Crystallographic refinement

Roberto A. Steiner
roberto.steiner@kcl.ac.uk

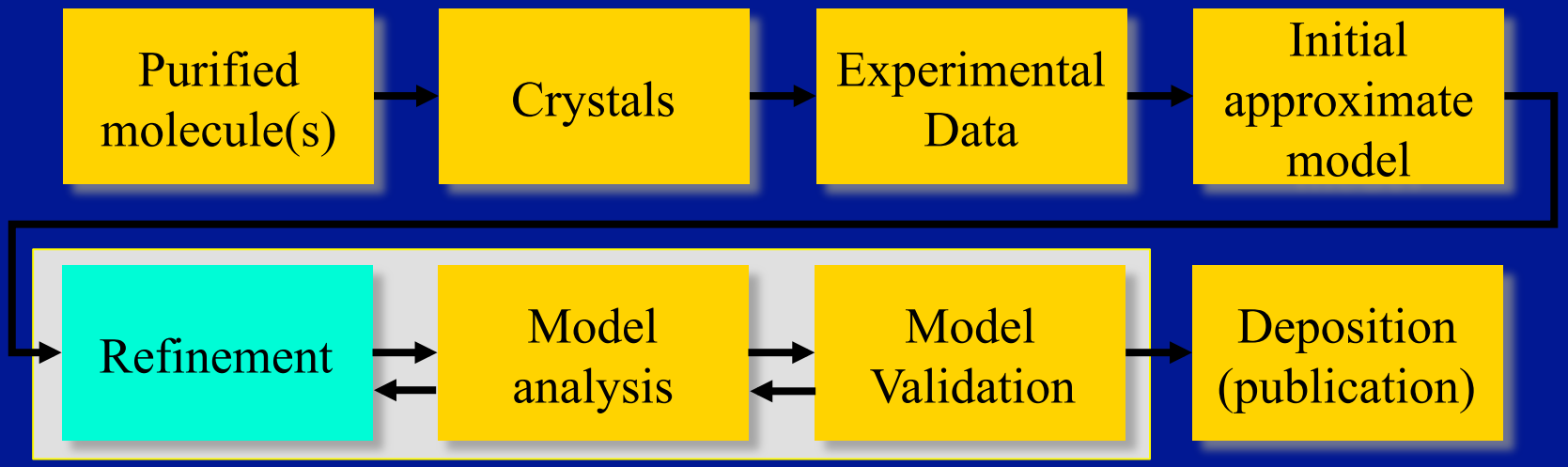
with many slides contributed by Pavel Afonine

Crystallography workshop - Trieste 23-27 April 2012

KING'S College LONDON
University of London

Key aspects of refinement

Crystallographic refinement is the process by which an initial structural model is modified to produce an updated model that is more consistent with the experimental data and known (bio)chemistry.

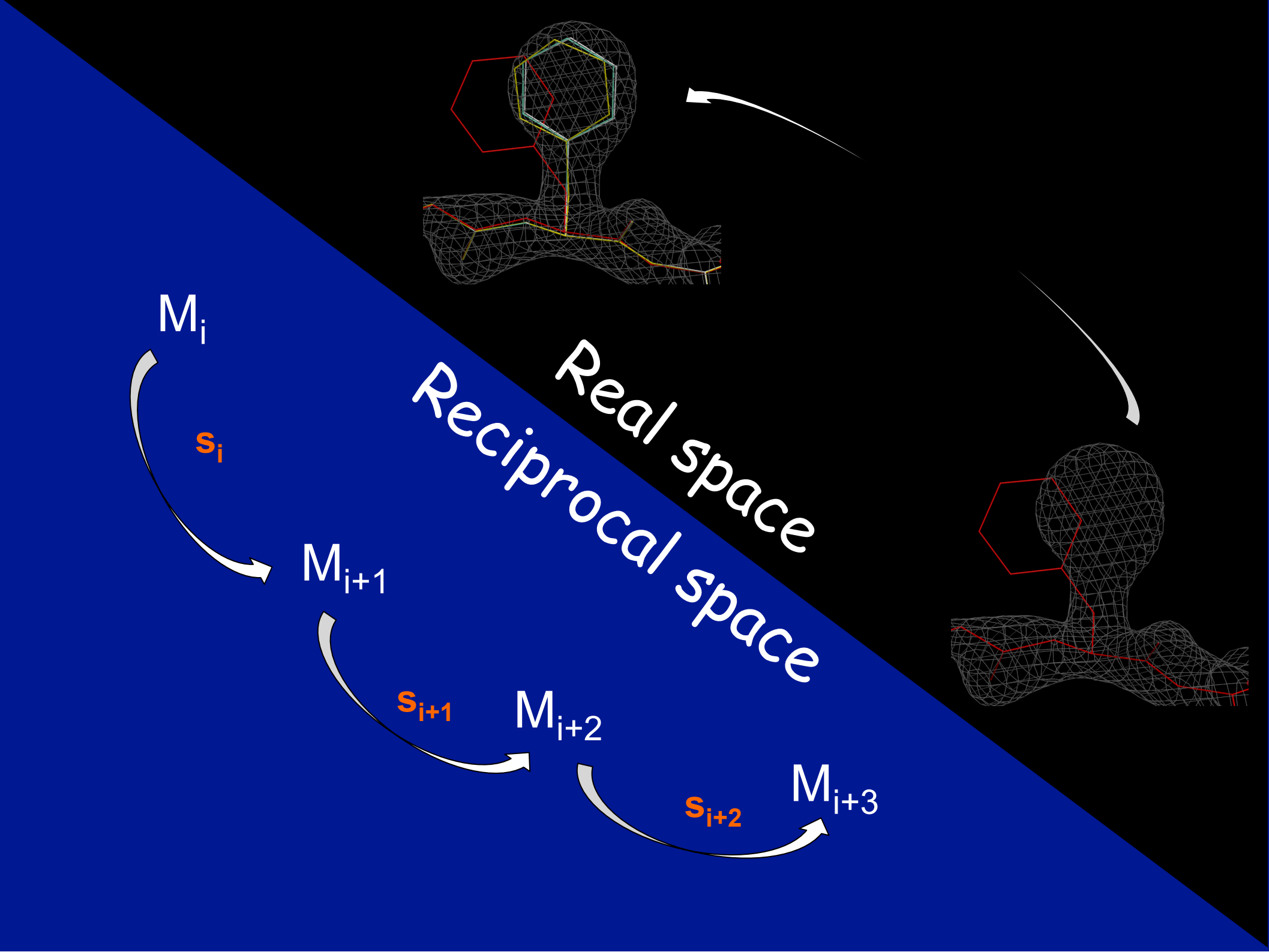


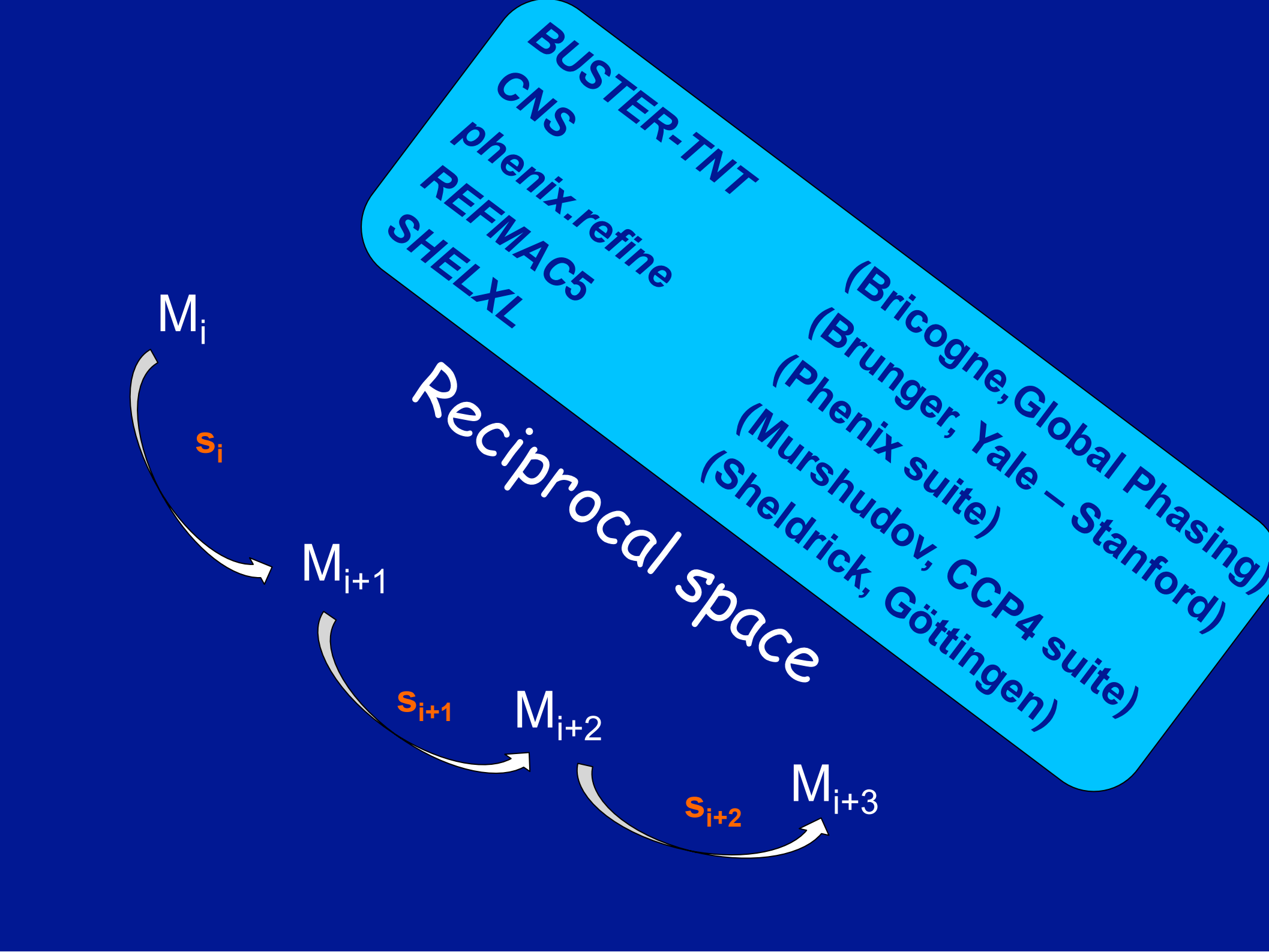
Updated model…what does that mean?

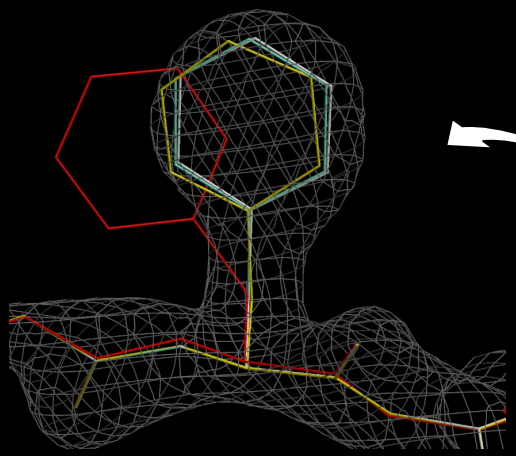
Atoms are added/removed
Atoms are moved
ADPs change their values
Occupancies can also change

Refinement is an iterative process which is generally terminated by the user.

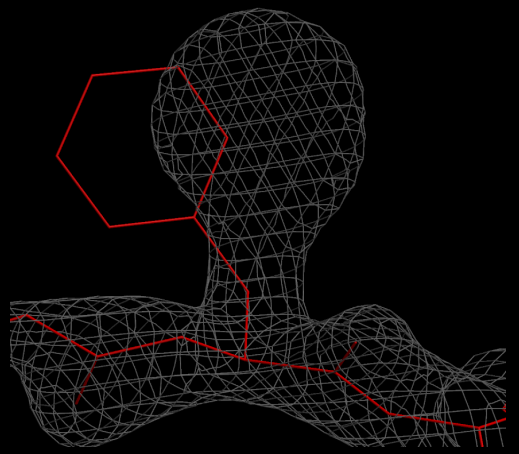
Phil Evans introduced the concept of refinement *at tedium* (until one is too bored to continue..)


Real space
Reciprocal space
M_i
s_i
M_{i+1}
s_{i+1}
M_{i+2}
s_{i+2}
M_{i+3}


BUSTER-TNT (Bricogne, Global Phasing)
CNS (Brunger, Yale – Stanford)
phenix.refine (Phenix suite)
REFMAC5 (Murshudov, CCP4 suite)
SHELXL (Sheldrick, Göttingen)
Reciprocal space
M_i
s_i
M_{i+1}
s_{i+1}
M_{i+2}
s_{i+2}
M_{i+3}



Real space



Coot
O
XtalView

(Emsley, Oxford)
(Jones, Uppsala)
(McRae, San Diego)

ISSN 0907-4449

Volume 68

Part 4

April 2012

Model building, refinement and validation

Proceedings of the CCP4 study weekend

Guest Editors

Roberto A. Steiner
Bernhard Rupp
Charles Ballard

Organisers

Shirley Miller
Damian Jones
Laura Johnston
Wendy Cotterill

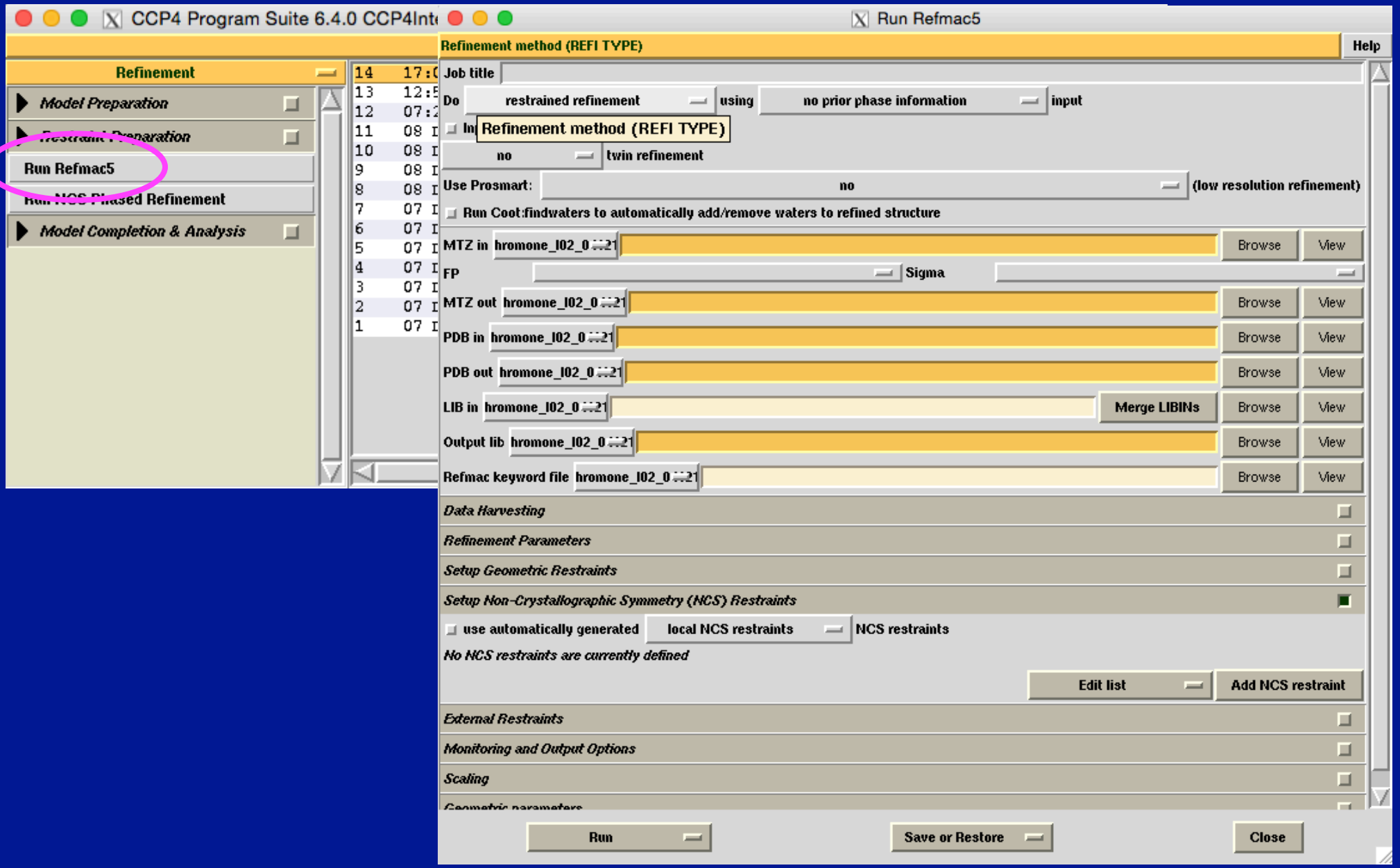
Acta Crystallographica Section D

Biological Crystallography

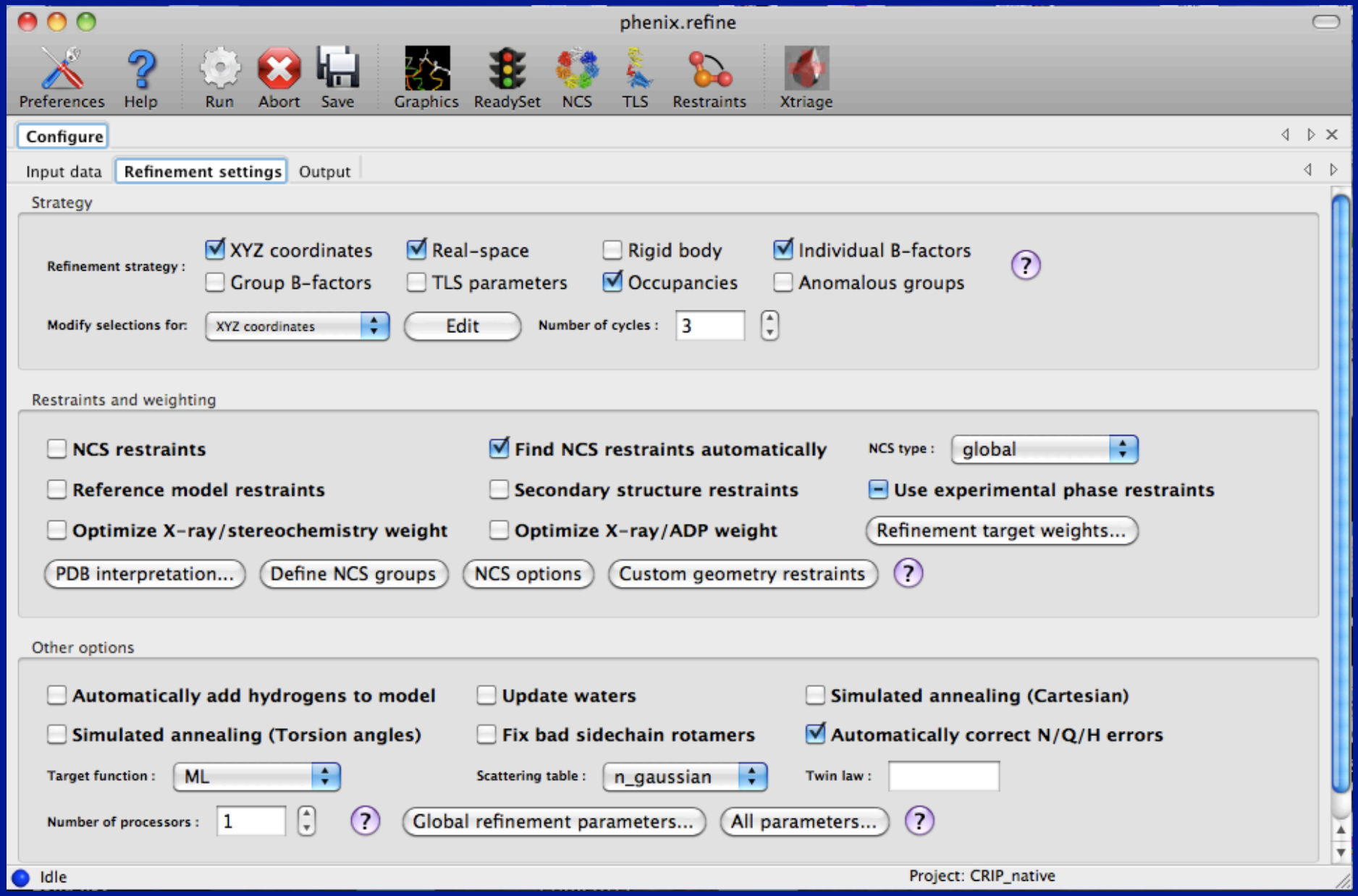
Editors: E. N. Baker and Z. Dauter

journals.iucr.org
International Union of Crystallography
Wiley-Blackwell

REFMAC5 - CCP4*i*

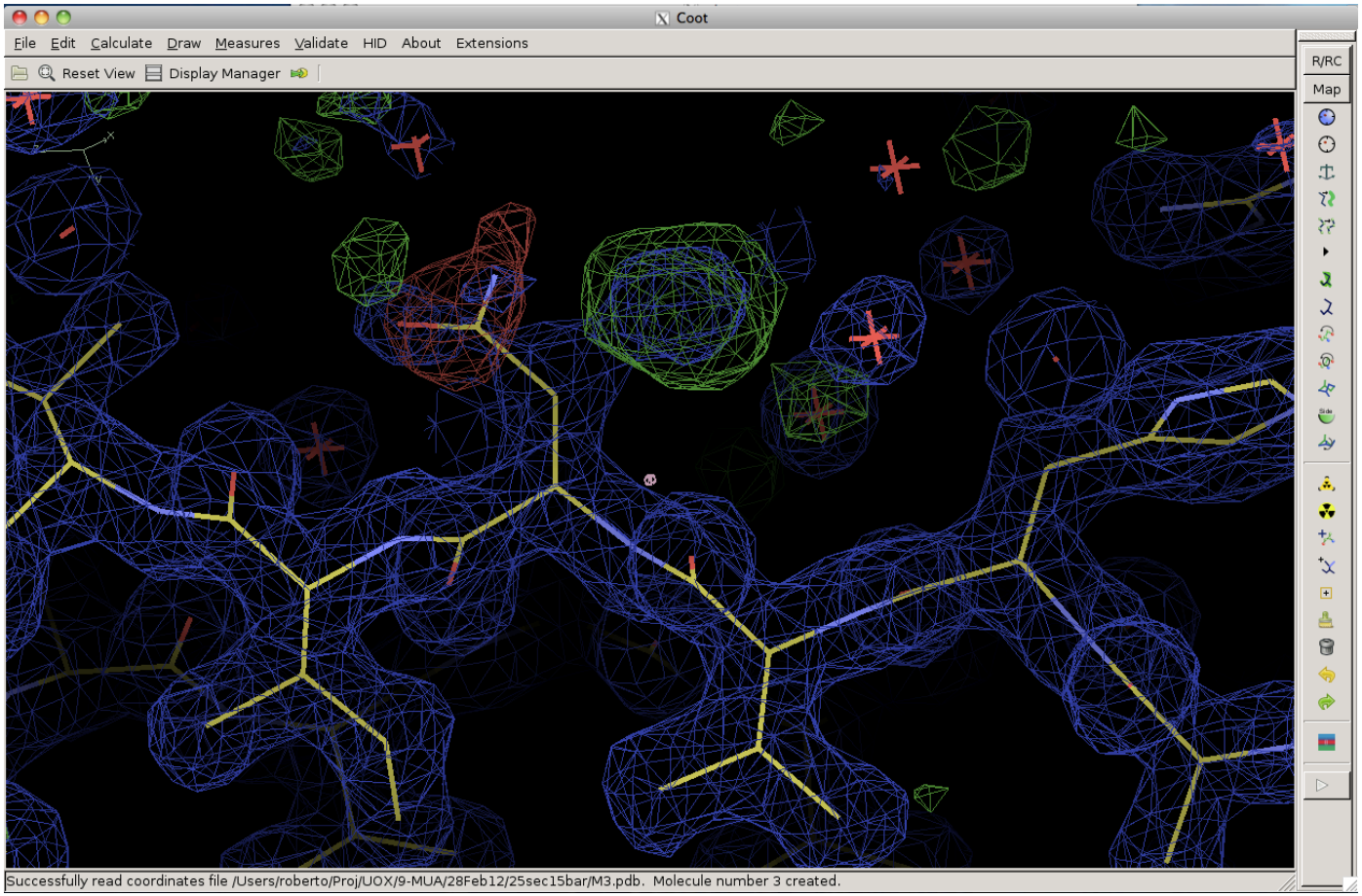


phenix.refine GUI



Real space refinement

$$\rho(xyz) = \frac{1}{V}\sum_{\text{hkl}} (2m|F_{\text{obs}}(hkl)| - D|F_{\text{calc}}(hkl)|)\exp[-2\pi i(hx+ky+lz) + i\varphi_{\text{calc}}(hkl)]$$



$$\rho(xyz) = \frac{1}{V}\sum_{\text{hkl}} (m|F_{\text{obs}}(hkl)| - D|F_{\text{calc}}(hkl)|)\exp[-2\pi i(hx+ky+lz) + i\varphi_{\text{calc}}(hkl)]$$

