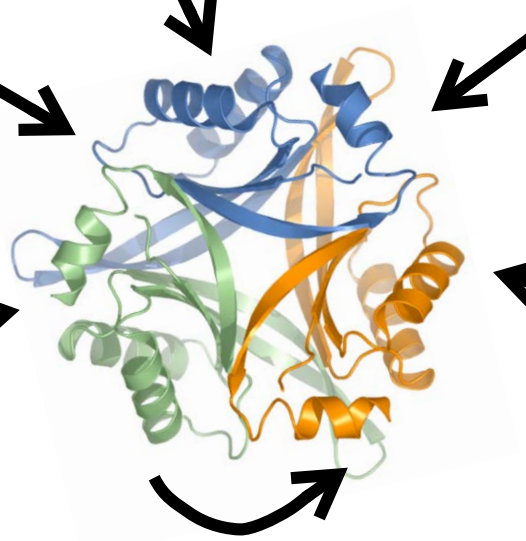


Mainchain distributions

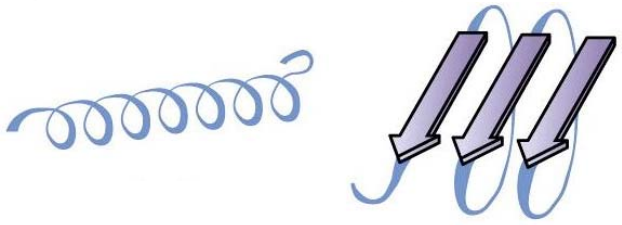




Related structures



Internal symmetry

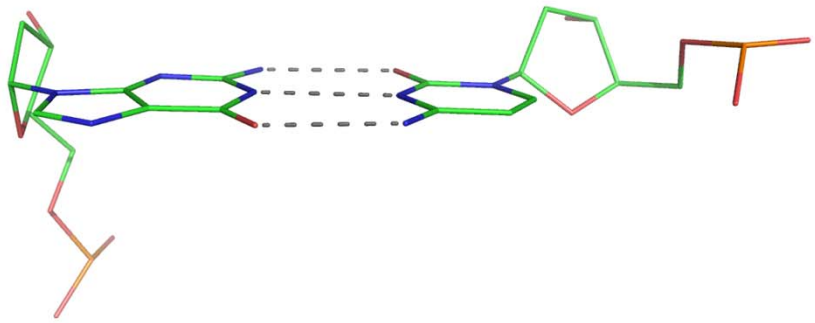


Secondary structure

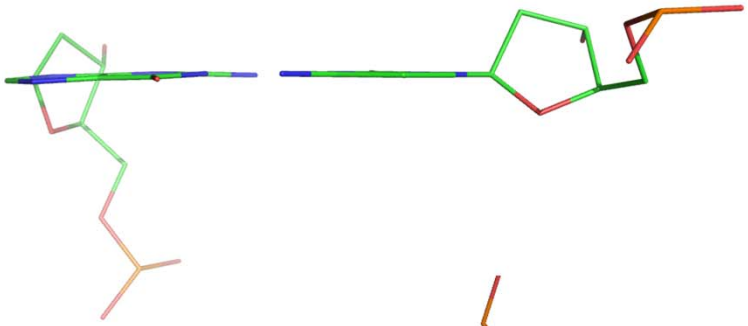
$$T_{RESTRAINTS} = T_{BOND} + T_{ANGLE} + \ldots + T_{NCS} + T_{RAMACHANDRAN} + T_{REFERENCE} + \ldots$$

DNA/RNA specific restraints

1. Hydrogen bonds between base pairs:
   - Bond length restraints
   - Bond angles restraints



1. Angle restraints between base planes



1. Parallelity of stacking nucleobases:
   - Parallelity restraints

Structure refinement is optimization problem

Model parameters

Parameters that describe crystal and its content

- Atomic model (coordinates, B-factors, occupancies)
- Non-atomic model (bulk-solvent, anisotropy, twining)

Scoring function (refinement target, optimization goal)