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

*REFMAC*5 for the refinement of macromolecular crystal structures

Garib N. Murshudov,[a]* Pavol Skubák,[b] Andrey A. Lebedev,[a] Navraj S. Pannu,[b] Roberto A. Steiner,[c] Robert A. Nicholls,[a] Martyn D. Winn,[d] Fei Long[a] and Alexei A. Vagin[a]

[a]Structural Biology Laboratory, Department of Chemistry, University of York, Heslington, York YO10 5YW, England, [b]Biophysical Structural Chemistry, Leiden University, PO Box 9502, 2300 RA Leiden, The Netherlands, [c]Randall Division of Cell and Molecular Biophysics, New Hunt's House, King's College London, London, England, and [d]STFC Daresbury Laboratory, Warrington WA4 4AD, England

Correspondence e-mail: garib@ysbl.york.ac.uk

This paper describes various components of the macromolecular crystallographic refinement program *REFMAC5*, which is distributed as part of the *CCP*4 suite. *REFMAC5* utilizes different likelihood functions depending on the diffraction data employed (amplitudes or intensities), the presence of twinning and the availability of SAD/SIRAS experimental diffraction data. To ensure chemical and structural integrity of the refined model, *REFMAC5* offers several classes of restraints and choices of model parameterization. Reliable models at resolutions at least as low as 4 Å can be achieved thanks to low-resolution refinement tools such as secondary-structure restraints, restraints to known homologous structures, automatic global and local NCS restraints, 'jelly-body' restraints and the use of novel long-range restraints on atomic displacement parameters (ADPs) based on the Kullback–Leibler divergence. *REFMAC5* additionally offers TLS parameterization and, when high-resolution data are available, fast refinement of anisotropic ADPs. Refinement in the presence of twinning is performed in a fully automated fashion. *REFMAC5* is a flexible and highly optimized refinement package that is ideally suited for refinement across the entire resolution spectrum encountered in macromolecular crystallography.

Received 14 July 2010
Accepted 10 January 2011

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Towards automated crystallographic structure refinement with *phenix.refine*

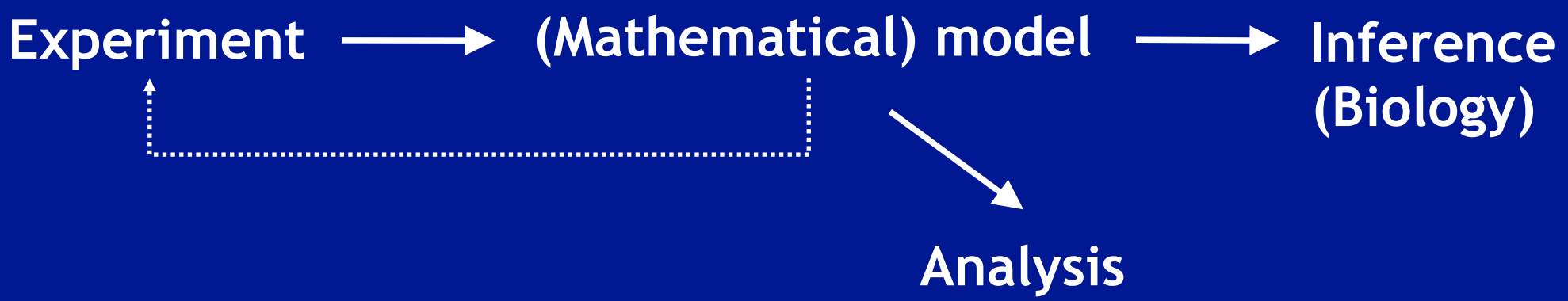
Pavel V. Afonine,[a]* Ralf W. Grosse-Kunstleve,[a] Nathaniel Echols,[a] Jeffrey J. Headd,[a] Nigel W. Moriarty,[a] Marat Mustyakimov,[b] Thomas C. Terwilliger,[b] Alexandre Urzhumtsev,[c,d] Peter H. Zwart[a] and Paul D. Adams[a,e]

[a]Lawrence Berkeley National Laboratory, One Cyclotron Road, MS64R0121, Berkeley, CA 94720, USA, [b]Los Alamos National Laboratory, M888, Los Alamos, NM 87545, USA, [c]IGBMC, CNRS–INSERM–UdS, 1 Rue Laurent Fries, BP 10142, 67404 Illkirch, France, [d]Département de Physique, Faculté des Sciences et des Technologies, Université Henri Poincaré, Nancy 1, BP 239, 54506 Vandoeuvre-lès-Nancy, France, and [e]Department of Bioengineering, University of California Berkeley, Berkeley, CA 94720, USA

phenix.refine is a program within the *PHENIX* package that supports crystallographic structure refinement against experimental data with a wide range of upper resolution limits using a large repertoire of model parameterizations. It has several automation features and is also highly flexible. Several hundred parameters enable extensive customizations for complex use cases. Multiple user-defined refinement strategies can be applied to specific parts of the model in a single refinement run. An intuitive graphical user interface is available to guide novice users and to assist advanced users in managing refinement projects. X-ray or neutron diffraction data can be used separately or jointly in refinement. *phenix.refine* is tightly integrated into the *PHENIX* suite, where it serves as a critical component in automated model building, final structure refinement, structure validation and deposition to the wwPDB. This paper presents an overview of the major *phenix.refine* features, with extensive literature references for readers interested in more detailed discussions of the methods.

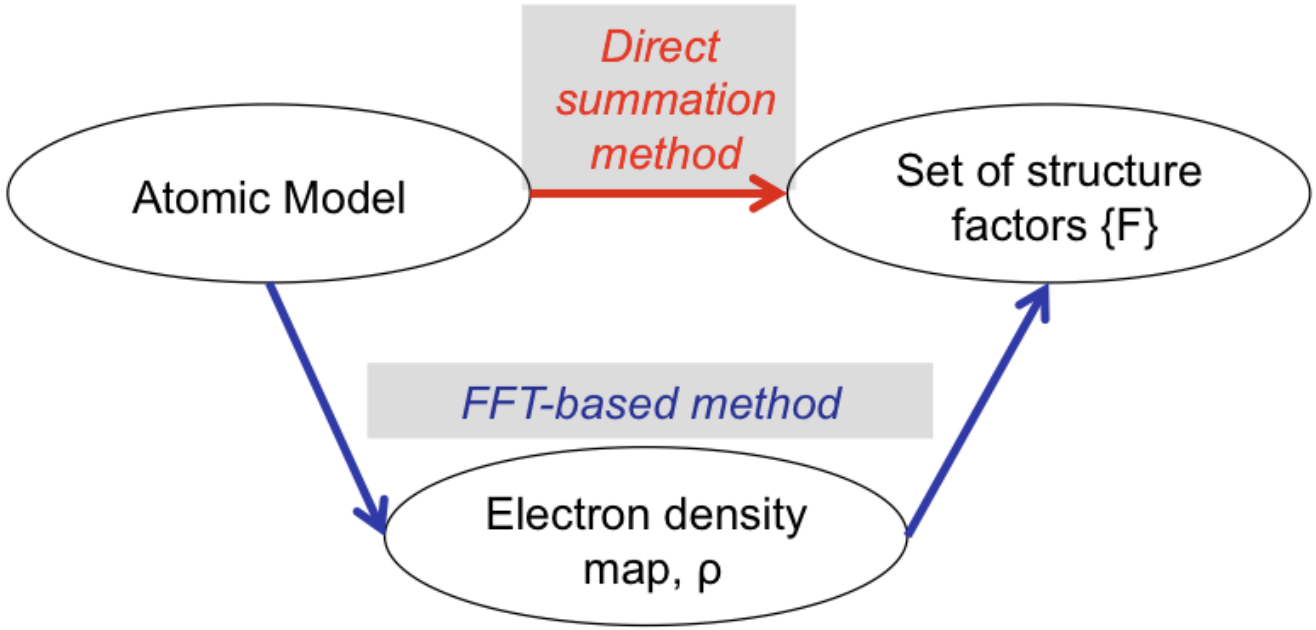
Received 27 September 2011
Accepted 11 January 2012

Model fitting



Note about structure factors calculation

- Two ways of computing structure factor from atomic model



- For macromolecules the *FFT-based method* is much faster than the direct summation method
- Most of macromolecular refinement programs use *FFT-based method*
- *FFT-based method* is based on a number of approximations and therefore it is less accurate than direct summation; however, inaccuracies introduced by these approximations are negligible in most of practical cases

Note about structure factors calculation

- **Structure factor formula (direct summation method)**

$$\mathbf{F}(h,k,l) = \sum_{n=1}^{Natoms} q_n f_n(s) \exp\left(-\frac{B_n s^2}{4}\right) \exp\left(2i\pi \mathbf{r}_n \mathbf{s}\right)$$

$$f(s) = \sum_{k=1}^{P} a_k \exp\left(-\frac{b_k s^2}{4}\right)$$ Gaussian approximation for atomic form-factor

q_n, B_n and $\mathbf{r}_n=(x_n,y_n,z_n)$ – atomic occupancy, isotropic B-factor and coordinates

P~5 (depends on approximation), a_k and b_k – parameters of approximation specific to atom type

$s^2 = \mathbf{h}^t\mathbf{G}^*\mathbf{h}$, h – column-vector of Miller indices, $\mathbf{G}^*$ - reciprocal-space metric tensor

✓ Calculation time ~ number of reflections * number of atoms

✓ Formula above yields exact values for F

Note about structure factors calculation

- **Structure factor formula (FFT-based summation)**

Fundamental formula $\mathbf{F}(h,k,l) = \int_{Vcell} \rho(\mathbf{r}) \exp\{2\pi i \mathbf{sr}\} dV$

Approximate way to compute this integral numerically:

$$\mathbf{F}(h,k,l) = \frac{V_{cell}}{N_X N_Y N_Z} \sum_{j_X}^{N_X-1} \sum_{j_Y}^{N_Y-1} \sum_{j_Z}^{N_Z-1} \rho(j_X, j_Y, j_Z) \exp\{2\pi i (hj_X + kj_Y + lj_Z)\}$$

which is discrete Fourier transform of electron density:

$$\rho(r) = \sum_{n=1}^{Natoms} q_n \sum_{k=1}^{P} a_k \left(\frac{4\pi}{b_k + B_n}\right)^{3/2} \exp\left(-\frac{4\pi^2 |\mathbf{r} - \mathbf{r}_n|^2}{b_k + B_n}\right)$$

sampled at grid N_X, N_Y, N_Z in a sphere of radius R (~2 Å) around each atom.

- ✓ Source of inaccuracy: replacement of continues integral with discrete summation and truncation of atomic density within a sphere R.

- ✓ Calculation time ~ density calculation + FFT ~ $K_{grid}*(V_{atom}/V_{crystal}) + K_{grid}*\ln(K_{grid})$, where $K_{grid} = N_X N_Y N_Z$

research papers

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Introduction to macromolecular refinement

Dale. E. Tronrud

Howard Hughes Medical Institute and Institute of Molecular Biology, University of Oregon, Eugene, OR 97403, USA

Correspondence e-mail: dale@uoxray.uoregon.edu

The process of refinement is such a large problem in function minimization that even the computers of today cannot perform the calculations to properly fit X-ray diffraction data. Each of the refinement packages currently under development reduces the difficulty of this problem by utilizing a unique combination of targets, assumptions and optimization methods. This review summarizes the basic methods and underlying assumptions in the commonly used refinement packages. This information can guide the selection of a refinement package that is best suited for a particular refinement project.

Received 5 April 2004
Accepted 21 September 2004

Key aspects of refinement

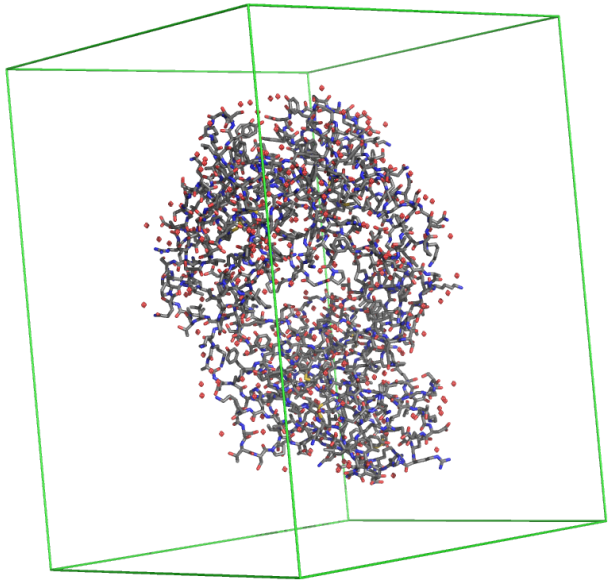
- Objective function
- Method of optimization
- Model parametrization
- Prior knowledge

Key aspects of refinement

- Objective function
- Method of optimization
- Model parametrization
- Prior knowledge

Model parameters or how we parameterize the crystal content

Crystal (unit cell)



Non-atomic model parameters (Bulk solvent, anisotropy, twinning)

- Macromolecular crystals contain ~20-80% of solvent (mostly disordered)
- Crystal-specific: description of anisotropy or twinning

Atomic model parameters

Non-atomic model (Bulk solvent and anisotropy)

- **Total model structure factor used in refinement, *R*-factor and map calculation:**

$$\mathbf{F}_{\mathrm{MODEL}} = k_{\mathrm{OVERALL}} e^{-\mathbf{s}\mathbf{U}_{\mathrm{CRYSTAL}}\mathbf{s}^t} \left(\mathbf{F}_{\mathrm{CALC\_ATOMS}} + k_{\mathrm{SOL}} e^{-\frac{B_{\mathrm{SOL}} s^2}{4}} \mathbf{F}_{\mathrm{MASK}} \right)$$

Anisotropy **Bulk-solvent contribution**

$\mathbf{U}_{\mathrm{CRYSTAL}}$ is 3x3 symmetric anisotropy scale matrix with 6 refinable parameters:

$$\begin{pmatrix} U_{11} & U_{12} & U_{13} \\ & U_{22} & U_{23} \\ & & U_{33} \end{pmatrix}$$

- symmetry constraints apply

Crystal System	Restrictions on U
Triclinic 1-2	None
Monoclinic 3-15	$U_{13}=U_{23}=0$ when $\beta=\alpha=90^\circ$ $U_{12}=U_{23}=0$ when $\gamma=\alpha=90^\circ$ $U_{12}=U_{13}=0$ when $\gamma=\beta=90^\circ$
Orthorhombic 16-74	$U_{12}=U_{13}=U_{23}=0$
Tetragonal 75-142	$U_{11}=U_{22}$ and $U_{12}=U_{13}=U_{23}=0$
Rhombohedral (trigonal) 143-167	$U_{11}=U_{22}=U_{33}$ and $U_{12}=U_{13}=U_{23}$
Hexagonal 168-194	$U_{11}=U_{22}$ and $U_{13}=U_{23}=0$
Cubic 195-230	$U_{11}=U_{22}=U_{33}$ and $U_{12}=U_{13}=U_{23}=0$ (=isotropic)

J. Mol. Biol. (1994) **243**, 100–115

Protein Hydration Observed by X-ray Diffraction

Solvation Properties of Penicillopepsin and Neuraminidase Crystal Structures

Jian-Sheng Jiang and Axel T. Brünger

The Howard Hughes Medical Institute and
Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520 U.S.A.

Solvation in macromolecular crystal structures was studied by analyzing X-ray diffraction data of two proteins, penicillopepsin and neuraminidase. The quality of several solvent models was assessed by complete cross-validation in order to prevent overfitting the diffraction data. Radial solvent distribution functions were computed from electron density maps using phases obtained from multiple isomorphous replacement and from the protein's atomic model combined with the best solvent model. Distribution functions were computed around hydrophilic and hydrophobic groups on the protein's surface. Averaging of the distribution functions was performed in order to reduce the influence of noise. The first solvation shell is characterized by a peak in the average distribution functions. At 1·8 Å resolution, polar groups show a sharp peak while non-polar groups show a broad one. The distinction between hydrophobic and hydrophilic solvation sites is lost when using lower resolution (2·8 Å) diffraction data. Higher-order solvation shells are not observed in the average distribution functions. We hope that site-specific radial distribution functions obtained from high-quality diffraction data will produce a picture of macromolecular solvation consistent with available experimental data and computational results.

Keywords: X-ray crystallography; solvation; refinement; cross-validation; radial distribution function

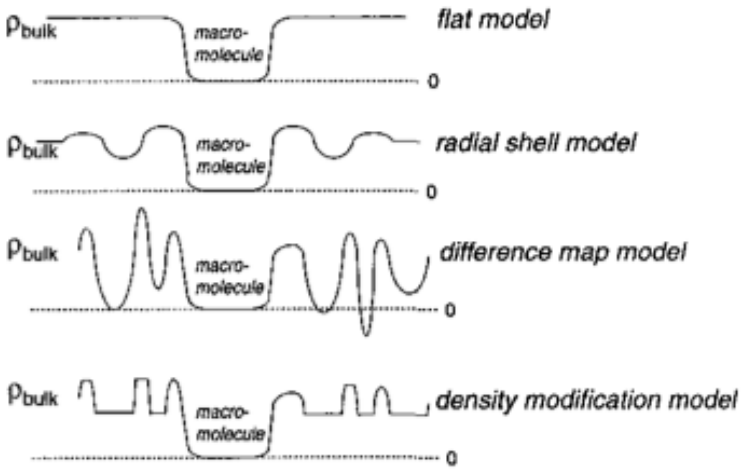


Figure 1. Schematic illustration for the 4 solvent models that were tested: flat model, radial shell model, difference map model and density modification model. The models are described in detail in the text.

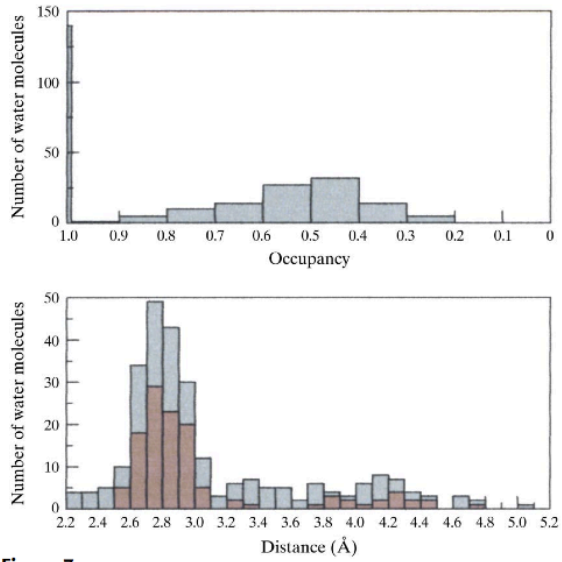
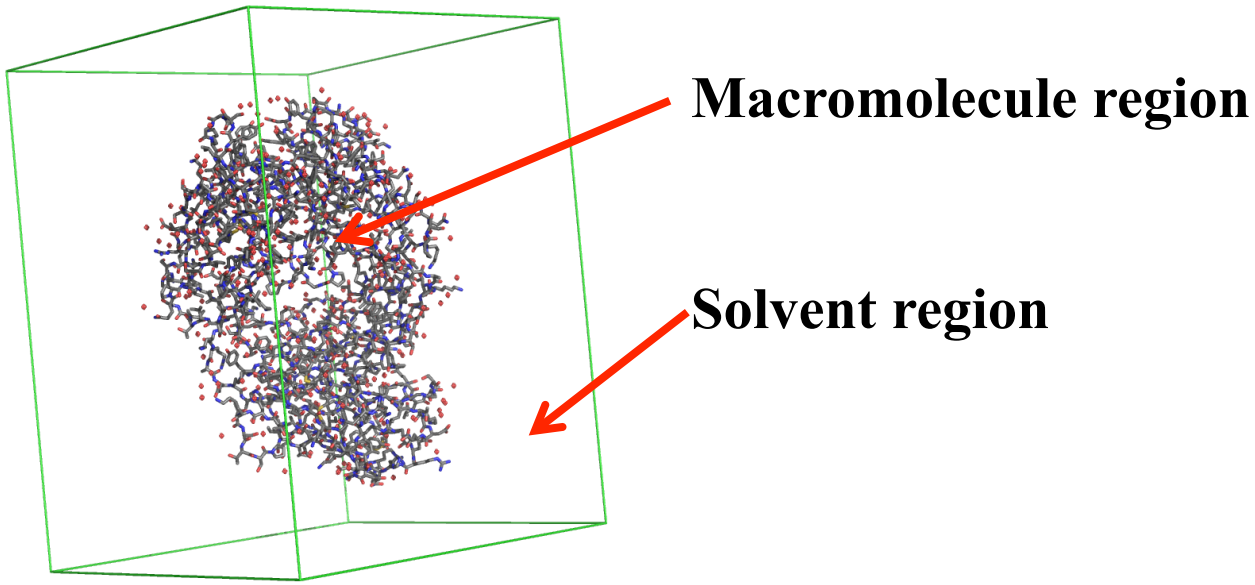


Figure 7
(*a*) Histogram of the occupancies of water molecules. (*b*) Histogram of the distribution of hydrogen-bond lengths for all water molecules (grey bars) and for ordered fully occupied waters. The distances from water molecules to the nearest N or O protein atom are plotted.