A function that relates model parameters and experimental data

- Least-squares (LS), Maximum-Likelihood (ML), R-factor, …

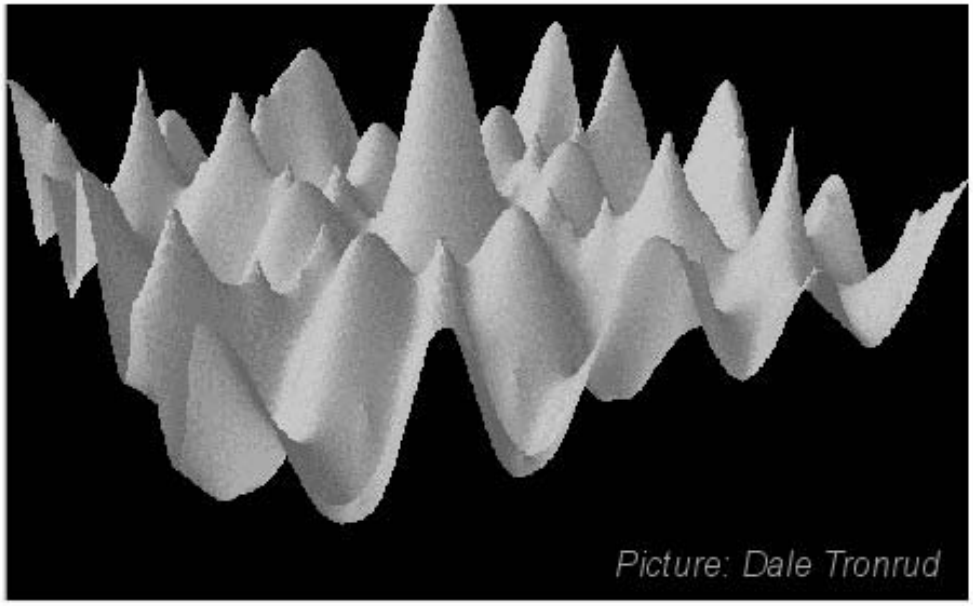
Optimization method

A tool that varies model parameters to optimize refinement target

- Gradient-based minimization
- Simulated Annealing (SA)
- Systematic (grid) searches
- Changing parameters by hand (using graphics, like Coot)

Complexity of refinement target

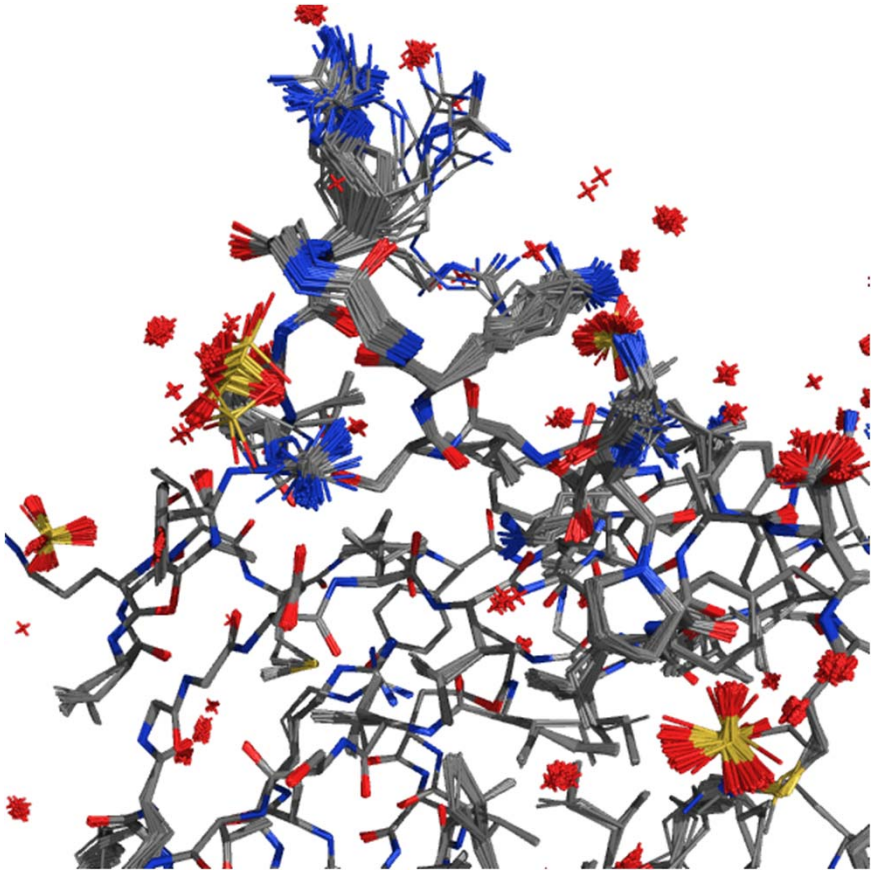
Refinement target profile is very complex: many local minima

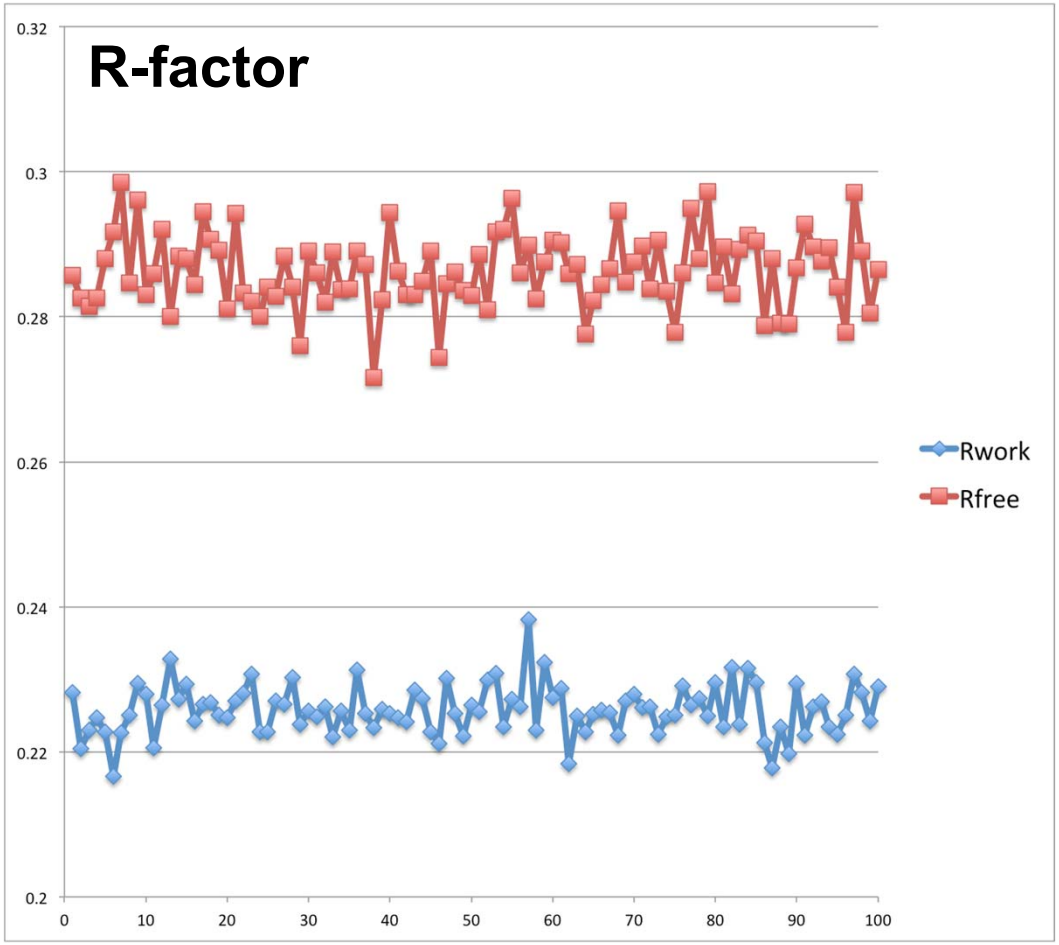


- Refinement goal is to reach the global minimum
  - Rarely achieved -> limited convergence radius
    - Refinement result strongly depends on starting point
    - Various optimization methods are in use

Refinement convergence

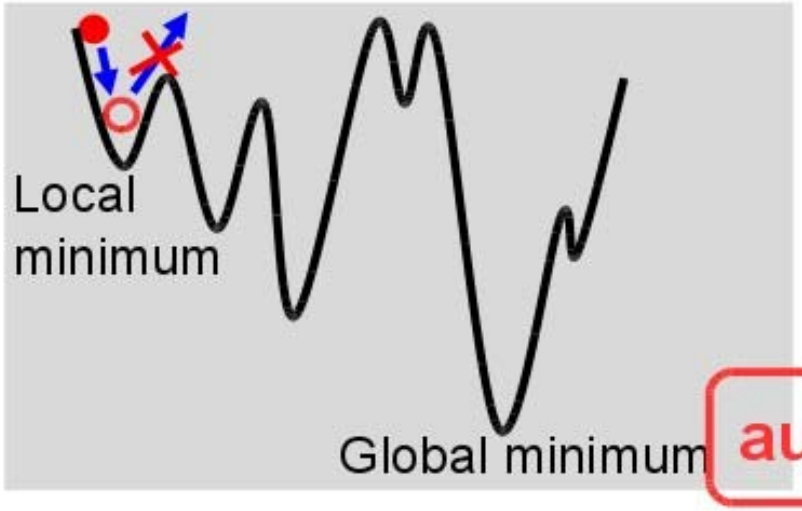
Result of many identical refinement runs starting with slightly perturbed model