[Steiner et al., (2001)]

Model parameters – Bulk solvent and anisotropy



Flat Bulk Solvent model (currently best available and most popular model):

- Electron density in solvent regions is flat with some average value k_{SOL} (e/Å^3)
- Solvent mask: a binary function: 0 in *Macromolecular* and 1 in *Solvent* region
- $\mathbf{F}_{MASK}$ are structure factors calculated from Bulk solvent mask
- Contribution to the model structure factor:

$$\mathbf{F}_{BULK} = k_{SOL} e^{-\frac{B_{SOL}\, s^2}{4}} \mathbf{F}_{MASK}$$

  - B_{SOL} is another bulk solvent parameter defining "how deeply bulk solvent penetrates into a macromolecular region"

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Bulk-solvent and overall scaling revisited: faster calculations, improved results

P. V. Afonine,[a]* R. W. Grosse-Kunstleve,[a] P. D. Adams[a,b] and A. Urzhumtsev[c,d]

[a]Lawrence Berkeley National Laboratory, One Cyclotron Road, MS64R0121, Berkeley, CA 94720, USA, [b]Department of Bioengineering, University of California, Berkeley, Berkeley, CA 94720, USA, [c]IGBMC, CNRS–INSERM–UdS, 1 Rue Laurent Fries, BP 10142, 67404 Illkirch, France, and [d]Université de Lorraine: Département de Physique – Nancy 1, BP 239, Faculté des Sciences et des Technologies, 54506 Vandoeuvre-lès-Nancy, France

Correspondence e-mail: pafonine@lbl.gov

Received 13 December 2012
Accepted 5 January 2013

A fast and robust method for determining the parameters for a flat (mask-based) bulk-solvent model and overall scaling in macromolecular crystallographic structure refinement and other related calculations is described. This method uses analytical expressions for the determination of optimal values for various scale factors. The new approach was tested using nearly all entries in the PDB for which experimental structure factors are available. In general, the resulting R factors are improved compared with previously implemented approaches. In addition, the new procedure is two orders of magnitude faster, which has a significant impact on the overall runtime of refinement and other applications. An alternative function is also proposed for scaling the bulk-solvent model and it is shown that it outperforms the conventional exponential function. Similarly, alternative methods are presented for anisotropic scaling and their performance is analyzed. All methods are implemented in the *Computational Crystallography Toolbox* (*cctbx*) and are used in *PHENIX* programs.

Non-atomic model parameters: Twinning

Atomic model parameters

Example of a PDB atom descriptors:

Position *Larger-scale disorder*

```
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
```

Local mobility (small harmonic vibration)

Atomic model parameters

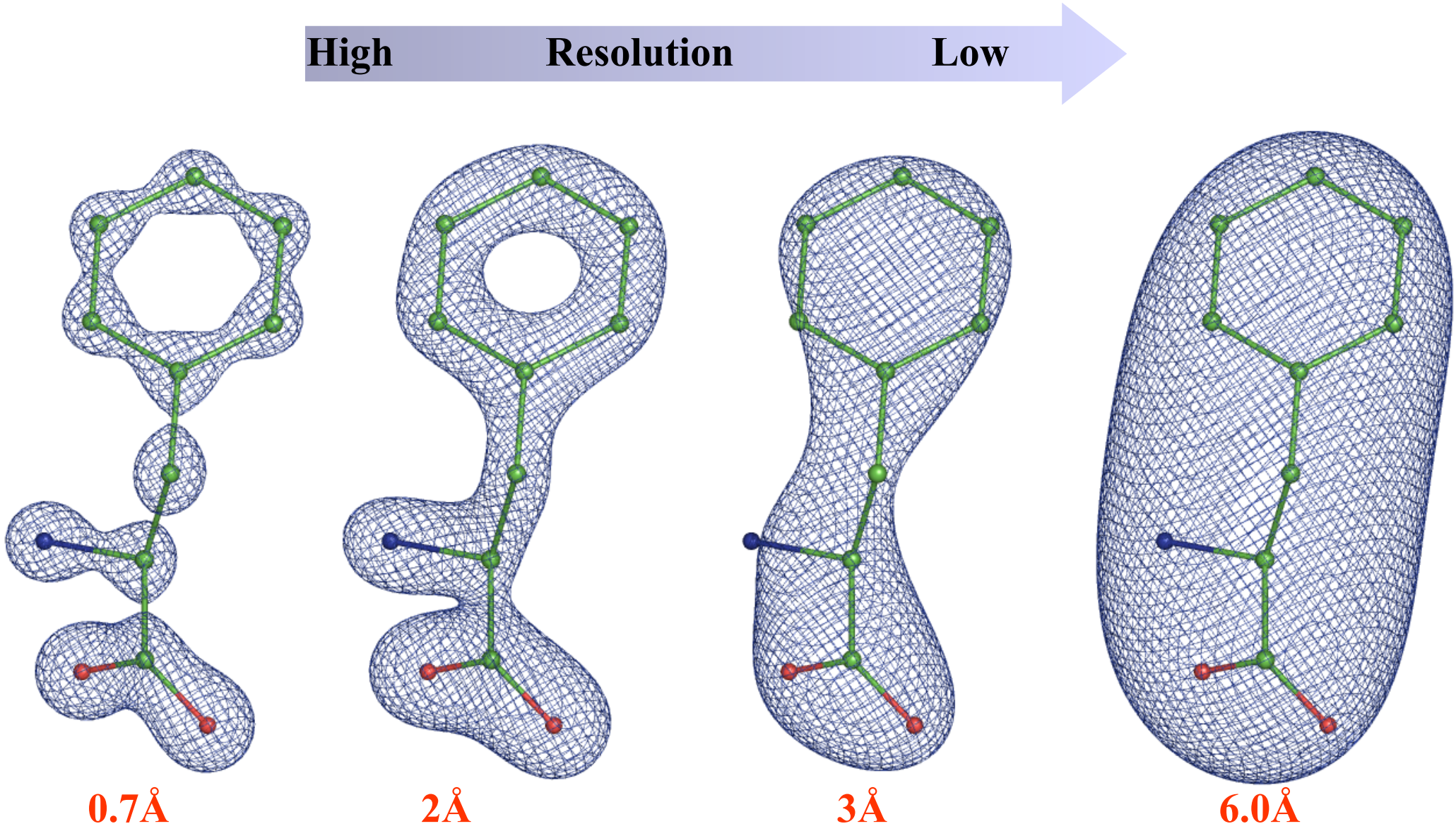
- Position (coordinates)
- Local mobility (ADP; Atomic Displacement Parameters or *B*-factors):

  Diffraction data represents time- and space-averaged images of the crystal structure: time-averaged because atoms are in continuous thermal motions around mean positions, and space-averaged because there are often small differences between symmetry copies of the asymmetric unit in a crystal. ADP is to model the *small* dynamic displacements as isotropic or anisotropic *harmonic* displacements.

- Larger-scale disorder (occupancies)

  Larger displacements (beyond harmonic approximation) can be modeled using occupancies ("alternative conformations/locations").

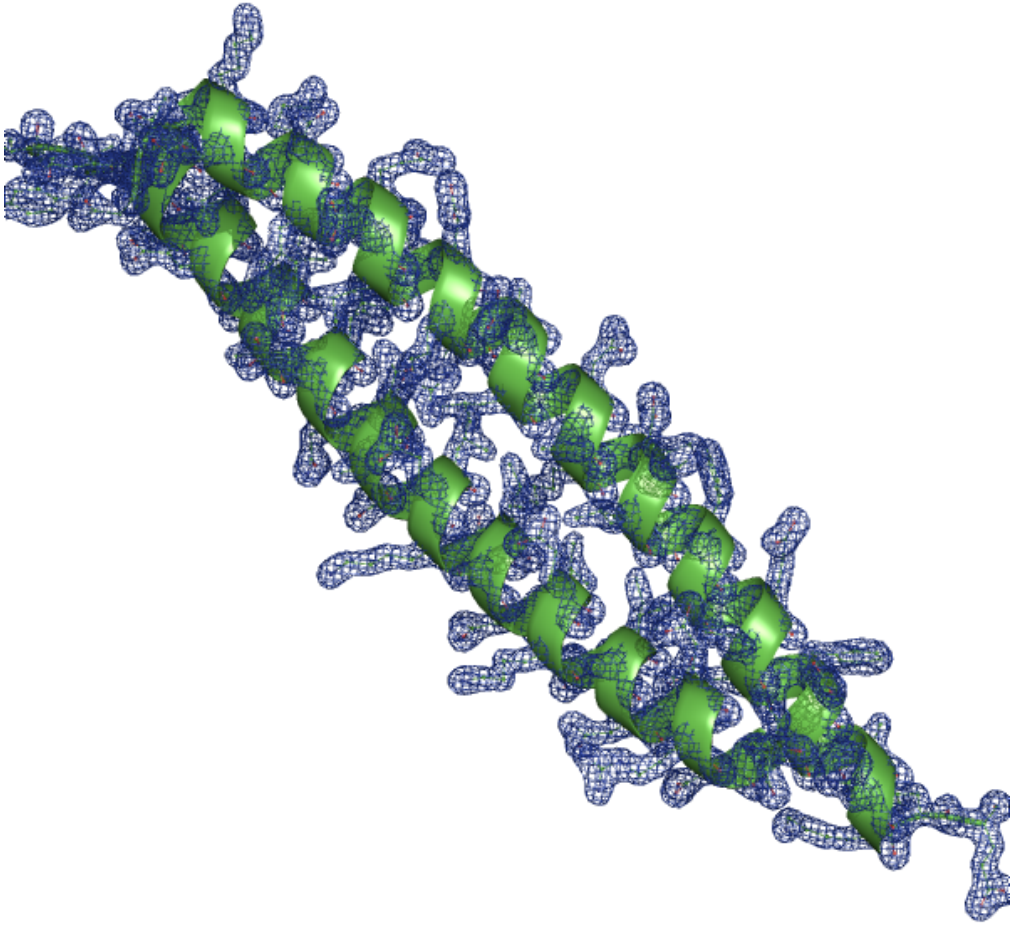
Data quality (resolution, completeness) defines how detailed the model is



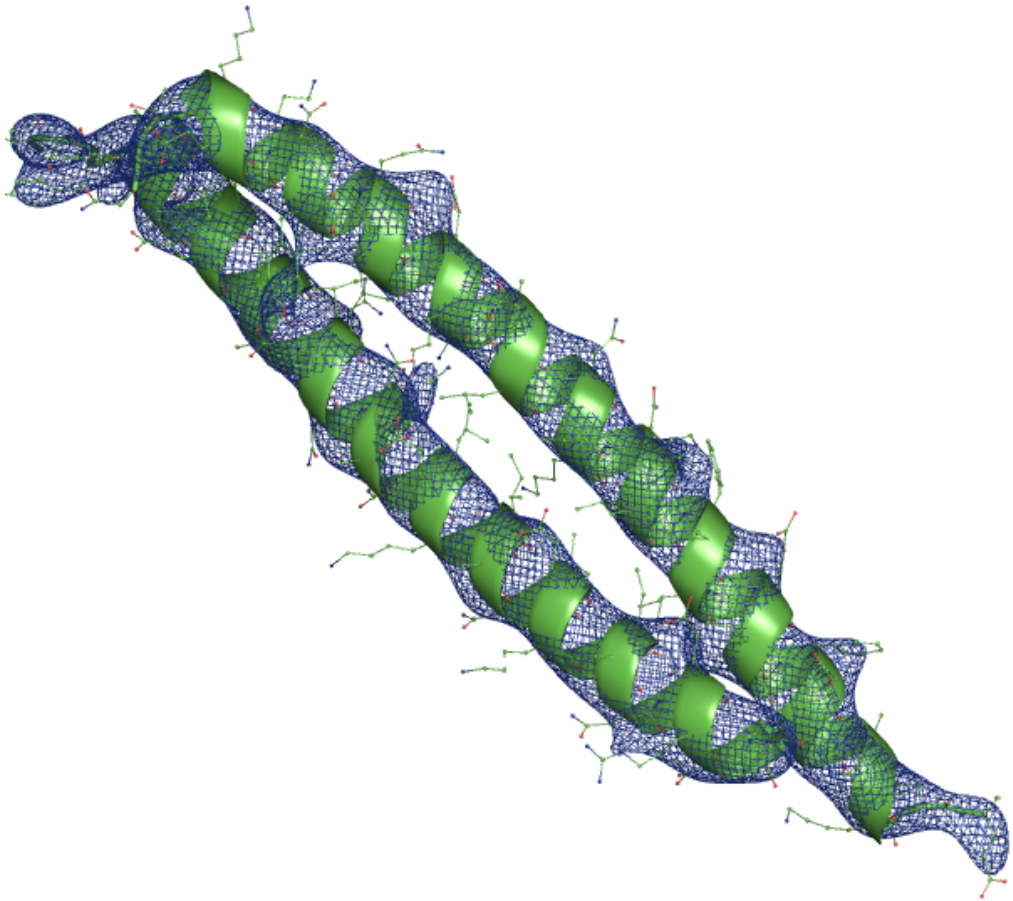
Model parametrization should match data quality (mostly resolution)

Data quality (resolution, completeness) defines how detailed the model is

High Resolution Low



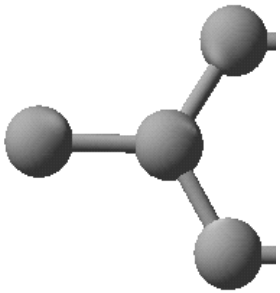
2Å



6.0Å

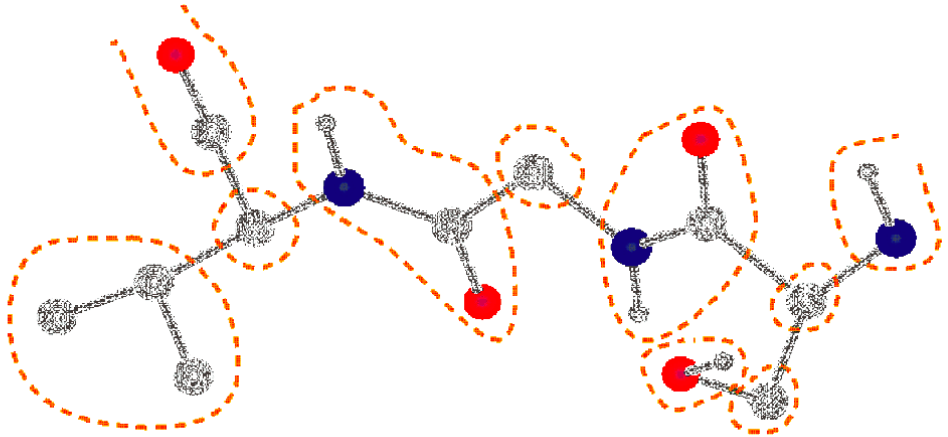
Model parameterization: coordinates

Individual atoms



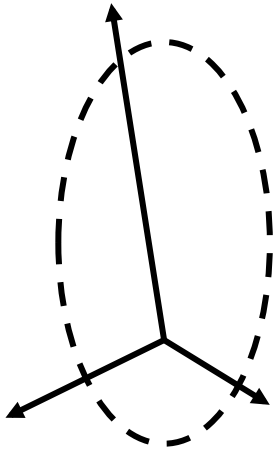
3 * Natoms

Constrained rigid bodies (torsion angle parameterization)



3 * Natoms / (7 …10)

Rigid body

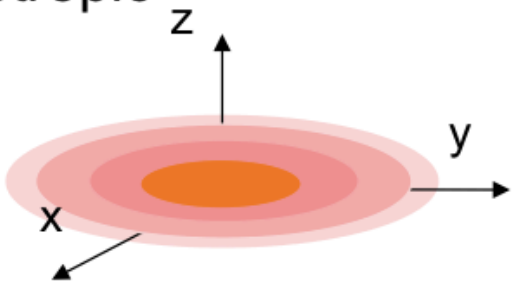


6 * Ngroups

High Resolution Low

Atomic Displacement Parameters (ADP or "B-factors")

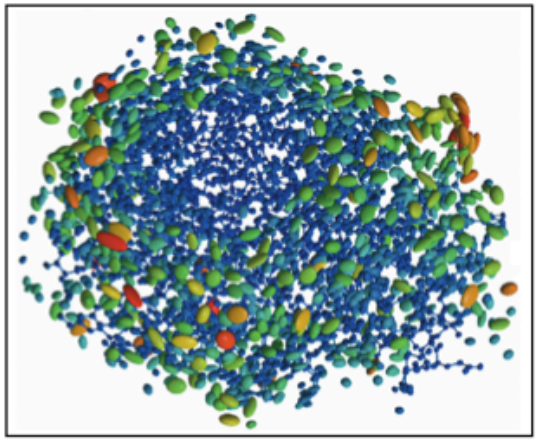
- Atomic displacements are anisotropic



$$\rho(\Delta \mathbf{r}) \sim \exp\{-\Delta \mathbf{r} \, , \, U^{-1} \Delta \mathbf{r}\}$$

$$U = \begin{pmatrix} U_{11} & U_{12} & U_{13} \\ U_{12} & U_{22} & U_{23} \\ U_{13} & U_{23} & U_{33} \end{pmatrix}$$

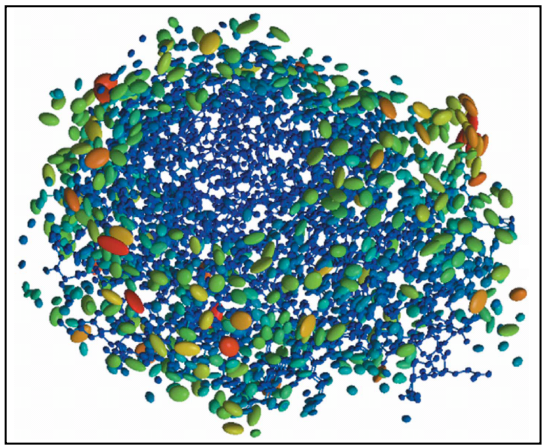
- Hierarchy of atomic displacements



atom
residue
domain
molecule
crystal

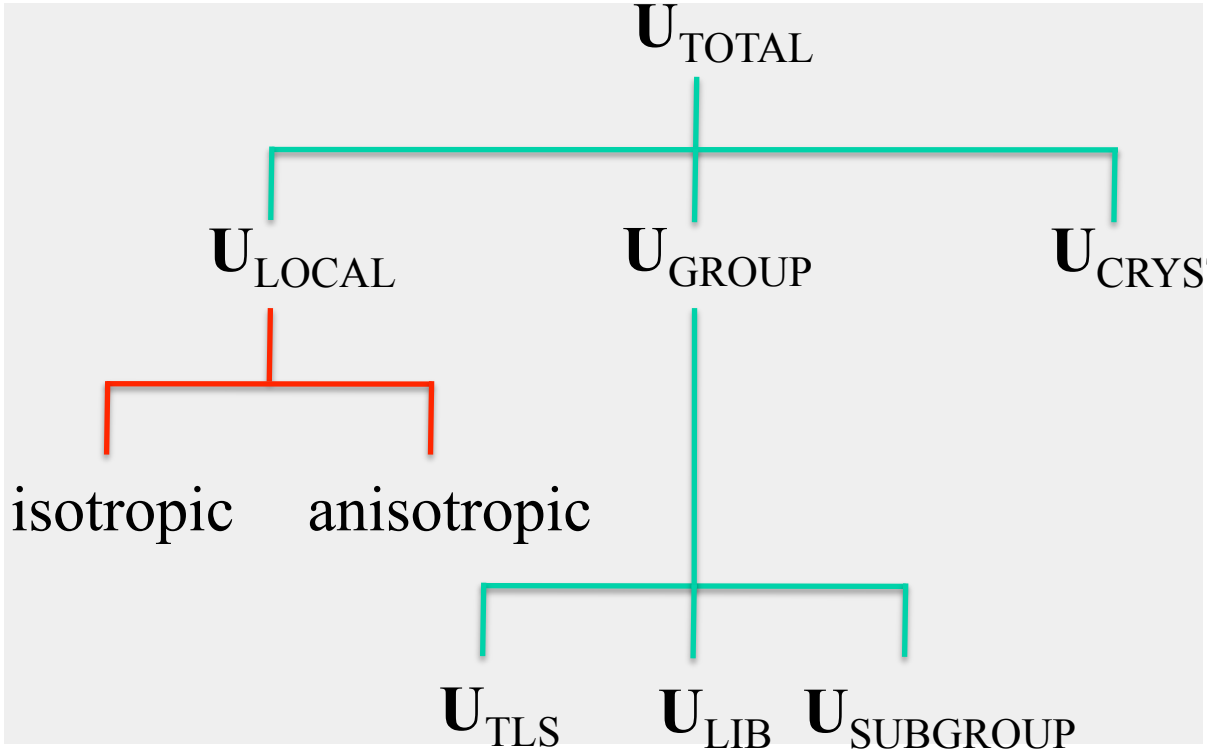
Atomic Displacement Parameters (ADP or "B-factors")