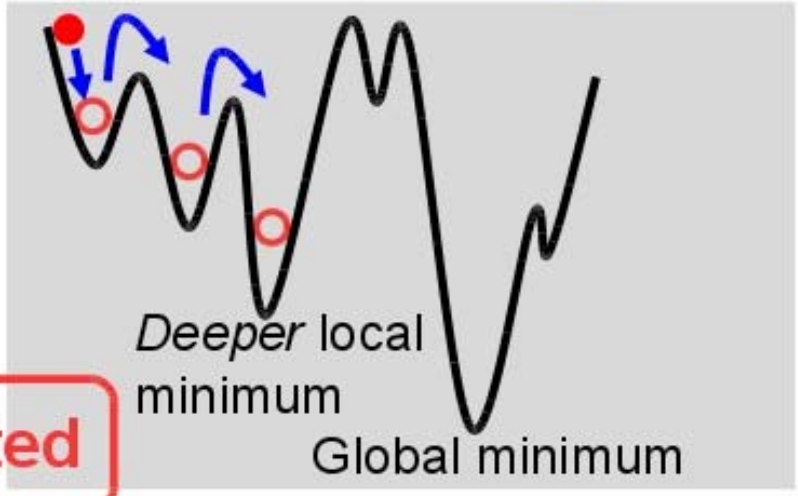


Refinement target optimization methods

- **Gradient-driven minimization**

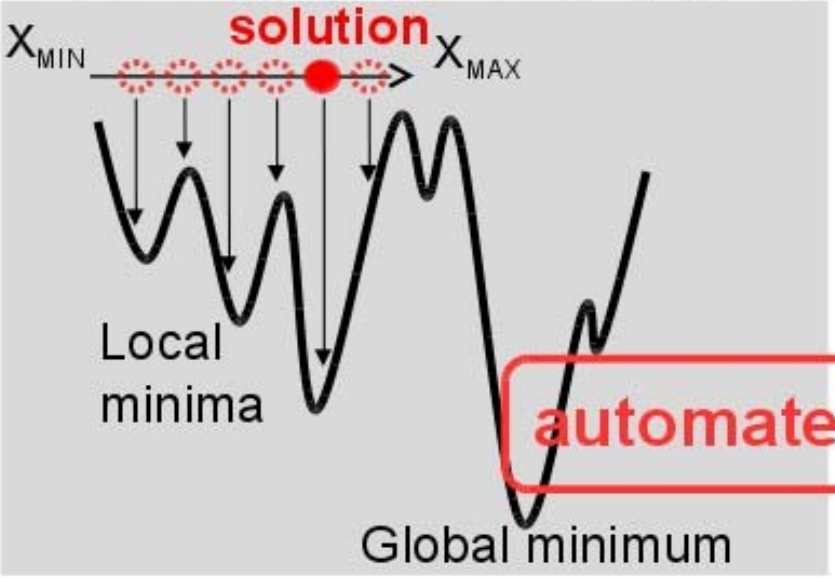


- **Simulated annealing**



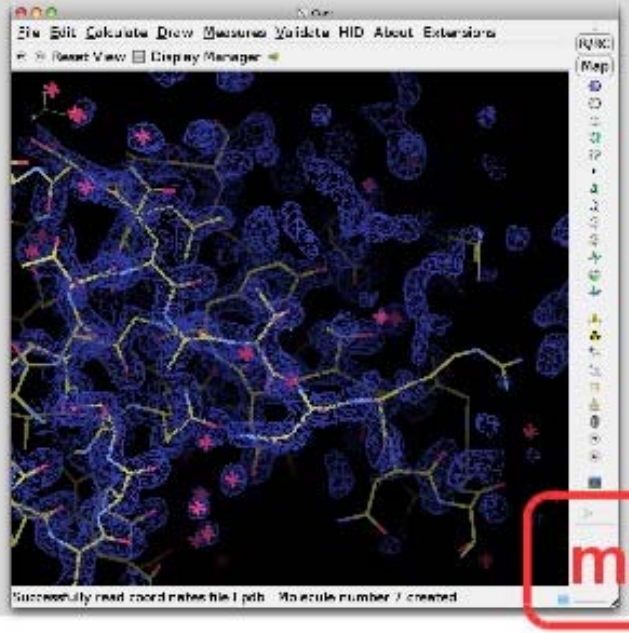
automated

- **Grid search** (Sample parameter space within known range [X_{MIN}, X_{MAX}])



automated

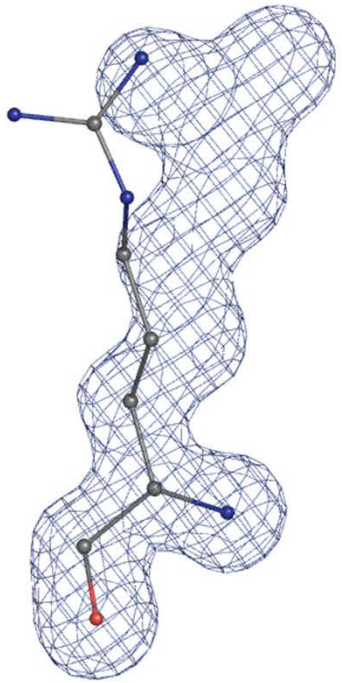
- **Hands & eyes** (via Coot)