- Hierarchy of atomic displacements



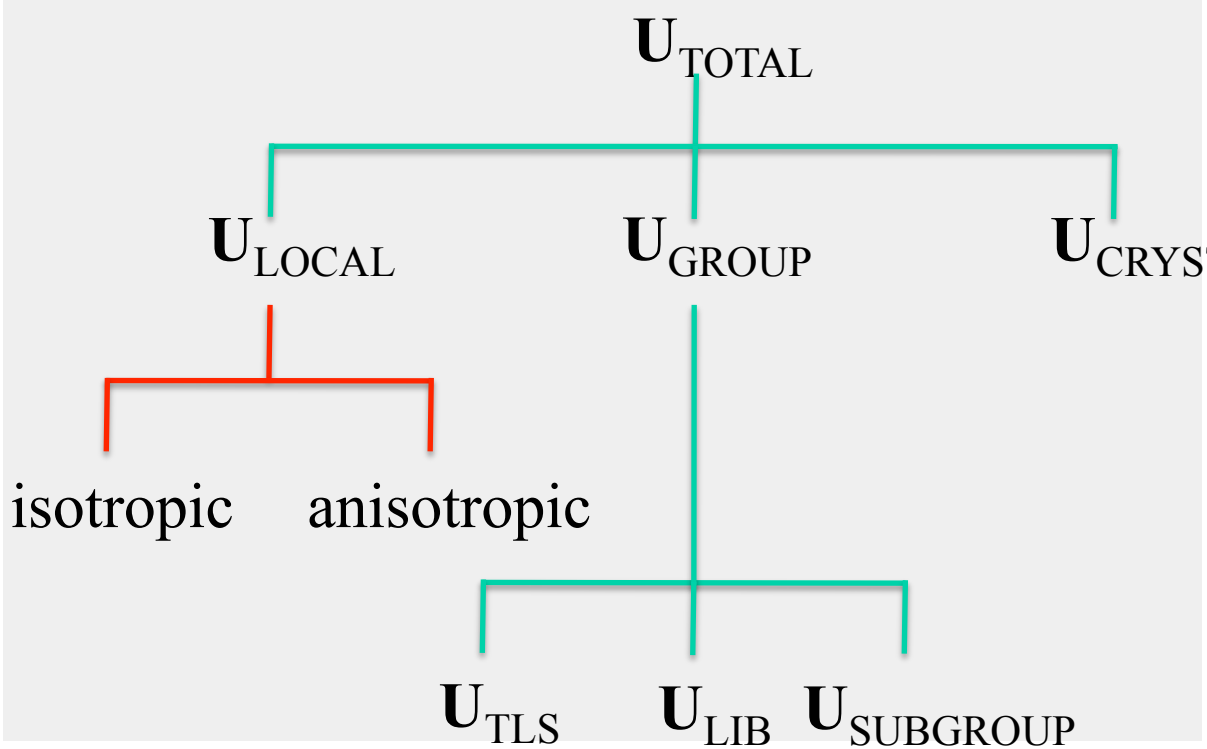
atom
residue
domain
molecule
crystal

Total ADP: $U_{TOTAL} = U_{CRYST} + U_{GROUP} + U_{LOCAL}$



Atomic Displacement Parameters (ADP or "B-factors")

- **Total ADP** $\mathbf{U}_{TOTAL} = \mathbf{U}_{CRYST} + \mathbf{U}_{GROUP} + \mathbf{U}_{LOCAL}$



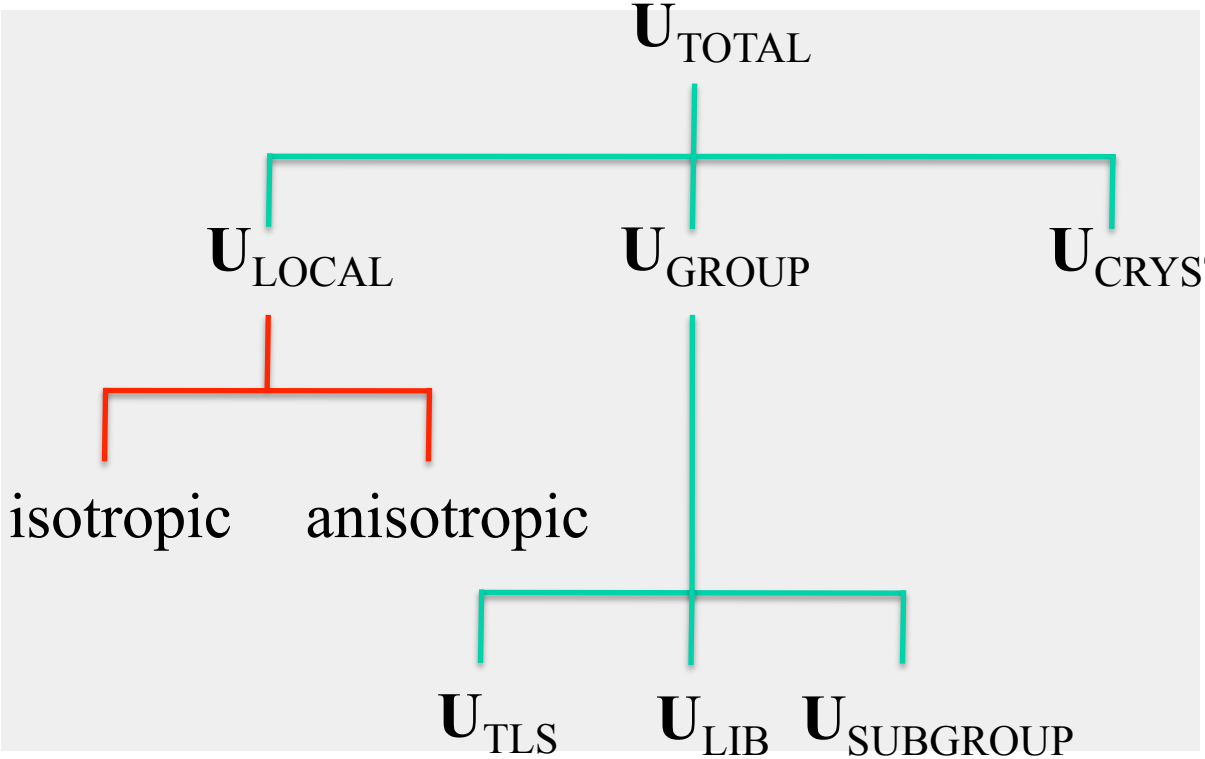
- $\mathbf{U}_{CRYST}$ – lattice vibrations; accounted for by overall anisotropic scale (6 parameters).

$$\mathbf{F}_{MODEL} = k_{OVERALL} e^{-\mathbf{s}\mathbf{U}_{CRYSTAL}\,\mathbf{s}^t} \left(\mathbf{F}_{CALC\_ATOMS} + k_{SOL} e^{-\frac{B_{SOL}\, s^2}{4}} \mathbf{F}_{MASK} \right)$$

Atomic Displacement Parameters: TLS

[Schomaker & Trueblood (1968) On the rigid-body motion of molecules in crystals Acta Cryst. B24, 63-76]

- **Total ADP** $U_{TOTAL} = U_{CRYST} + U_{GROUP} + U_{LOCAL}$



U_{TLS} – rigid body collective displacements of whole molecules, domains, secondary structure elements.

$$U_{TLS} = \mathbf{T} + \mathbf{A}\mathbf{L}\mathbf{A}^t + \mathbf{A}\mathbf{S} + \mathbf{S}^t\mathbf{A}^t$$

(20 TLS parameters per group); **T**, **L** and **S** are 3x3 tensors. **T** and **L** are symmetric, **S** is not.

- **T** describes anisotropic translational displacement (units: Å^2).
- **L** describes rotational displacement (libration) of the rigid group (units: rad^2).
- **S** describes the correlation between the rotation and translation of a rigid body that undergoes rotation about three orthogonal axes that do not intersect at a common point.
- **A** is anti-symmetric tensor; a function of atomic coordinates and TLS origin.

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Use of TLS parameters to model anisotropic displacements in macromolecular refinement

M. D. Winn,[a]* M. N. Isupov[b] and G. N. Murshudov[a,c]

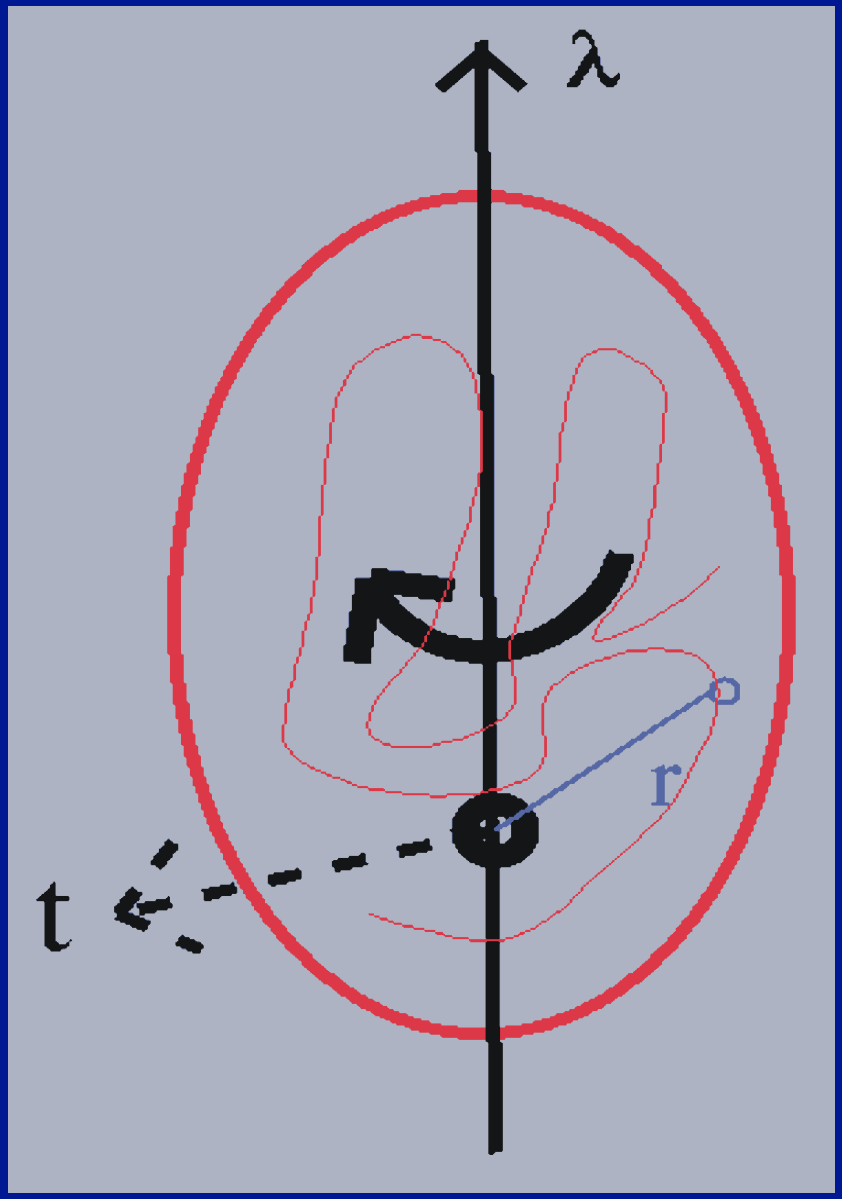
[a]Daresbury Laboratory, Daresbury, Warrington WA4 4AD, England, [b]Department of Chemistry and Biological Sciences, University of Exeter, Exeter EX4 4QD, England, and [c]Chemistry Department, University of York, Heslington, York YO1 5DD, England

Correspondence e-mail: m.d.winn@dl.ac.uk

Received 30 May 2000
Accepted 19 October 2000

An essential step in macromolecular refinement is the selection of model parameters which give as good a description of the experimental data as possible while retaining a realistic data-to-parameter ratio. This is particularly true of the choice of atomic displacement parameters, where the move from individual isotropic to individual anisotropic refinement involves a sixfold increase in the number of required displacement parameters. The number of refinement parameters can be reduced by using collective variables rather than independent atomic variables and one of the simplest examples of this is the TLS parameterization for describing the translation, libration and screw-rotation displacements of a pseudo-rigid body. This article describes the implementation of the TLS parameterization in the macromolecular refinement program *REFMAC*. Derivatives of the residual with respect to the TLS parameters are expanded in terms of the derivatives with respect to individual anisotropic U values, which in turn are calculated using a fast Fourier transform technique. TLS refinement is therefore fast and can be used routinely. Examples of TLS refinement are given for glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and a transcription activator GerE, for both of which there is data to only 2.0 Å, so that individual anisotropic refinement is not feasible. GAPDH has been refined with between one and four TLS groups in the asymmetric unit and GerE with six TLS groups. In both cases, inclusion of TLS parameters gives improved refinement statistics and in particular an improvement in R and free R values of several percent. Furthermore, GAPDH and GerE have two and six molecules in the asymmetric unit, respectively, and in each case the displacement parameters differ significantly between molecules. These differences are well accounted for by the TLS parameterization, leaving residual local displacements which are very similar between molecules and to which NCS restraints can be applied.

Rigid-body motion



General displacement of a rigid-body point can be described as a rotation along an axis passing through a fixed point together with a translation of that fixed point.

$$\underline{u} = \underline{t} + D\underline{r}$$

for small librations

$$\underline{u} \approx \underline{t} + \underline{\lambda} \times \underline{r}$$

D = rotation matrix

$\underline{\lambda}$ = vector along the rotation axis of magnitude equal to the angle of rotation

TLS parameters

Dyad product:

$$\underline{uu}^T = \underline{tt}^T + \underline{t\lambda}^T \times \underline{r}^T - \underline{r} \times \underline{\lambda t}^T - \underline{r} \times \underline{\lambda\lambda}^T \times \underline{r}^T$$

ADPs are the time and space average

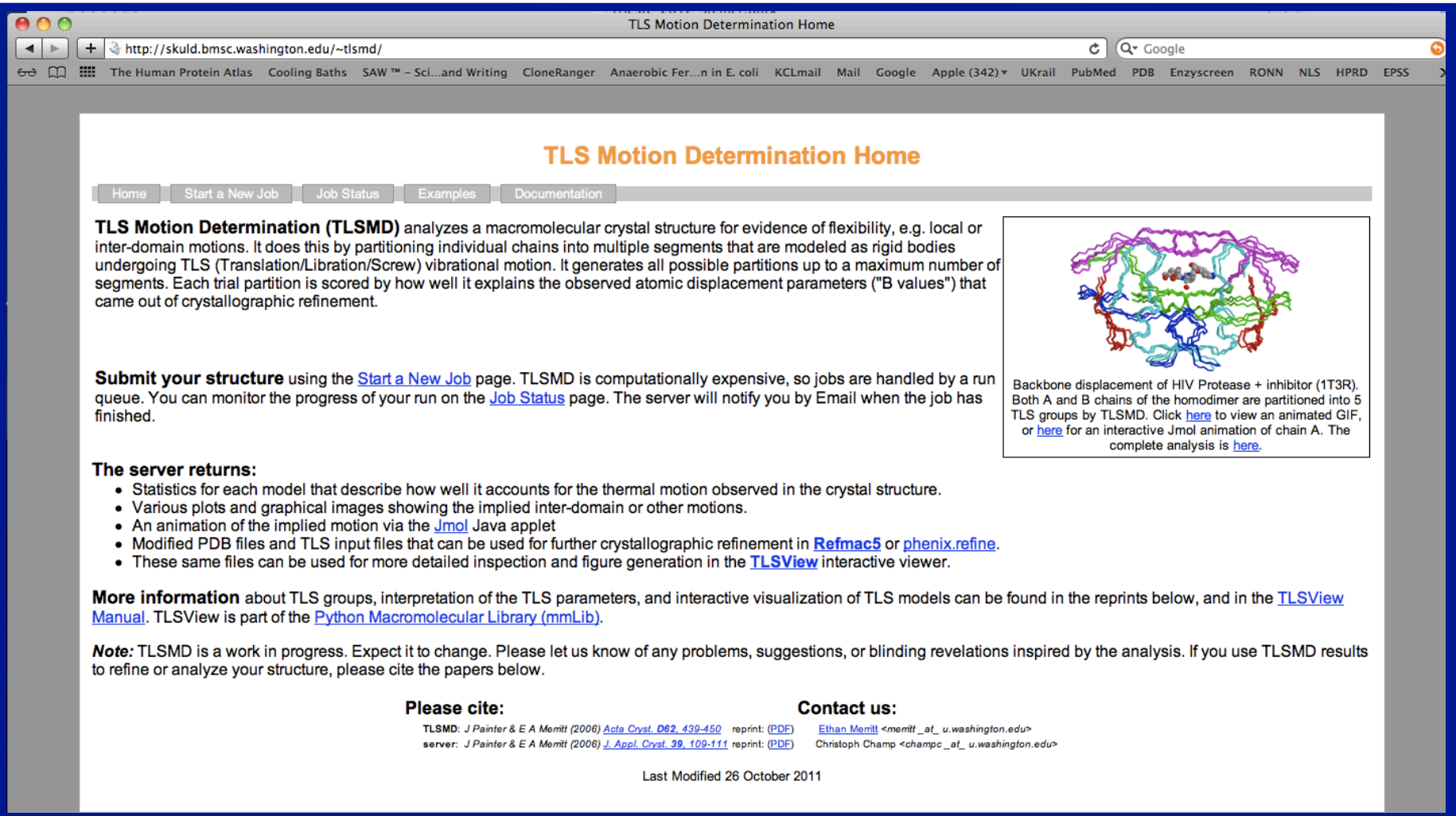
$$U_{TLS} = \langle uu^T \rangle = T + S^T \times \underline{r}^T - \underline{r} \times S - \underline{r} \times L \times \underline{r}^T$$

$T = \langle \underline{tt}^T \rangle$ 6 parameters, TRANSLATION

$L = \langle \underline{\lambda\lambda}^T \rangle$ 6 parameters, LIBRATION

$S = \langle \underline{\lambda t}^T \rangle$ 8 parameters, SCREW-ROTATION

Choice of TLS groups and resolution

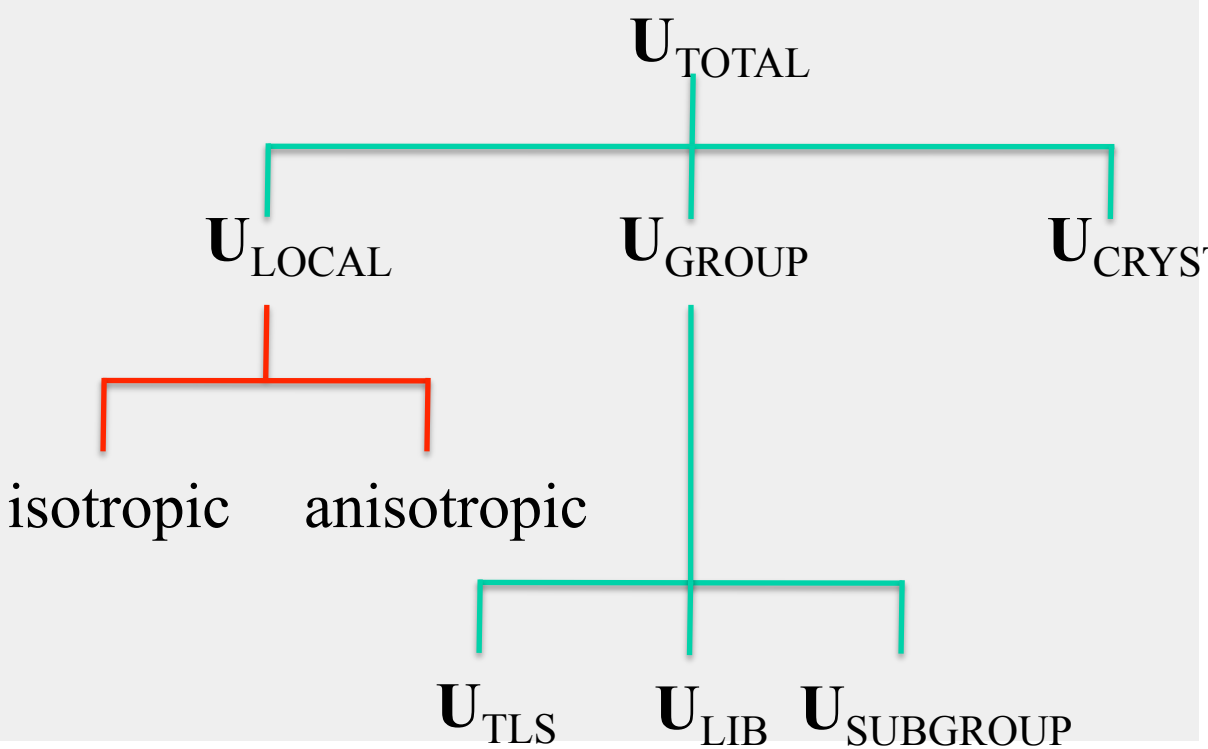


phenix.refine **also offers a TLS selection routine**

Resolution is not a problem. There are only 20 more parameters per TLS group

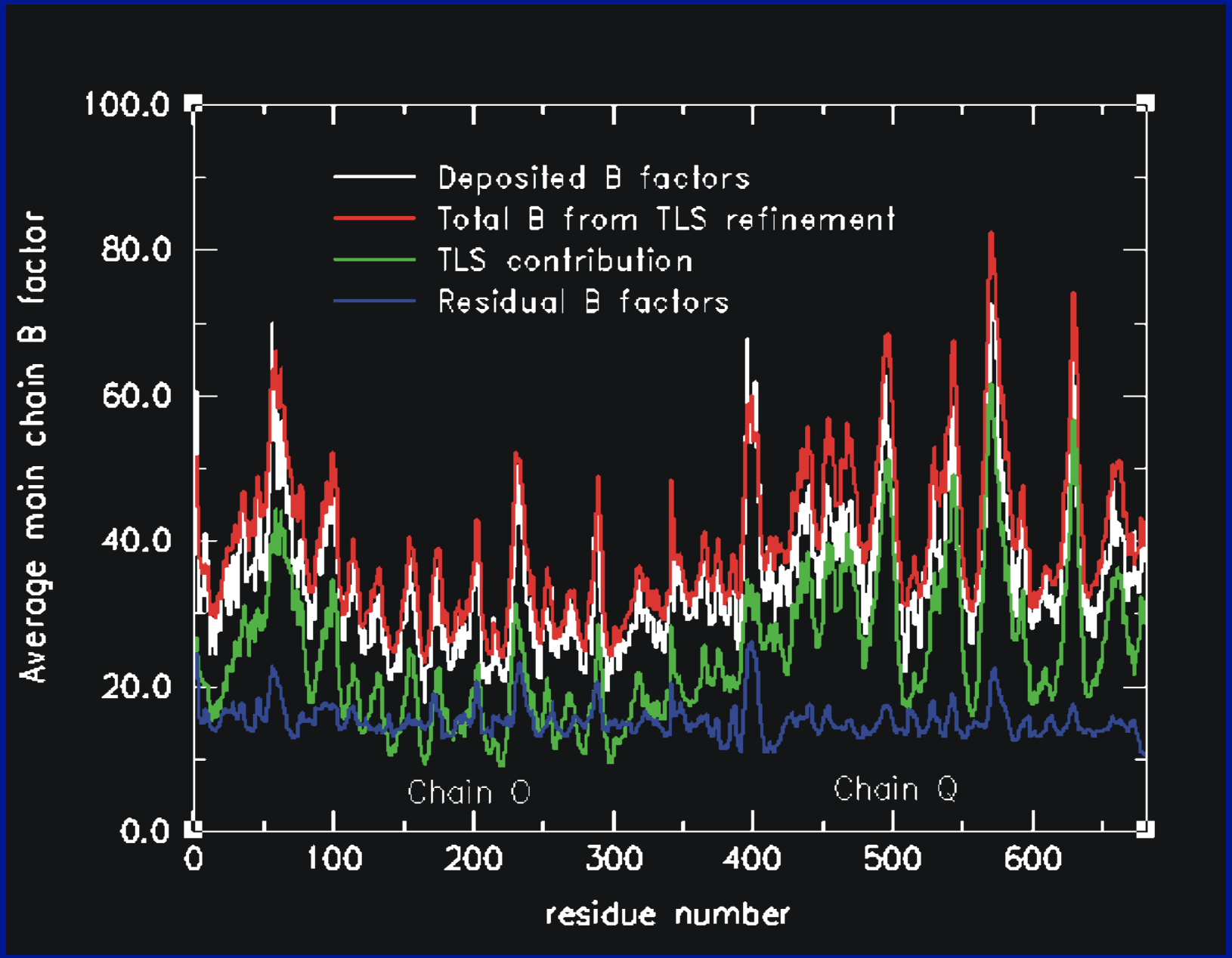
Atomic Displacement Parameters (ADP or "B-factors")