manual

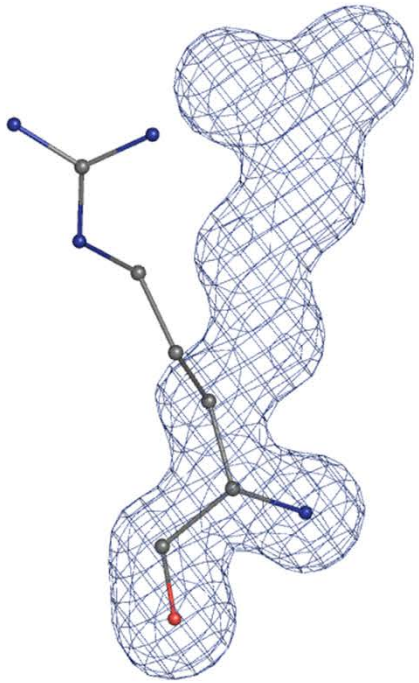
Refinement convergence

Minimization



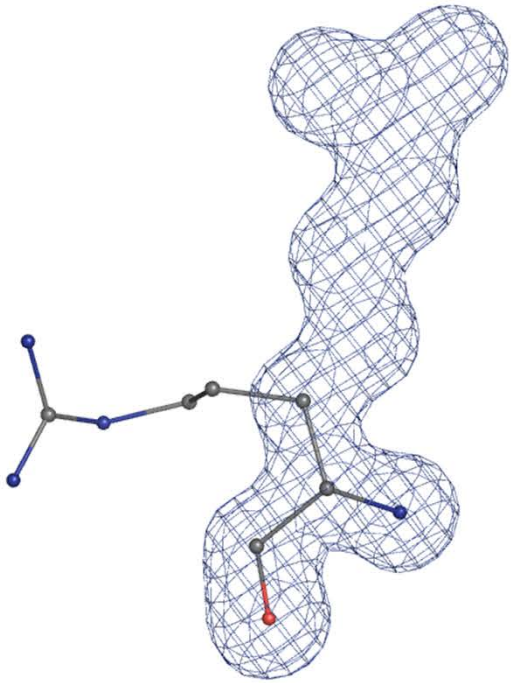
Minimization or SA can fix it

Simulated Annealing



Beyond convergence radius of minimization

Real-space grid search



Beyond convergence radius of minimization and SA

Refinement convergence

Quality Figures: the R–value

Refinement programs target at minimisation of the R–value = the agreement between measured amplitudes (Fobs(hkl)) and calculated from the model (Fcalc(hkl)).

$$R = \frac{\sum_{hkl} (|F_{obs}| - |F_{calc}|)}{\sum_{hkl} (|F_{obs}|)}$$

|Fobs| are represented by the reflection data (observations), |Fcalc| are calculated from (x,y,z) and B-values of the atoms of the model.

- For small molecules, R–values between 2% and 5% are normal,
- For macromolecules, the range is approximately 20%–30%.
- Rule of thumb: R–value about 1/10 of the resolution (a 2.5Å structure should have an R–value of 25%).

Refinement and Overfitting

The amplitudes lack some information (their phase) and are far from ideal
The difference can be nearly arbitrarily reduced by adding more and more atoms or allowing positions that do not make much sense → overfitting of data.
A measure to reduce overfitting is the Rfree–value:

- About 5%–10% of the reflections are excluded from minimisation of the R–value.
- They remain unconsidered and are like an “independent judge”:
- After refinement, the R free value is calculated with the excluded reflections.
- The two values must not differ too much.