- **Total ADP** $U_{TOTAL} = U_{CRYST} + U_{GROUP} + U_{LOCAL}$



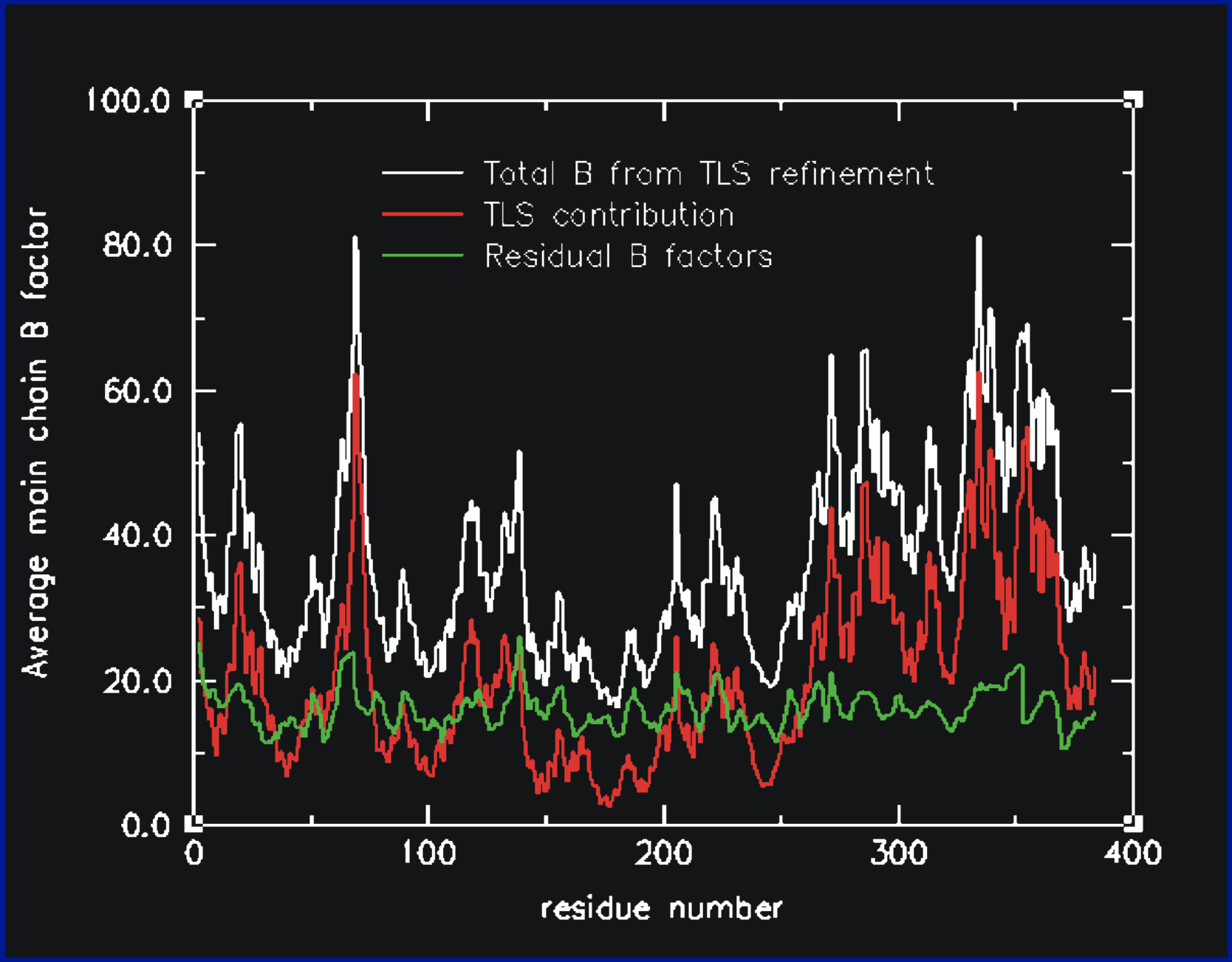
- U_{LOCAL} – local vibration of individual atoms.
  - Depending on data amount and quality, it can be less precise (isotropic) or more precise (anisotropic).
  - These vibrations are expected to be very small due to assumption of rigidity of interatomic bonds (vibrating atoms cannot stretch the bond much).

Contributions to equivalent isotropic *B*s

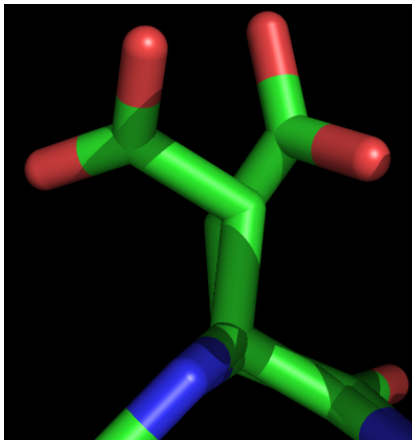


[Howlin, B. & al. (1993) TLSANL: TLS parameter-analysis program for segmented anisotropic refinement of macromolecular structures, J. Appl. Cryst. 26, 622-624]

Contribution to equivalent isotropic *B*s



Occupancy: large-scale disorder that cannot be modeled with harmonic model (ADP)



- Occupancy is the fraction of molecules in the crystal in which a given atom occupies the position specified in the model.
- If all molecules in the crystal are identical, then occupancies for all atoms are 1.00.

- We may refine occupancy because sometimes a region of the molecules may have several distinct conformations.
- Refining occupancies provides estimates of the frequency of alternative conformations.

```
ATOM      1  N   AARG A 192      -5.782  17.932  11.414  0.72  8.38      N
ATOM      2  CA  AARG A 192      -6.979  17.425  10.929  0.72 10.12      C
ATOM      3  C   AARG A 192      -6.762  16.088  10.271  0.72  7.90      C
ATOM      7  N   BARG A 192     -11.719  17.007   9.061  0.28  9.89      N
ATOM      8  CA  BARG A 192     -10.495  17.679   9.569  0.28 11.66      C
ATOM      9  C   BARG A 192      -9.259  17.590   8.718  0.28 12.76      C
```

Key aspects of refinement

- **Objective function**
- **Method of optimization**
- **Model parametrization**
- **Prior knowledge**

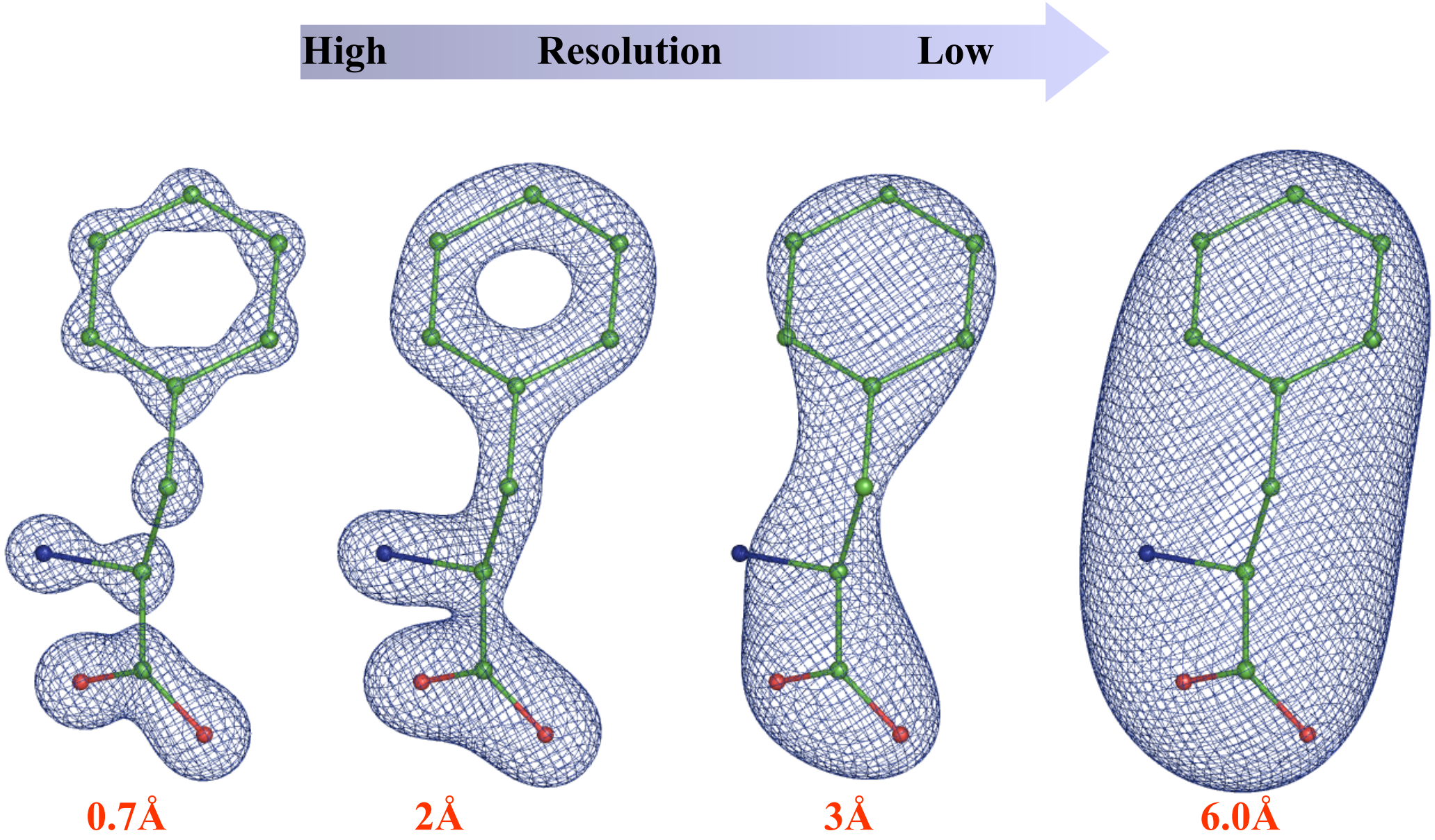
Refinement target function

- **Structure refinement** is a process of changing a model parameters in order to optimize a goal (target) function:

$$T = F(\text{Experimental data}, \text{Model parameters}, \textit{A priori} \text{ knowledge})$$

  - Experimental data – a set of diffraction amplitudes Fobs (and phases, if available).
  - Model parameters: coordinates, ADPs, occupancies, bulk-solvent, …
  - *A priori* knowledge (restraints or constraints) – additional information that may be introduced to compensate for the insufficiency of experimental data (finite resolution, poor data-to-parameters ratio)

- Typically: $T = T_{DATA} + w * T_{RESTRAINTS}$
  - T_{DATA} relates model to experimental data
  - $T_{RESTRAINTS}$ represents *a priori* knowledge
  - w is a weight to balance the relative contribution of T_{DATA} and $T_{RESTRAINTS}$


High
Resolution
Low
0.7Å
2Å
3Å
6.0Å

Observations/Variables ratio

$$f = \sum_{\underline{h}} w(\underline{h})(|F_o| - |F_c|)^2$$

Least-squares crystallographic function

NOT ENOUGH INFORMATION !!!

2.0 Å resolution
2500 non-H atoms (325 aa)

22000 reflections

x,y,z,isoADPs param — obs/var ratio = 2.2

x,y,z,anisoADPs param — obs/var ratio = 0.9777

Restraints

$$f = \sum_{\underline{h}} w(\underline{h})(|F_o| - |F_c|)^2$$

Least-squares crystallographic function

$$+$$

$$\sum_{b} w(b)(B_o - B_c)^2$$

$$+$$

$$\sum_{a} w(a)(A_o - A_c)^2$$

$$+$$

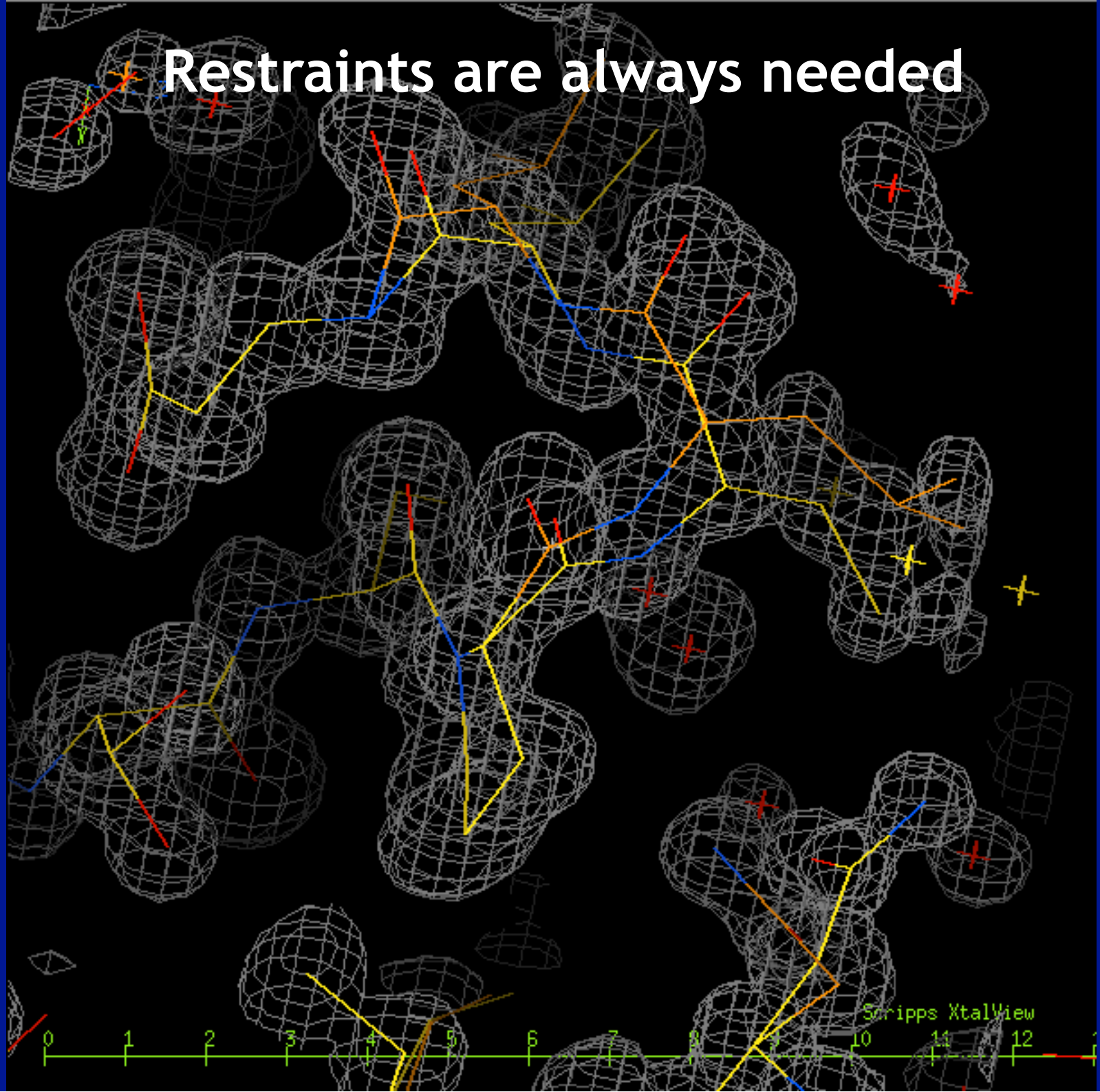
$$\sum \ldots\ldots$$

Restraint functions

Bond lengths
Angles
…

[Waser, J. (1963), Least-squares refinement with subsidiary conditions, Acta Cryst. 16, 1091-1094]
[Konnert, J. (1976), A restrained-parameter structure-factor least-squares refinement procedure for large asymmetric units, Acta Cryst. A32, 614-617]

Restraints are always needed



R and R_{free} statistics

$$f = \Sigma\, w(\underline{h})(|F_o| - |F_c|)^2$$

The 'conventional' R factor compares the observed structure amplitudes $|F_{obs}|$ to those calculated from the current model $|F_{calc}|$. It is defined as

$$R = \frac{\sum_h \left| |F_{obs}| - |F_{calc}| \right|}{\sum_h |F_{obs}|}. \qquad (1)$$

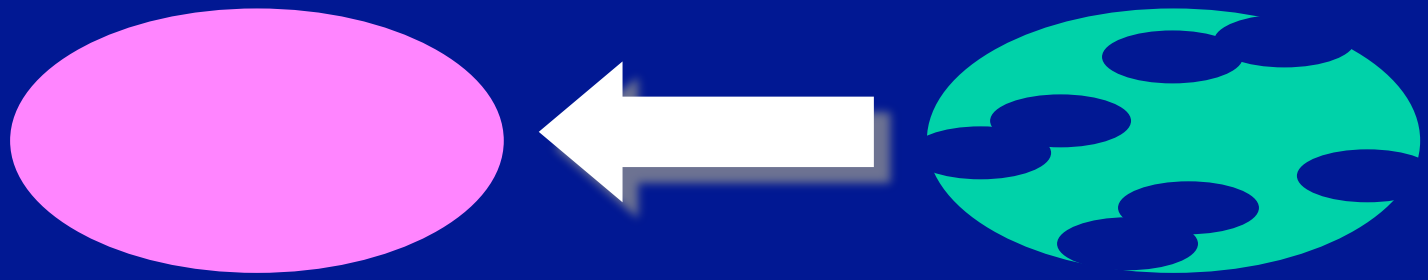
As with other R factors, some authors express it as a percentage. Thus '$R = 20\%$' is the same as '$R = 0.2$'.

The R factor is calculated over a group of reflections h, which may be all the observed reflections, or a particular group. Frequently an R factor is calculated over small ranges or 'bins' of resolution, to give an idea of the performance of the model as resolution is increased.

Problems with LS

$$f = \sum_{\underline{h}} w(\underline{h})(|F_o| - |F_c|)^2$$

Least-squares crystallographic function



Assumption

The distribution of amplitudes is Gaussian with completely known variances

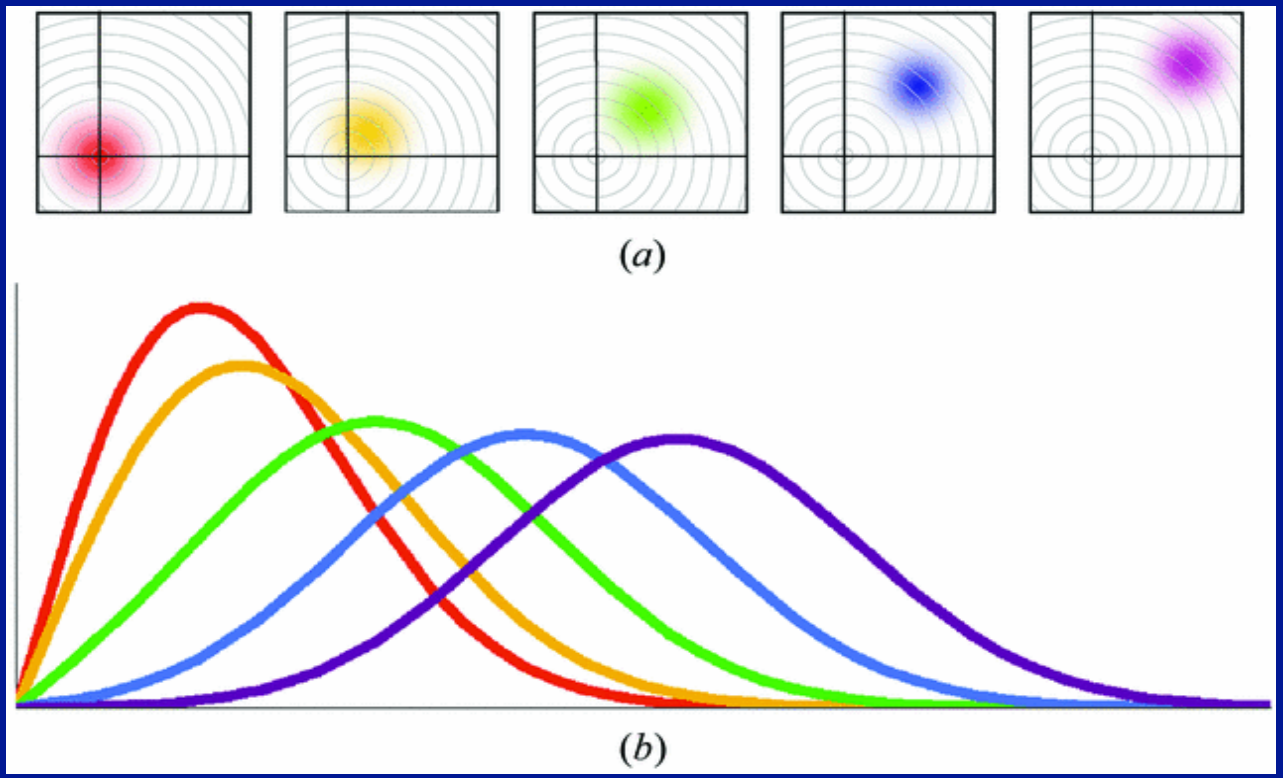
The above assumption can be considered reasonable only at the very final stages of refinement

Rice distribution

$P(\underline{F}_o;\underline{F}_c)$
2D Gaussian

Integrating over the nuisance variable ϕ

$P(|F_o|;|F_c|)$



[McCoy, A.J. (2004) Liking likelihood, Acta Cryst. D60, 2169-2183]

Bayesian approach

The best model is the one which has the highest probability given a set of observations and a certain prior knowledge.

Bayes' theorem

$$P(M;O) = P(M)P(O;M)/P(O)$$

Application of Bayes' theorem

Screening for disease D.

On average 1 person in 5000 dies because of D. P(D)=0.0002
Let P be the event of a positive test for D.
P(P;D)=0.9, i.e. 90% of the times the screening identifies the disease.
P(P;not D)=0.005 (5 in 1000 persons) false positives.

What is the probability of having the disease if the test says it is positive?

P(D;P)=P(D)P(P;D)/P(P)
P(P)=P(P;D)P(D)+P(P;not D)P(not D) = (0.9)(0.0002)+(0.005)(1-0.0002)=0.005179
P(D;P)=(0.0002)(0.9)/(0.005179)=0.0348
Less than 3.5% of persons diagnosed to have the disease actually have it.