Reciprocal-space structure refinement programs

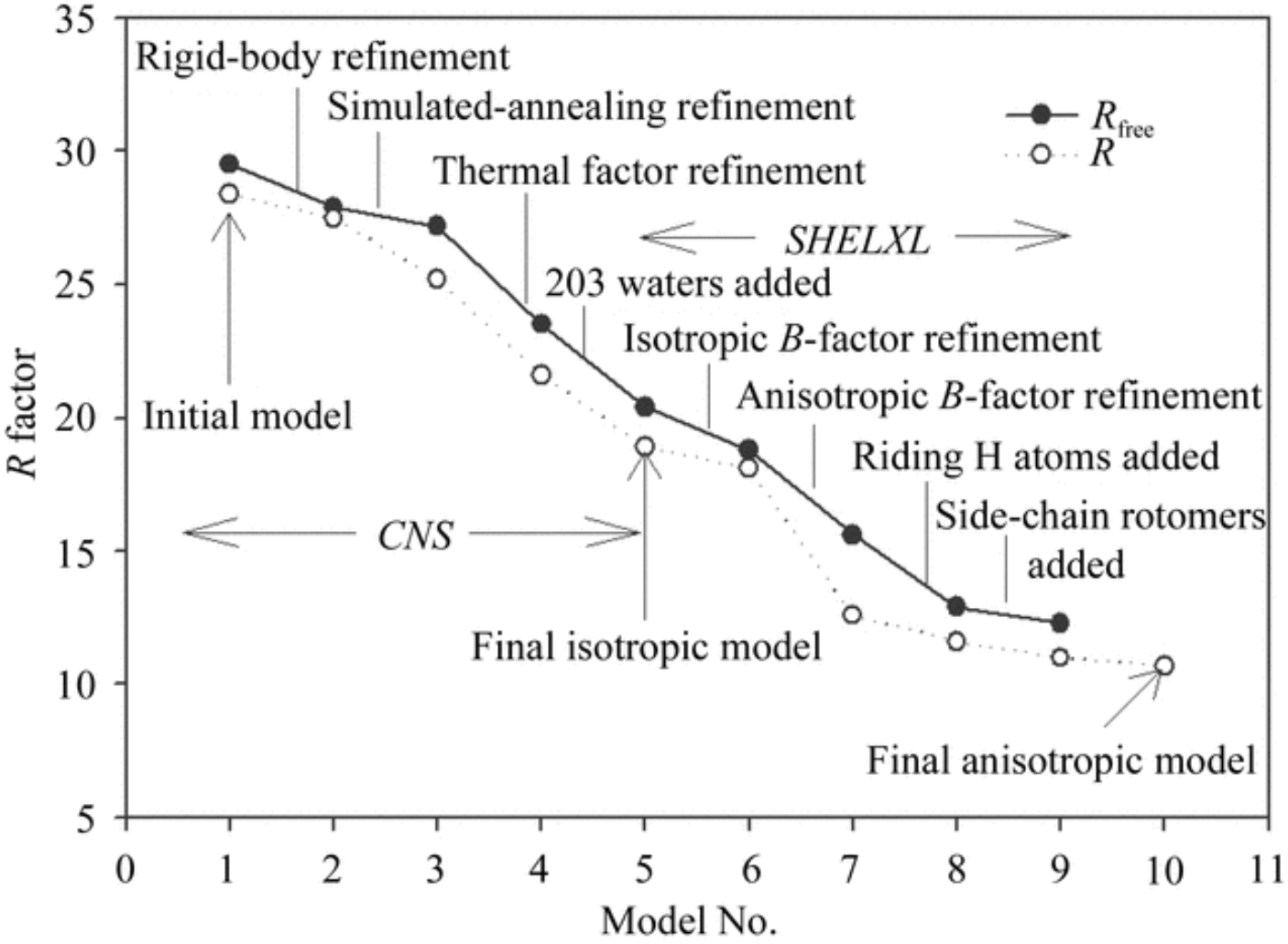
Available refinement programs

- SHELX (1970)
- CNS/Xplor (1987)
- BUSTER-TNT (TNT – 1987, BUSTER-TNT - around 2000)
- REFMAC (1997)
- Phenix : phenix.refine (2005)

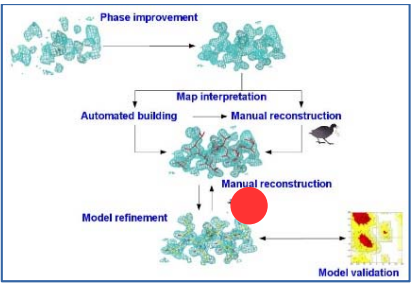
phenix.refine: automation is the focus

- **Typical structure refinement work-flow (past)**

Acta Cryst. (2002). D58, 2009-2017, Yousef et al.



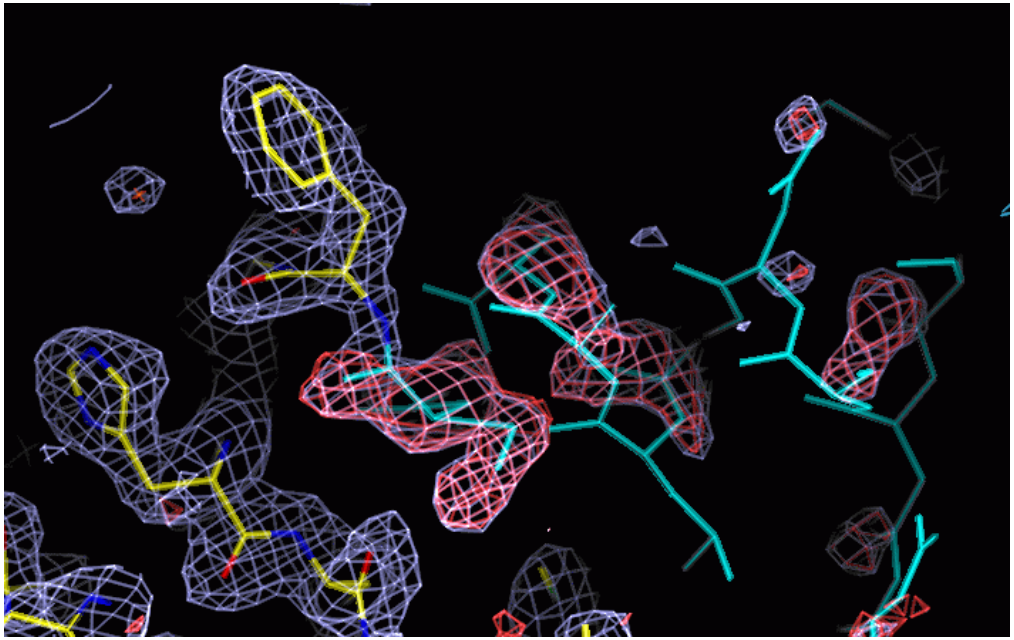
- **Modern software should do all these steps automatically**



Model reconstruction

manual

Dans la pratique 2 types de cartes



Représentée par des surfaces d'isodensité

1. Carte de densité du modèle

$(3F_{obs} - 2F_{calc}), \varphi_{calc}$
$(2F_{obs} - F_{calc}), \varphi_{calc}$

Les atomes manquants dans le modèle apparaissent grâce à l'information apportée par les amplitudes F_{obs}