Maximum likelihood and the Bayesian view

The best model is the most consistent with the data

Statistically this can be expressed by the likelihood L(O,M)

Bayes' theorem

$$P(M;O) = P(M)\frac{P(O;M)}{P(O)} = P(M)L(O;M)$$

$$L(O;M)$$

$$\max P(M;O) \Leftrightarrow \min -\log P(M;O) = \min [-\log P(M) - \log L(O;M)]$$

[Probability Theory: The Logic of Science by E.T.Jaynes; http://bayes.wustl.edu]
[Bricogne, G. & al. (1997), Methods in Enzymology. 276]
[Murshudov, G.N. & al. (1997), Refinement of macromolecular structures by the maximum-likelihood method, Acta Cryst. D53, 240-255]

Independence

$$\max P(M;O) \Leftrightarrow \min -\log P(M;O) = \min [-\log P(M) - \log L(O;M)]$$

Prior knowledge contibutions and observations
are assumed to be independent (this is a limitation)

$$P(M) = \prod_R P_j(M) \quad \Longrightarrow \quad -\log P(M) = -\sum_R \log P_j(M)$$

$$L(O;M) = \prod_N L_i(O;M) \quad \Longrightarrow \quad -\log L(O;M) = -\sum_N \log L_i(O;M)$$

Target function

A function that relates model parameters to experimental data. Typically looks like this:

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

- **Least-Squares (reciprocal space)**

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

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

- **Maximum-Likelihood (reciprocal space; much better option for macromolecules)**

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

Target function

- Maximum-Likelihood (reciprocal space; option of choice for macromolecules)

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

  - α and β account for model imperfection:
    - α is proportional to the error in atomic parameters and square of overall scale factor;
    - β is proportional to the amount of missing (unmodelled) atoms.

  - α and β are estimated using test reflections by minimization of ML function w.r.t. α and β in each relatively thin resolution bin where α and β can be assumed constant.
    - This is why ML-bases refinement requires *test set reflections*(*) that should be defined sensibly:
      - Each resolution bin should contain at least 50 randomly distributed test reflections.

(*) *Test reflections* – a fraction of reflections (5-10%) put aside for cross-validation.

T_{DATA}: Least-Squares vs Maximum-Likelihood

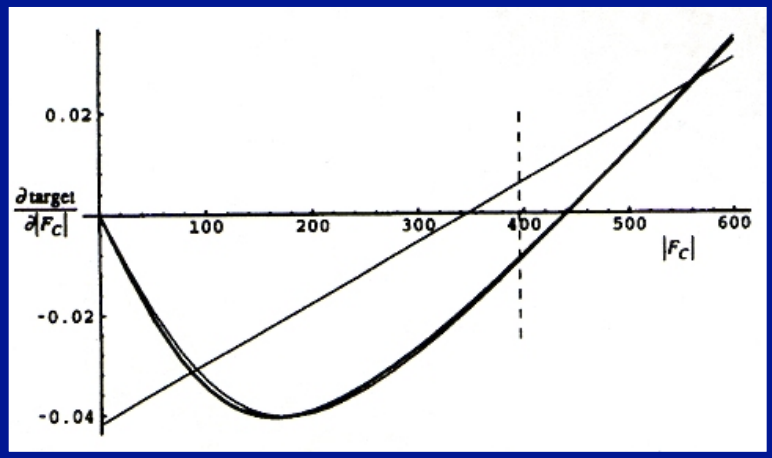
- **Why Maximum-Likelihood target is better than Least-Squares** (in a nutshell):
  - ML accounts for model incompleteness (missing, unmodeled atoms) while LS doesn't;
  - ML automatically downweights the terms corresponding to reflections with the poor fit (poorly measured inaccurate F_{OBS}, high resolution reflections at the beginning of refinement, etc.)

- ***R*-factors in LS and ML refinement**:
  - *R*-factor is expected to decrease during LS based refinement, since the LS target and *R*-factor formula are very similar:

$$R = \frac{\sum |F_{OBS} - F_{MODEL}|}{\sum F_{MODEL}} \qquad LS = \sum_{s} \left(F_{OBS} - F_{MODEL}\right)^2$$

  - In ML based refinement the *R*-factor may eventually decrease (and this is what typically happens in practice) but this is not implied by the ML target function

LS vs ML



[Pannu, N.S. & Read, R.J. (1996), Improved structure refinement through maximum-likelihood , Acta Cryst. A52, 659-668]

Summary objective function

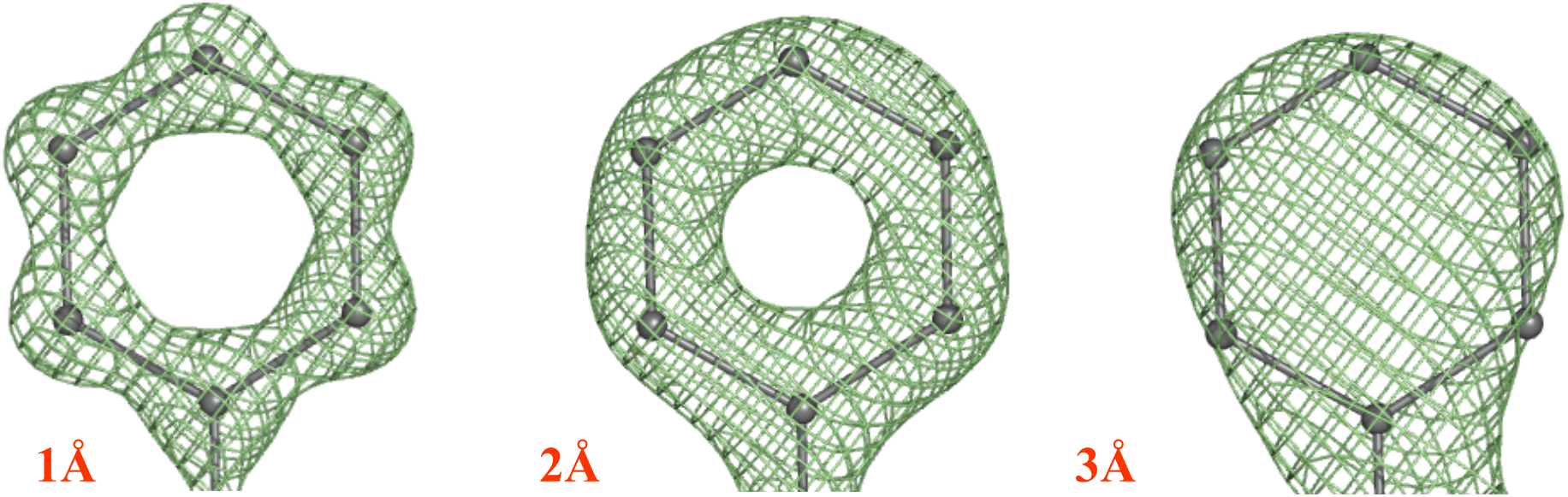
- ML target functions are typically superior to LS target functions
- There are limitations in current ML implementations
- LS is acceptable when the model is complete (SHELXL uses LS with direct summation – no FFT)

Key aspects of refinement

- Objective function
- Method of optimization
- Model parametrization
- Prior knowledge

Restraints in refinement of individual coordinates

Fourier images at different data resolution:



- At lower resolution the electron density is not informative enough to keep the molecule geometry sensible
- Therefore there is a need to bring in some additional a priori knowledge that we may have about the molecules in order to keep the geometry …
- This knowledge is typically expressed *either* as an additional term to the refinement target (***restraints*** term):

$$E_{TOTAL} = w * E_{DATA} + E_{RESTRAINTS}$$

or strict requirement that the model parameter must exactly match the prescribed value and never change during refinement (***constraints***).

Restraints in refinement of individual coordinates

- *A priori* chemical knowledge (restraints) is introduced to keep the model chemically correct while fitting it to the experimental data at lower resolution (less resolution, stronger the weight *W*):

$$E_{TOTAL} = w * E_{DATA} + E_{RESTRAINTS}$$

$$E_{RESTRAINTS} = E_{BOND}+E_{ANGLE}+E_{DIHEDRAL}+E_{PLANARITY}+E_{NONBONDED}+E_{CHIRALITY} + E_{NCS}+E_{RAMACHANDRAN}+E_{REFERENCE}+\ldots$$

- Higher resolution – less restraints contribution (can be completely unrestrained for well ordered parts at subatomic resolution).
- Typically, each term in $E_{RESTRAINTS}$ is a harmonic (quadratic) function: $E = \Sigma\ weight * (X_{model} - X_{ideal})^2$
- *weight* = $1/\sigma(X)^2$ is the inverse variance, in least-squares methods (*e.g.* 0.02 Å for a bond length)
- Making $\sigma(X)$ too small is NOT equivalent to constraints, but will make weight infinitely large, which in turn will stall the refinement.

Restraints: bonds and angles

- Bond distances:

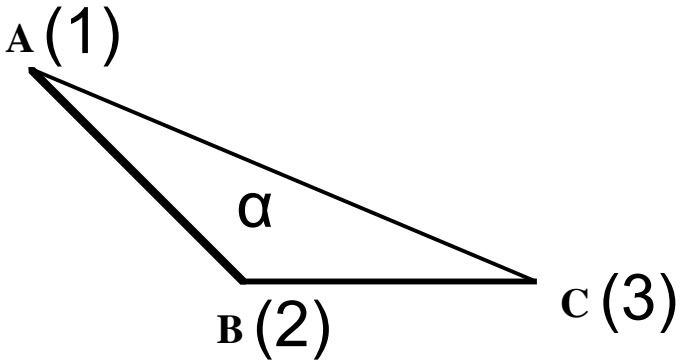
$E = \Sigma_{bonds}\ weight * (d_{model} - d_{ideal})^2$

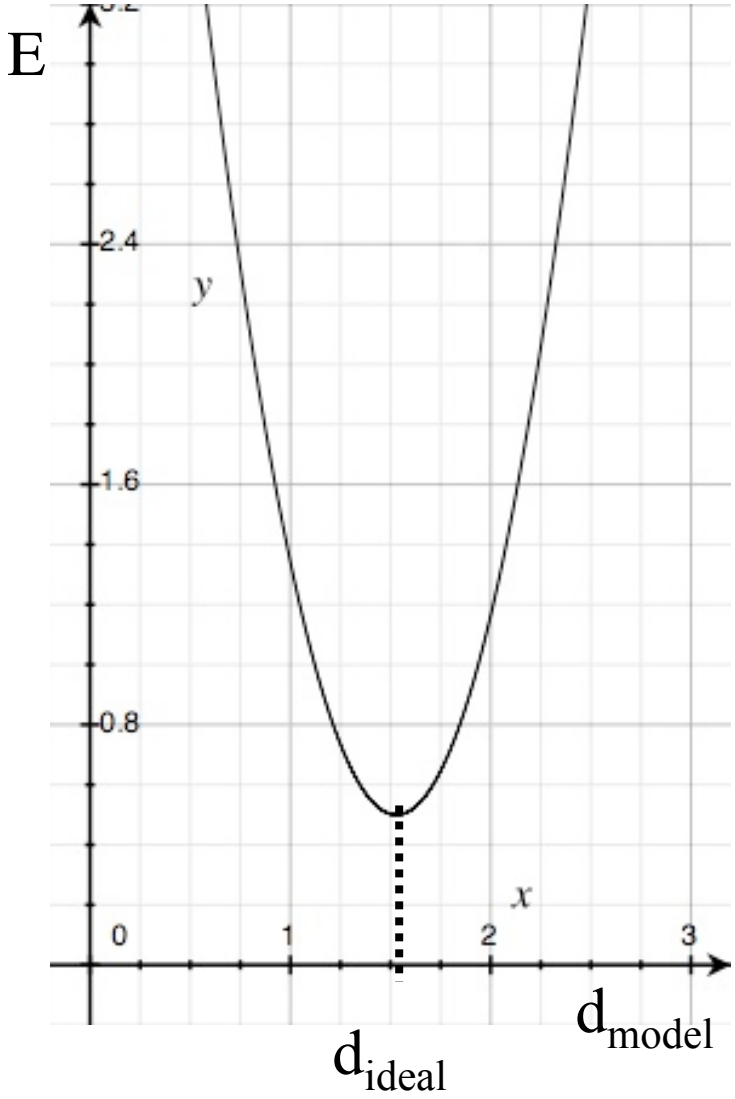
- Bond angles:

$E = \Sigma_{angles}\ weight * (\alpha_{model} - \alpha_{ideal})^2$

Alternatively, one can restrain1-3 distances:

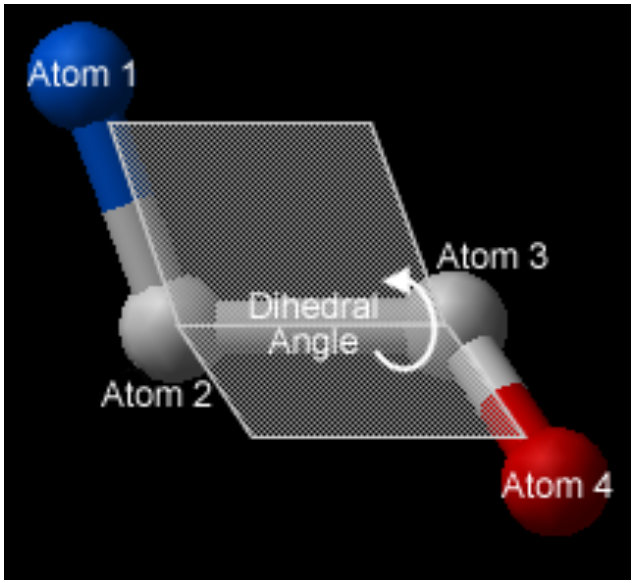
$E = \Sigma_{1\text{-}3\text{-}pairs}\ weight * (d_{model} - d_{ideal})^2$

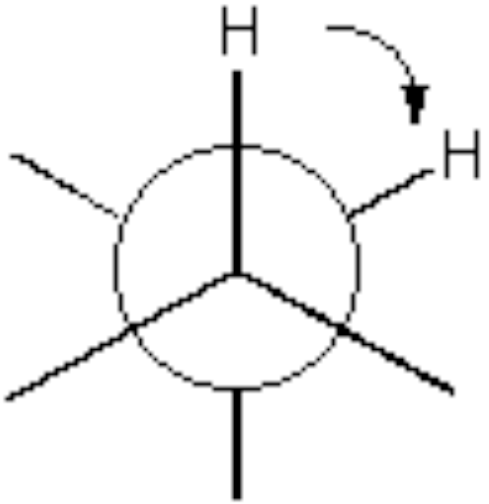




Restraints: dihedral (torsion) angles

- Dihedral or torsion angle is defined by 4 sequential bonded atoms 1-2-3-4
  - Dihedral = angle between the planes 123 and 234
  - Torsion = looking at the projection along bond B-C, the angle over which one has to rotate A to bring it on top of D (clockwise = positive)





- Three possible ways to restraining dihedrals:
  - $E = \Sigma_{dihedrals}\ weight * (\chi_{ideal} - \chi_{model})^2$ (if only one target value for the dihedral)
  - $E = \Sigma_{dihedrals}\ weight * (1 + \cos(n\,\chi_{model} + \chi_{shift}))$ (n = periodicity)
  - $E = \Sigma_{1\text{-}4\text{-}pairs}\ weight * (d_{model} - d_{ideal})^2$
    (sign ambiguity unless $\chi = 0^\circ$ or 180°, *i.e.* both χ and $-\chi$ give rise to the same 1-4 distances)

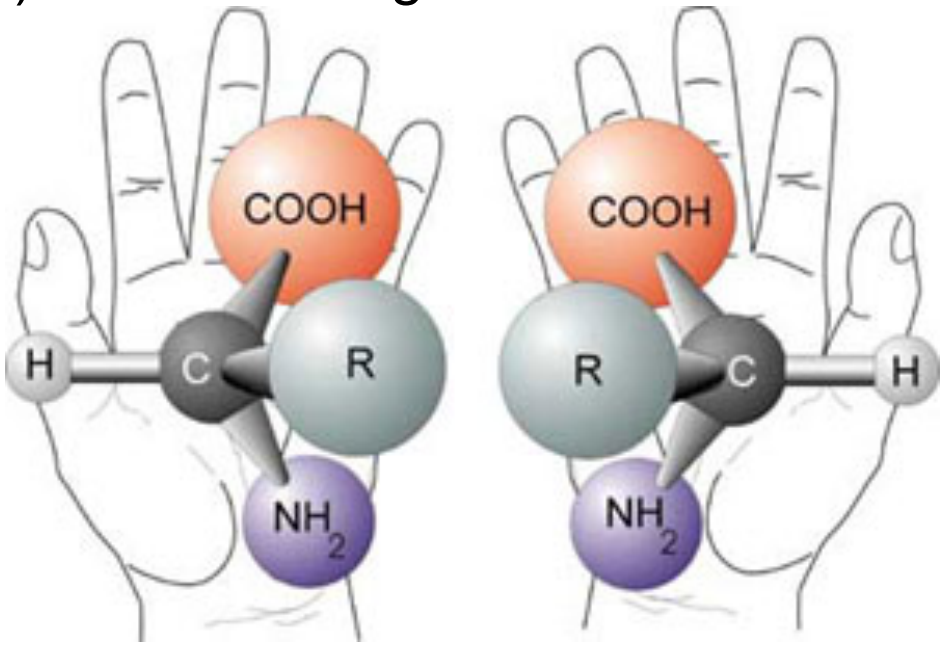
Restraints: chirality

- A chiral molecule has a non-superposable mirror image
- Chirality restraints (example: for Cα atoms) defined through chiral volume:

$$V = (\mathbf{r}_N - \mathbf{r}_{CA}) \bullet [(\mathbf{r}_C - \mathbf{r}_{CA}) \times (\mathbf{r}_{CB} - \mathbf{r}_{CA})]$$

sign depends on handedness ($V_D = -V_L$)

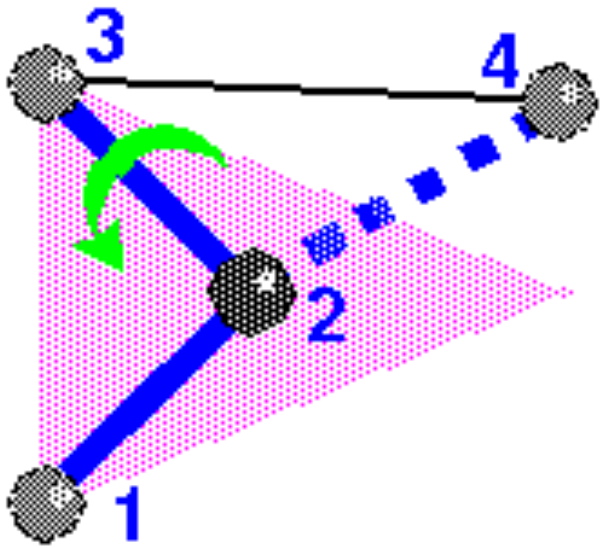
$$E = \Sigma_{chiral}\ weight * (V_{model} - V_{ideal})^2$$



- Alternatively, chirality restraints can be defined by an "improper torsion" ("improper", because it is not a torsion around a chemical bond)

Example: for Cα: torsion (Cα-N-C-Cβ) = +35˚ for L-aa, -35˚ for D-aa

$$E = \Sigma_{chiral}\ weight * (\chi_{ideal} - \chi_{model})^2$$

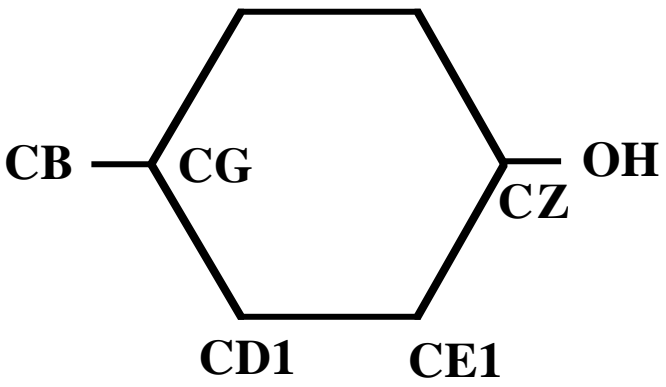


Restraints: planarity

- Planarity (double bonds, aromatic rings):
  - Identify a set of atoms that has to be in plane, and then for each set, minimise sum of distances to the best-fitting plane through the atoms

    $E = \Sigma_{planes} \Sigma_{atoms\_in\_plaine}$ *weight* * $(\underline{\mathbf{m}} \bullet \underline{\mathbf{r}} - d)^2$

  - Restrain the distances of all atoms in the plane to a dummy atom that lies removed from the plane
  - Define a set of ("fixed", "non-conformational") dihedral angles (or improper torsions) with target values of 0˚ or 180˚:



(CB-CG-CD1-CE1) = 180

(CG-CD1-CE1-CZ) = 0

(CD1-CE1-CZ-OH) = 180

(CD1-CE1-CZ-CE2) = 0

(CE1-CZ-CE2-CD2) = 0

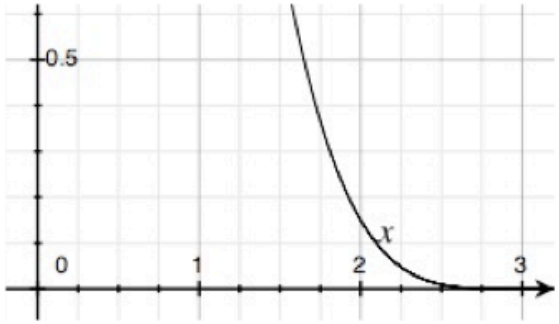
(CZ-CE2-CD2-CG) = 0

(CE2-CD2-CG-CD1) = 0

(CD2-CG-CD1-CE1) = 0

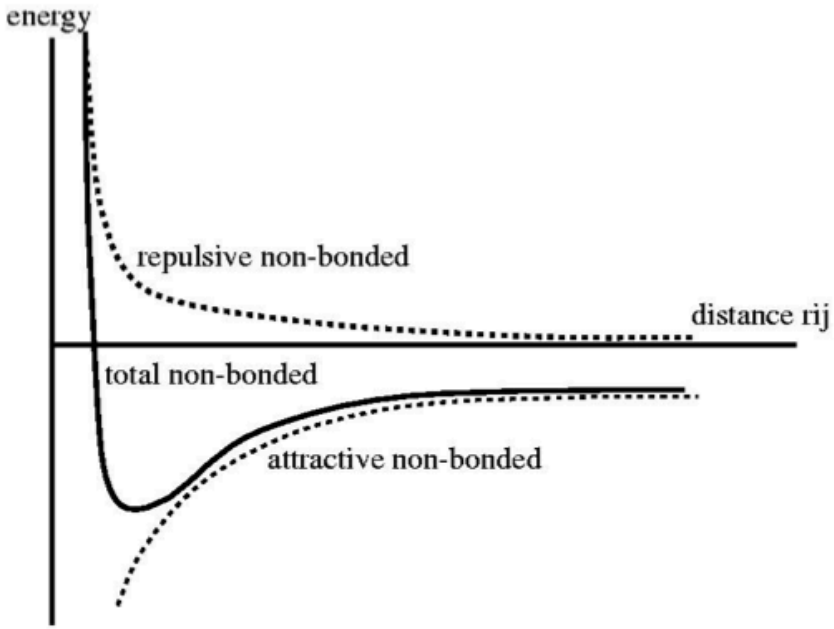
Restraints: non-bonded

- Simple repulsive term: $E = \Sigma_{nb}\ weight * (d_{model} - d_{min})^4$ (only if $d_{model} < d_{min}$)



- Combined function: Van der Waals and electrostatics terms

$$E = E_{attractive} + E_{repulsive} + E_{electrostatic} =$$
$$\Sigma_{nb} (Ad_{model}^{-12} - Bd_{model}^{-6} + Cq_1q_2/d_{model})$$



Sources of target ("ideal") values for constraints and restraints

- Libraries (for example, Engh & Huber) created out of small molecules that are typically determined at much higher resolution, use of alternative physical methods (spectroscopies, etc).