2. « Carte différence »

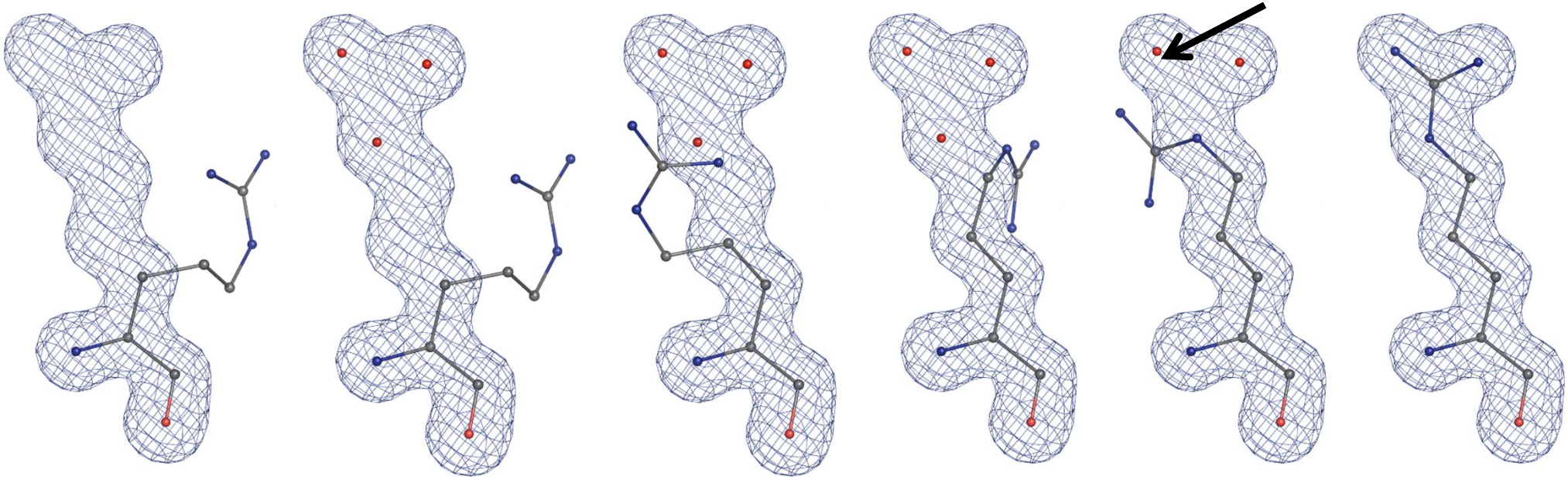
$(F_{obs} - F_{calc}), \varphi_{calc}$

Mise en valeur des différences entre le modèle et les données de diffraction

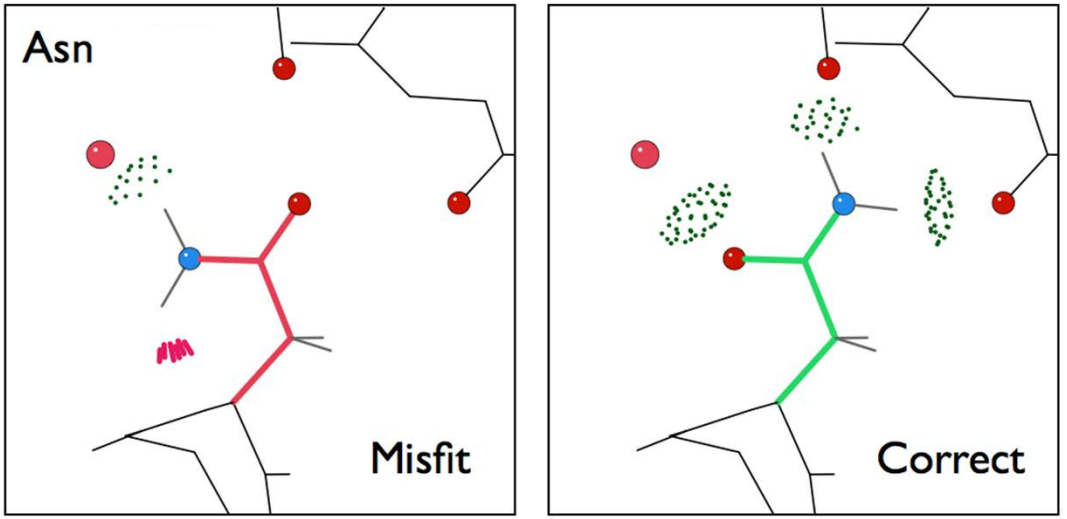
Local real-space refinement / Asn-Gln-His correction

Fix rotamer first…

… remove water if necessary

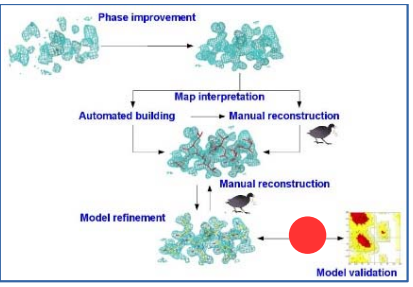


- **Move to density**
- **Correct flipped N/Q/H residues**

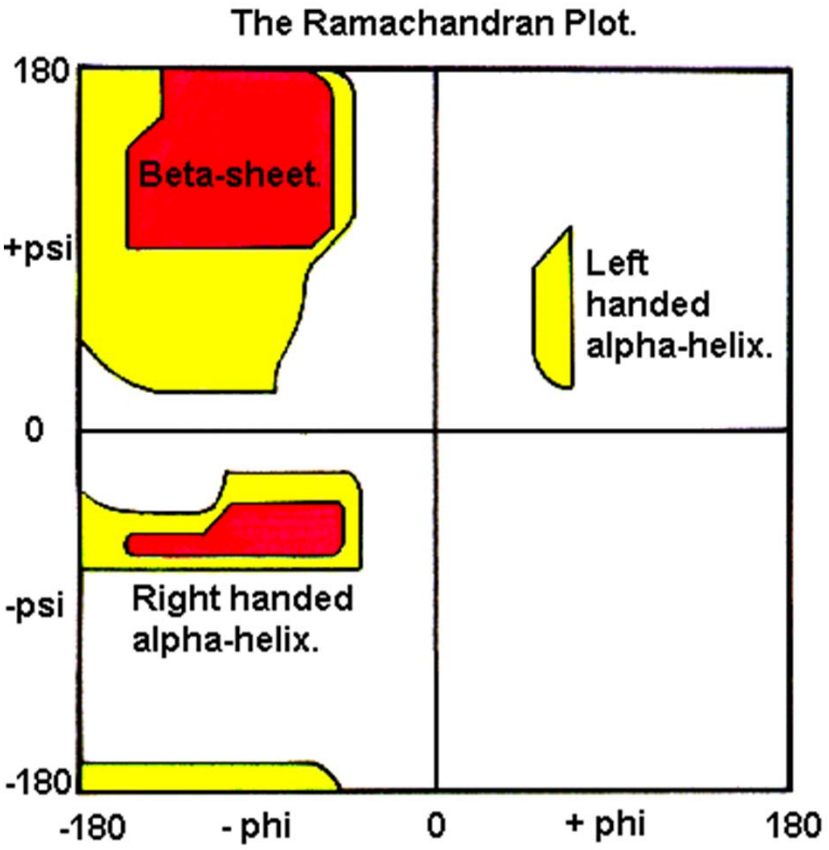


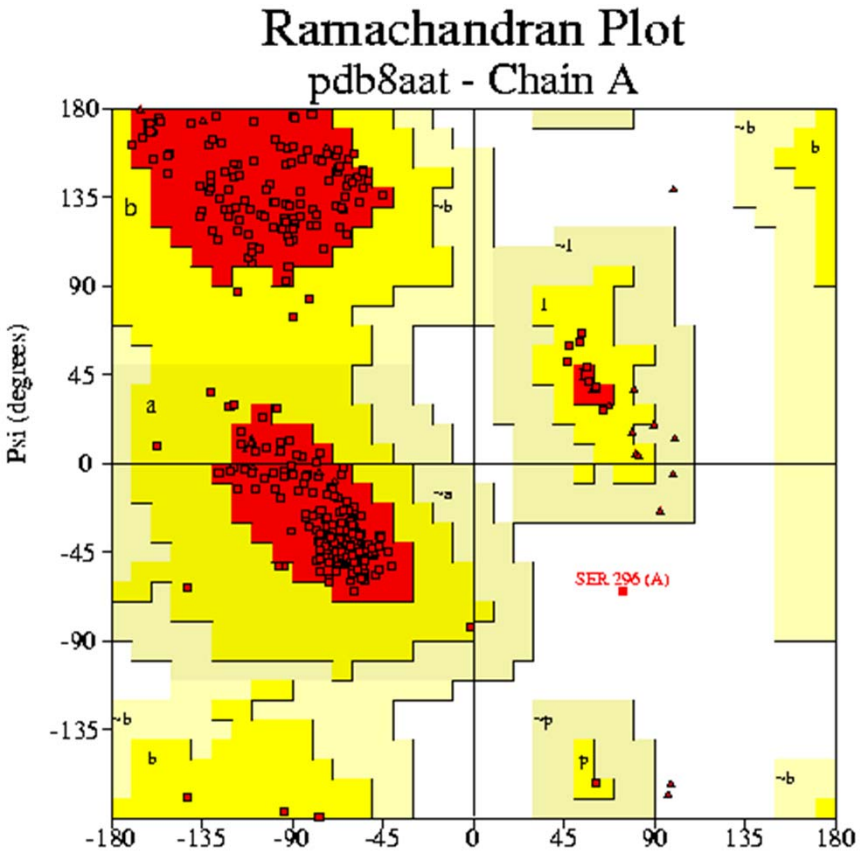
manual

automated



Validation





MolProbity (http://molprobity.biochem.duke.edu/)