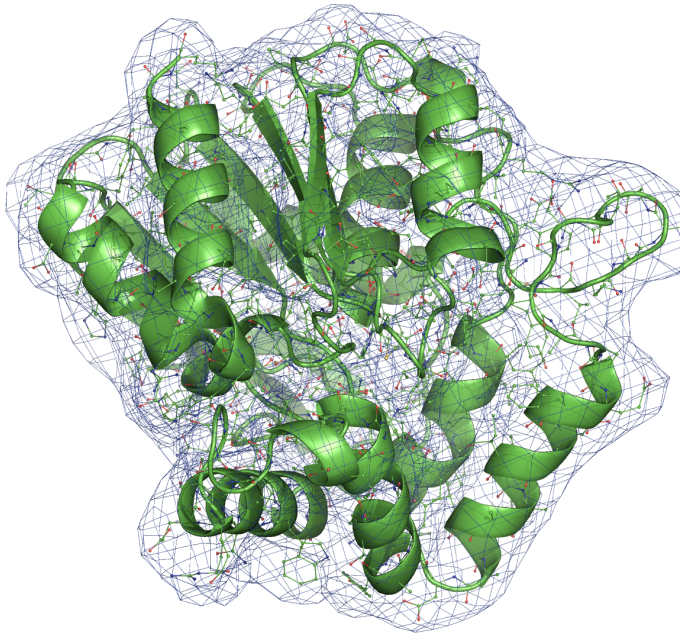
- Analysis of macromolecular structures solved at ultra-high resolution

- Pure conformational considerations (Ramachandran plot), tabulated secondary structure parameters

- QM (quantum-chemical) calculations

Specific restraints for refinement at low and very low resolution

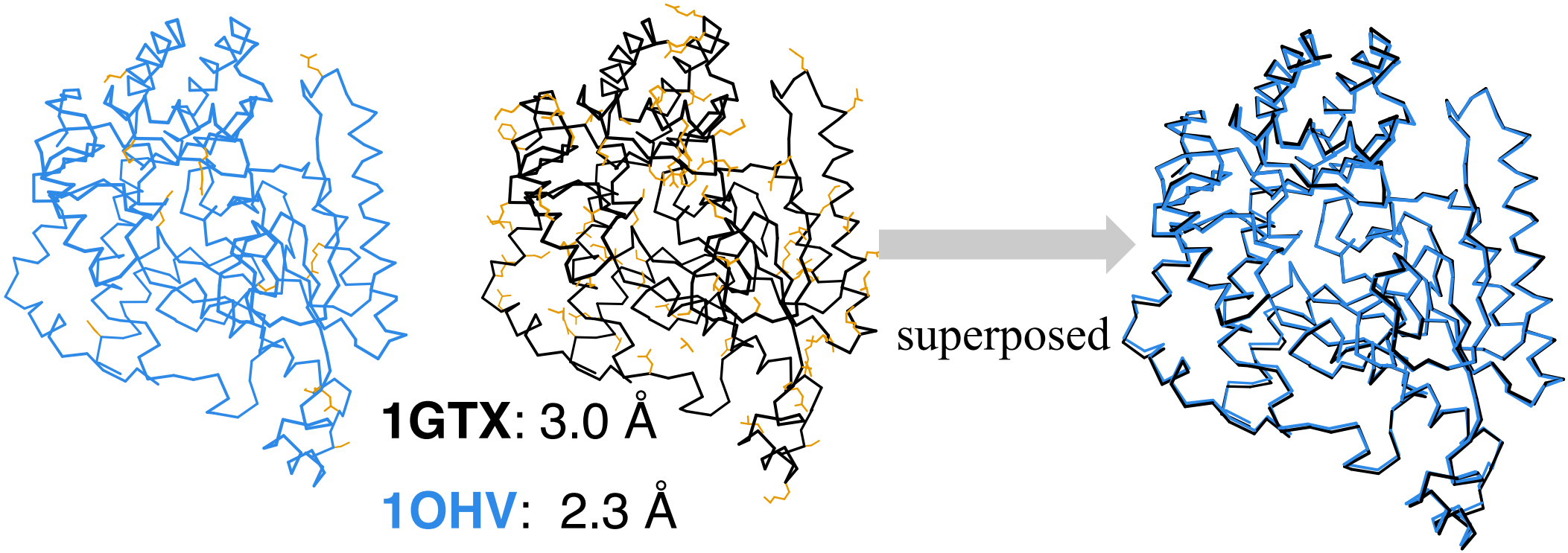
- At low(ish) resolution the electron density map is not informative enough and a set of local restraints are insufficient to maintain known higher order structure (secondary structure), and the amount of data is too small compared to refinable model parameters …



- … therefore one needs to bring in more information in order to assure the overall correctness of the model:
  - Reference model
  - Secondary structure restraints
  - Ramachandran restraints
  - NCS restraints/constraints

Specific restraints for refinement at low and very low resolution

- **Reference model:**
  - If you are lucky, there may be a higher resolution structure available that is similar to low resolution structure
  - Use higher resolution information to direct low-resolution refinement



- Reference point restraint for isolated atoms (water / ions): sometime density peak may not be strong enough to keep an atom in place (due to low resolution or low site occupancy, for example), so it can drift away from it. Use harmonic restraint to peak position.

Specific restraints for refinement at low and very low resolution

- **Secondary structure restraints**
  - H-bond restraints for alpha helices, beta sheets, RNA/DNA base pairs
  - This requires correct annotation of secondary structure elements:
    - It can be done automatically using programs like DSSP / KSDSSP
    - Or… manually….or with *ProSMART*

research papers

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Low-resolution refinement tools in *REFMAC5*

Robert A. Nicholls, Fei Long and Garib N. Murshudov*

Structural Studies Division, MRC Laboratory of Molecular Biology, Cambridge CB2 0QH, England

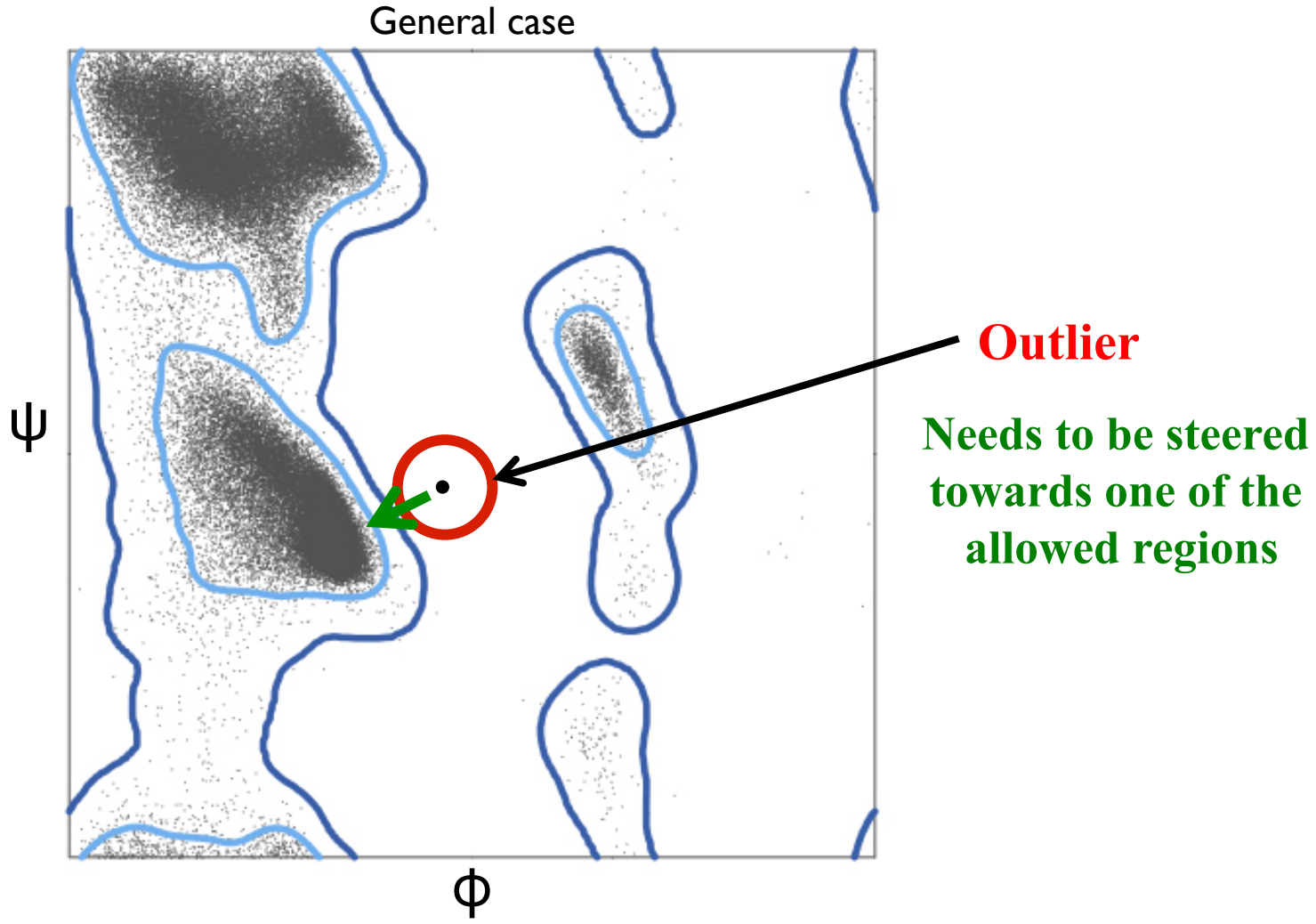
Correspondence e-mail: garib@mrc-lmb.cam.ac.uk

Two aspects of low-resolution macromolecular crystal structure analysis are considered: (i) the use of reference structures and structural units for provision of structural prior information and (ii) map sharpening in the presence of noise and the effects of Fourier series termination. The generation of interatomic distance restraints by *ProSMART* and their subsequent application in *REFMAC5* is described. It is shown that the use of such external structural information can enhance the reliability of derived atomic models and stabilize refinement. The problem of map sharpening is considered as an inverse deblurring problem and is solved using Tikhonov regularizers. It is demonstrated that this type of map sharpening can automatically produce a map with more structural features whilst maintaining connectivity. Tests show that both of these directions are promising, although more work needs to be performed in order to further exploit structural information and to address the problem of reliable electron-density calculation.

Received 2 November 2011
Accepted 28 December 2011

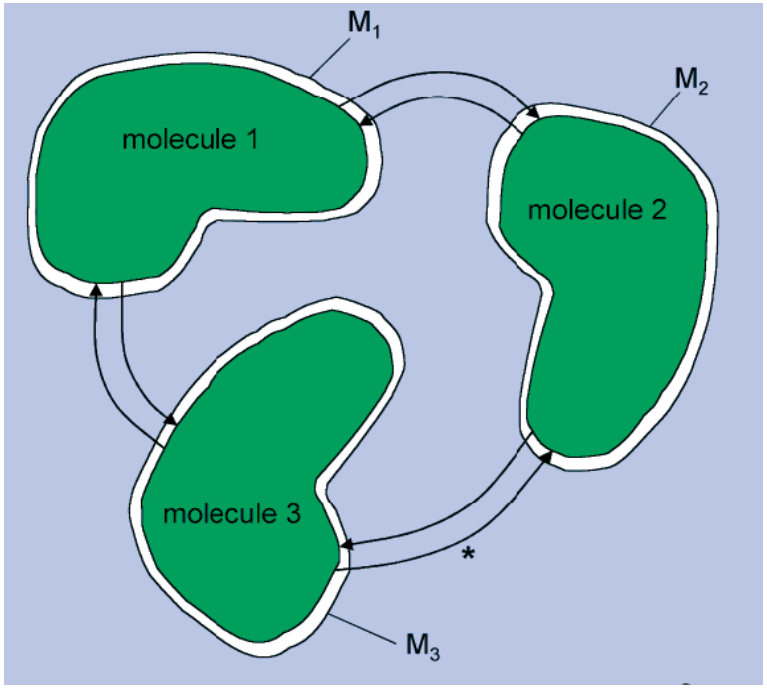
Specific restraints for refinement at low and very low resolution

- **Ramachandran restraints**
  - steer outliers towards favored region
  - should only be used at low resolution
  - should never be used at higher resolution, since it is one of the few precious validation tools (sometimes compare to “real-space analog of Rfree”)



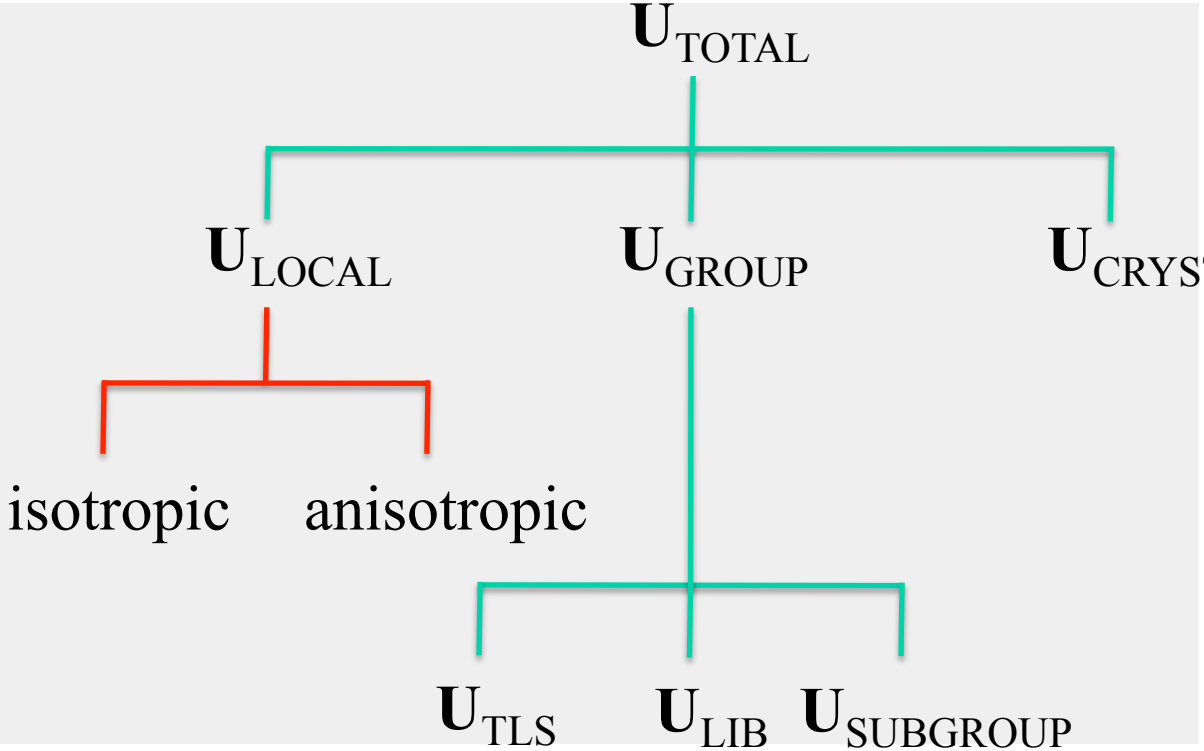
Specific restraints for refinement at low and very low resolution: NCS

- **NCS (non-crystallographic symmetry) restraints/constraints**
  - Multiple copies of a molecule/domain in the asymmetric unit that are assumed to have similar conformations (and sometimes B-factors)
  - Restrain positional deviations from the average structure
    $E = \Sigma_{atoms}\ weight * \Sigma_{NCS}\ |\mathbf{r} - \langle\mathbf{r}\rangle|^2$
  - Different weights for different parts of the model possible

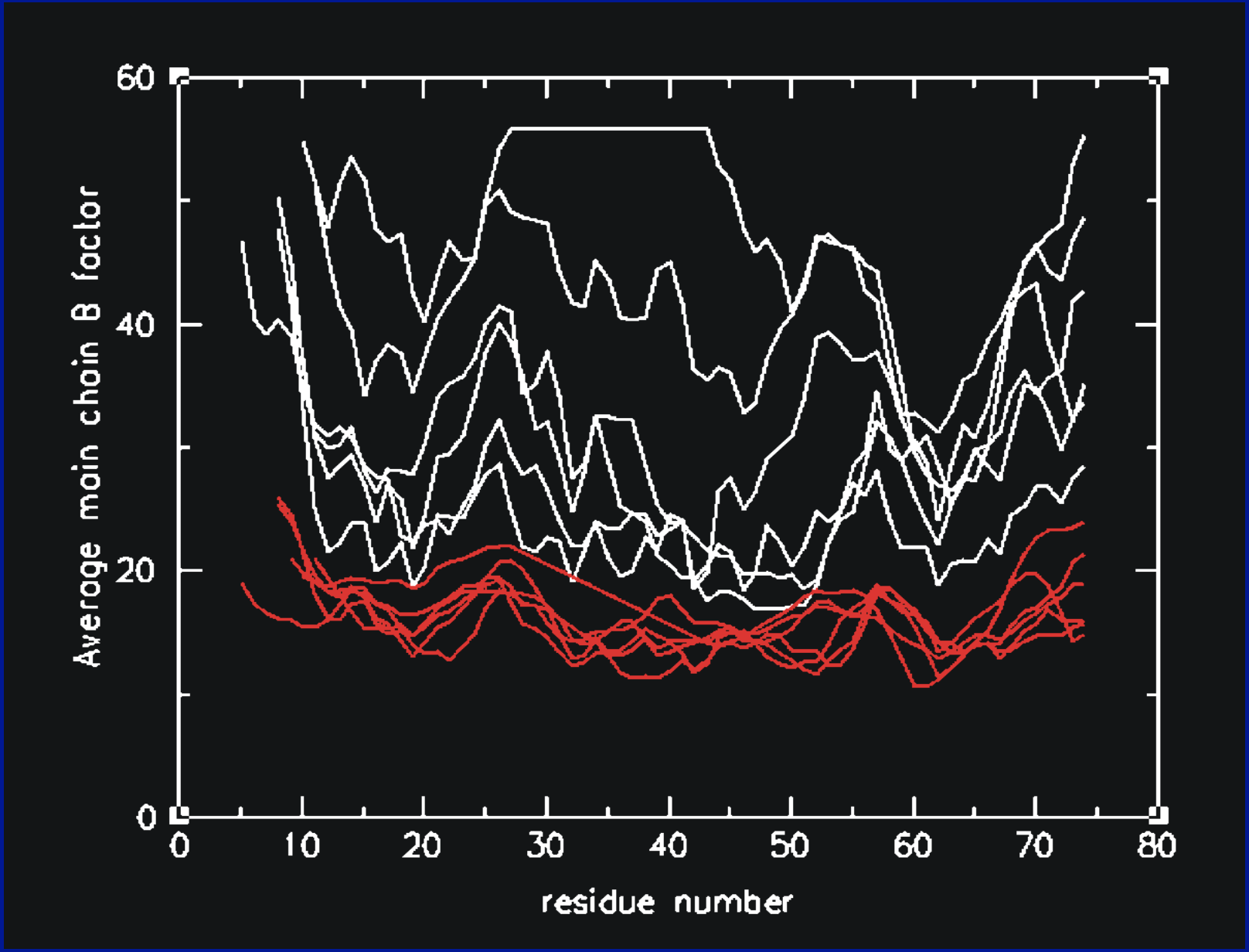


- **NCS (non-crystallographic symmetry) restraints/constraints**
  - Similarly for B-factors: $E = \Sigma_{atoms}\ weight * \Sigma_{NCS} (B - \langle B \rangle)^2$
    - In case when TLS is used, the NCS is applied to $\mathbf{U_{LOCAL}}$

Total ADP: $\mathbf{U}_{TOTAL} = \mathbf{U}_{CRYST} + \mathbf{U}_{GROUP} + \mathbf{U}_{LOCAL}$

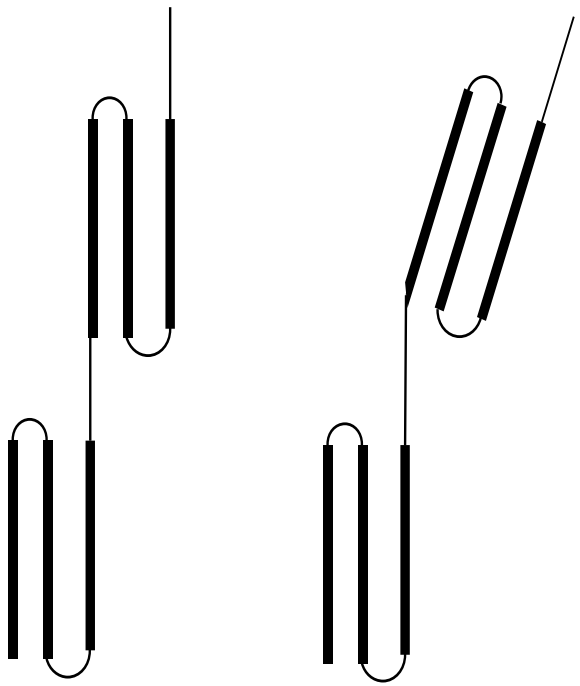


*B*s from NCS related chains



Specific restraints for refinement at low and very low resolution: NCS

- Potential problem when using position-based NCS restraints:
  - Restraining whole will introduce substantial errors (hinge does not obey NCS)



- Solution:
  - Need to use finer-grained NCS groups (in this example treat each domain separately), OR
  - Instead of restraining atomic positions, restrain the orientation of atom with respect to its neighbours ➔ construct restraint target in torsion angle space.

Ramachandran, secondary structure and NCS restraints: when to use ?

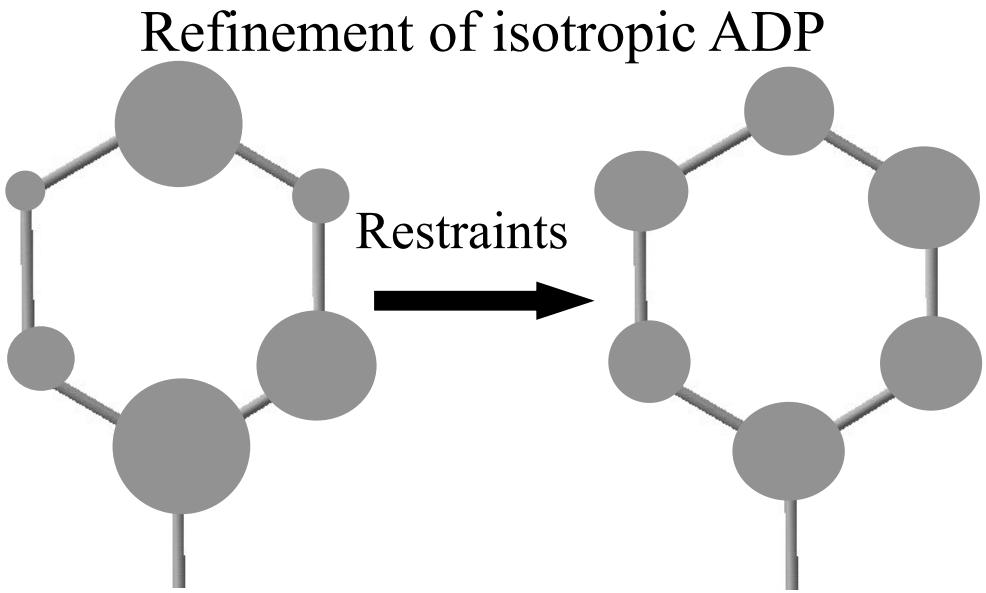
- Ramachandran and secondary structure restraints should be used only at very low resolution(*), when you essentially should use it to assure correctness of your structure (~3-3.5A or even lower, depends on data and model quality)

- NCS restrains:
  - Unlike Ramachandran and secondary-structure, NCS restraints should be used at higher resolution (2A and lower)
  - Some big crystallography names state that NCS should always be used in refinement (if available)
    - This is not quite true: at higher resolution, say lower than 2A, using NCS may rather harm then help, because it may wipe out the naturally occurring differences between NCS-related copies visible at that resolutions
  - Suggestion: simply try refining with and without NCS restraints and see what works better – this is the most robust way to find out!

(*) Urzhumtsev, A., Afonine, P.V. & Adams P.D. (2009). On the use of logarithmic scales for analysis of diffraction data. Acta Cryst. D65, 1283-1291.

Restraints in refinement of individual isotropic ADP

$$E_{TOTAL} = w * E_{DATA} + E_{RESTRAINTS}$$



- Similarity restraints: $E = \Sigma_{\text{all pairs of bonded atoms}}\ weight * (B_i - B_j)^2$

- Knowledge-based restraints: $E = \Sigma_{\text{all pairs of bonded atoms}}\ weight * (|B_i - B_j| - \Delta_{ij})^2$

  where Δ_{ij} comes from a library of values collected from well-trusted structures for given type of atoms.

Restraints in refinement of individual isotropic ADP

$$E_{TOTAL} = w * E_{DATA} + E_{RESTRAINTS}$$

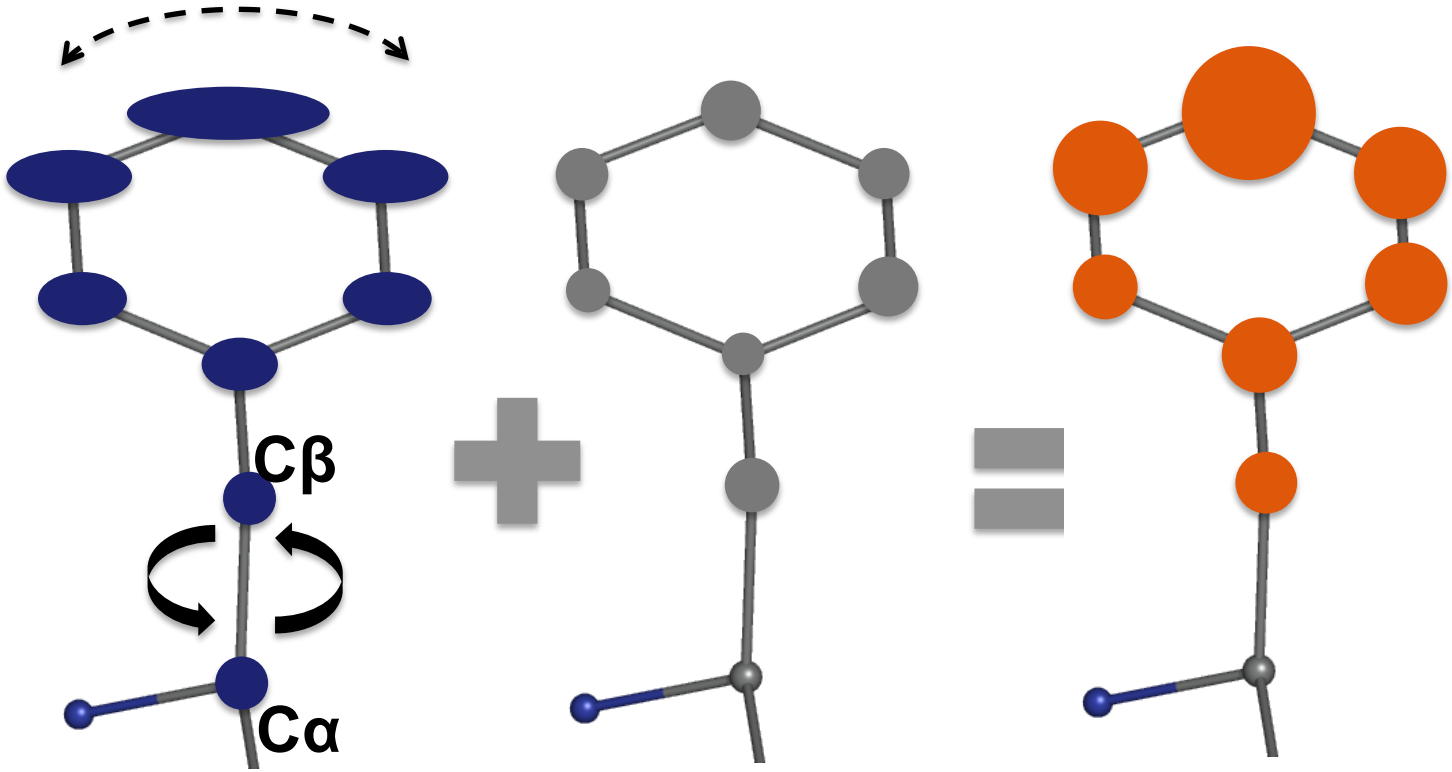
- A better way of defining restraints for isotropic ADPs is based on the following facts:
  - A bond is almost rigid, therefore the ADPs of bonded atoms are similar (Hirshfeld, 1976);
  - ADPs of spatially close (non-bonded) atoms are similar (Schneider, 1996);
  - The difference between the ADPs of bonded atoms, is related to the absolute values of ADPs. Atoms with higher ADPs can have larger differences (Ian Tickle, CCP4 BB, March 14, 2003).

$$\mathbf{E}_{\mathrm{RESTRAINTS}} = \sum_{\mathbf{i=1}}^{\mathbf{N}_{\mathrm{ALL\ ATOMS}}} \left[\sum_{\mathbf{j=1}}^{\mathbf{M}_{\mathrm{ATOMS\ IN\ SPEHRE}}} \frac{\mathbf{1}}{\mathbf{r}_{\mathbf{ij}}^{\mathbf{distance\_power}}} \frac{\left(\mathbf{U_i - U_j}\right)^{\mathbf{2}}}{\left(\frac{\mathbf{U_i + U_j}}{\mathbf{2}}\right)^{\mathbf{average\_power}}} \right|_{\mathbf{sphereR}} \right]$$

- Distance power, average power and sphere radius are some empirical parameters

Restraints in refinement of individual ADP

- **A nuance about using similarity restraints**
  - Total ADP is: $\mathbf{U}_{TOTAL} = \mathbf{U}_{CRYST} + \mathbf{U}_{GROUP} + \mathbf{U}_{LOCAL}$
  - Similarity restraints should be applied to $\mathbf{U}_{LOCAL}$
  - Applying it to $\mathbf{U}_{TOTAL}$ is much less justified



Rigid-body libration around Cα-Cβ bond vector, $\mathbf{U}_{GROUP}$

Small local atomic vibrations, $\mathbf{U}_{LOCAL}$

Resulting isotropic equivalent, $\mathbf{U}_{TOTAL}$

Example of constraints

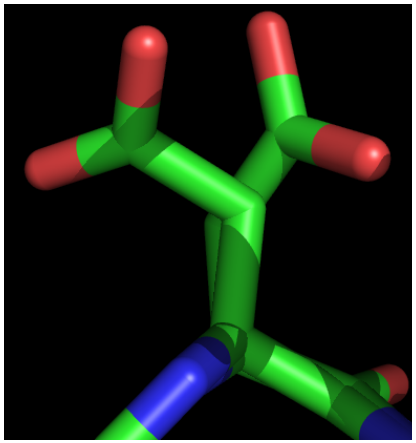
- Rigid body refinement: mutual positions of atoms within a rigid groups are forced to remain the same, while the rigid group can move as a whole. 6 refinable parameters per rigid group (3 translations + 3 rotations).
- Constrained rigid groups: torsion angle parameterization. Reduction of refinable parameters by a factor between 7 and 10.
- Occupancies of atoms in alternative conformations: occupancies of alternate conformers must add up to 1.
- Group ADP refinement: mutual distribution of all B-factors within the group must remain the same. One refinable B-factor per group.
- Constrained NCS refinement: a number of *N* NCS related molecules or domains are assumed to be identical. Reduction of refinable parameters by a factor N.

- Do not confuse restraints and constraints

  Constraints: model property = ideal value

  Restraints: model property ~ ideal value

Constraints in occupancy refinement



- As it stands, occupancy refinement is always a constrained refinement…
- When we do not refine occupancy we essentially constrain its value to whatever value comes from input model (typically 1)

- Refining occupancies of alternative conformations we apply two constraints:
  - Occupancies of atoms within each conformer must be equal
  - Sum of occupancies for each set of matching atoms taken over all conformers must add to 1. Ideally, it should be less than or equal to 1, since we may not be including all existing conformers; however inequality constraints are very hard to handle in refinement.

```
ATOM      1  N   AARG A 192      -5.782  17.932  11.414  0.72  8.38           N
ATOM      2  CA  AARG A 192      -6.979  17.425  10.929  0.72 10.12           C
ATOM      3  C   AARG A 192      -6.762  16.088  10.271  0.72  7.90           C
ATOM      7  N   BARG A 192     -11.719  17.007   9.061  0.28  9.89           N
ATOM      8  CA  BARG A 192     -10.495  17.679   9.569  0.28 11.66           C
ATOM      9  C   BARG A 192      -9.259  17.590   8.718  0.28 12.76           C
```

Refinement target weight (MORE DETAILS)

- Refinement target $E_{TOTAL} = w * E_{DATA} + E_{RESTRAINTS}$
  - the weight w is determined automatically
  - in most of cases the automatic choice is good

- If automatic choice is not optimal there are two possible refinement outcomes:
  - structure is over-refined: *Rfree*-*Rwork* is too large. This means the weight w is too small making the contribution of E_{DATA} too large.
  - weight w is too large making the contribution of restraints too strong. This results increase of *Rfree* and/or *Rwork*.
  - A possible approach to address this problem is to perform a grid search over an array of w values and choose the one w that gives the best *Rfree* and *Rfree*-*Rwork*.
- A random component is involved in w calculation. Therefore an ensemble of identical refinement runs each done using different random seed will result in slightly different structures. The *R*-factor spread depends on resolution and may be as large as 1…2%.

Dictionary

research papers

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

*REFMAC*5 dictionary: organization of prior chemical knowledge and guidelines for its use

Alexei A. Vagin, Roberto A. Steiner,‡ Andrey A. Lebedev, Liz Potterton, Stuart McNicholas, Fei Long and Garib N. Murshudov*

Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5YW, England

‡ Current address: IFOM – The FIRC Institute of Molecular Oncology, Via Adamello 16, 20139 Milano, Italy

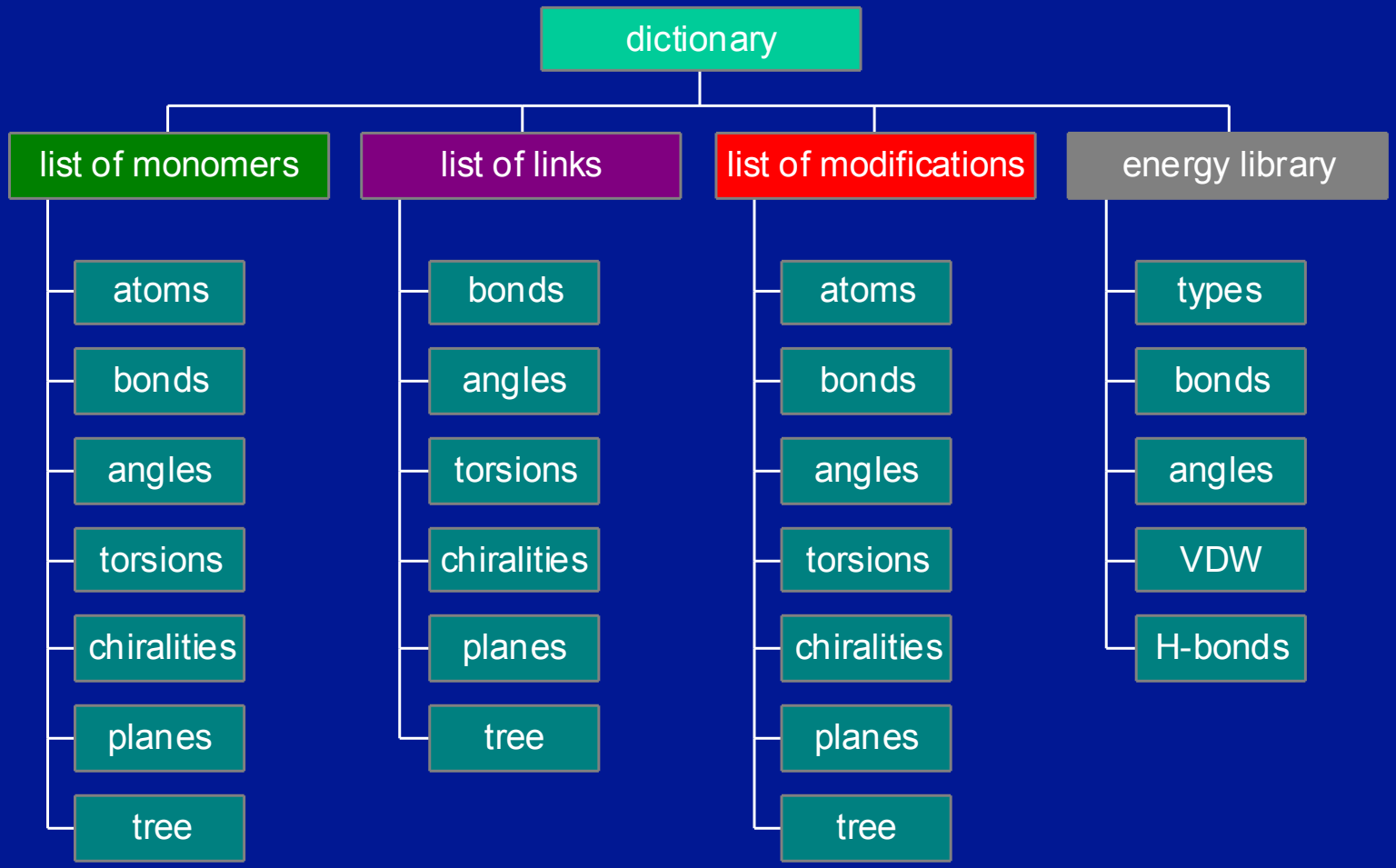
Correspondence e-mail: garib@ysbl.york.ac.uk

One of the most important aspects of macromolecular structure refinement is the use of prior chemical knowledge. Bond lengths, bond angles and other chemical properties are used in restrained refinement as subsidiary conditions. This contribution describes the organization and some aspects of the use of the flexible and human/machine-readable dictionary of prior chemical knowledge used by the maximum-likelihood macromolecular-refinement program *REFMAC*5. The dictionary stores information about monomers which represent the constitutive building blocks of biological macromolecules (amino acids, nucleic acids and saccharides) and about numerous organic/inorganic compounds commonly found in macromolecular crystallography. It also describes the modifications the building blocks undergo as a result of chemical reactions and the links required for polymer formation. More than 2000 monomer entries, 100 modification entries and 200 link entries are currently available. Algorithms and tools for updating and adding new entries to the dictionary have also been developed and are presented here. In many cases, the *REFMAC*5 dictionary allows entirely automatic generation of restraints within *REFMAC*5 refinement runs.

Received 19 April 2004
Accepted 22 September 2004

The use of prior knowledge requires its organised storage.

Organisation of dictionary

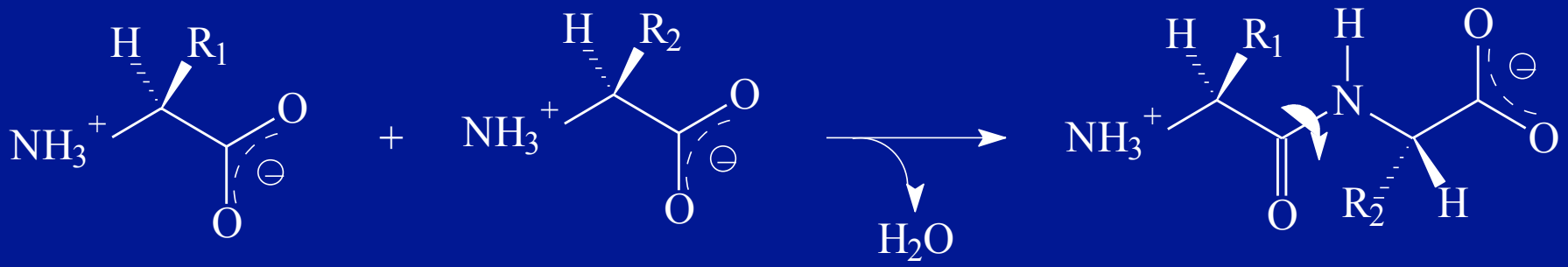


DICTIONARY http://www.ysbl.york.ac.uk/~alexei/dictionary.html

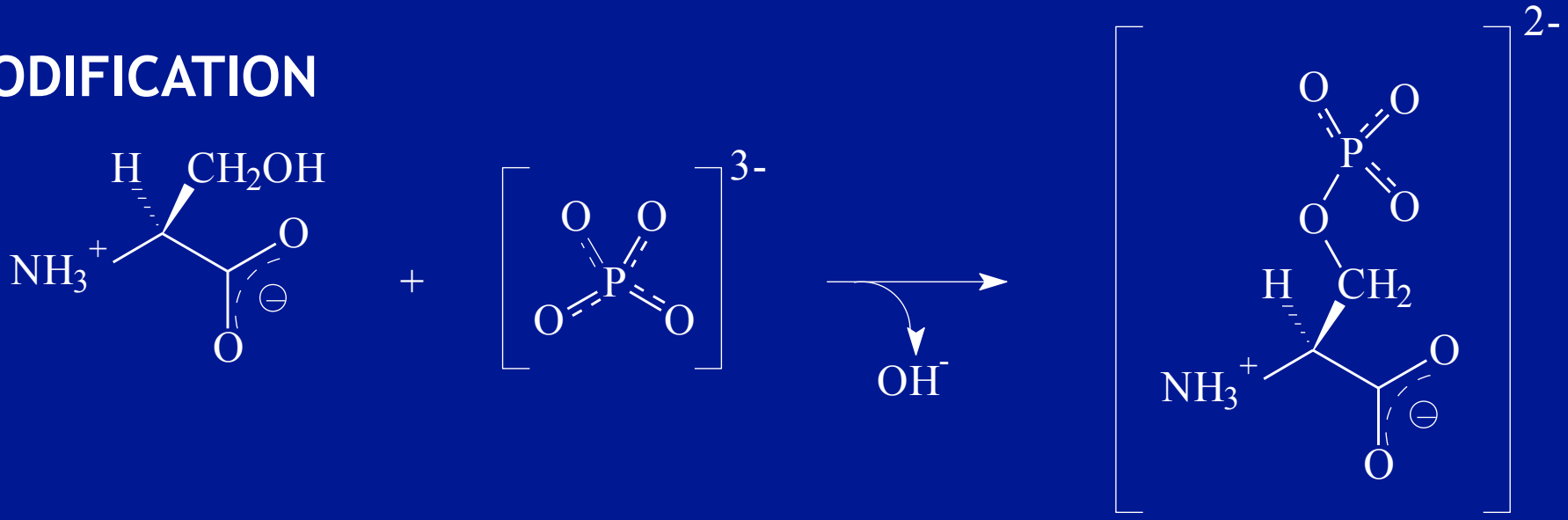
LIBCHECK http://www.ysbl.york.ac.uk/~alexei/libcheck.html

Links and Modifications

LINK



MODIFICATION



Description of monomers

In the files:

a/A##.cif

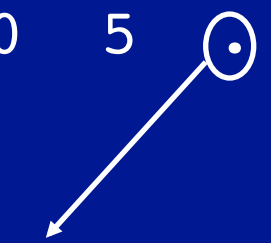
Monomers are described by the following catagories:

```
_chem_comp
_chem_comp_atom
_chem_comp_bond
_chem_comp_angle
_chem_comp_tor
_chem_comp_chir
_chem_comp_plane_atom
```

Monomer library (_chem_comp)

```
loop_
_chem_comp.id
_chem_comp.three_letter_code
_chem_comp.name
_chem_comp.group
_chem_comp.number_atoms_all
_chem_comp.number_atoms_nh
_chem_comp.desc_level

ALA    ALA    'ALANINE '    L-peptide     10   5   .
```



Level of description
. = COMPLETE
M = MINIMAL

Monomer library (_chem_comp_atom)

```
loop_
_chem_comp_atom.comp_id
_chem_comp_atom.atom_id
_chem_comp_atom.type_symbol
_chem_comp_atom.type_energy
_chem_comp_atom.partial_charge
 ALA       N     N     NH1        -0.204
 ALA       H     H     HNH1        0.204
 ALA       CA    C     CH1         0.058
 ALA       HA    H     HCH1        0.046
 ALA       CB    C     CH3        -0.120
 ALA       HB1   H     HCH3        0.040
 ALA       HB2   H     HCH3        0.040
 ALA       HB3   H     HCH3        0.040
 ALA       C     C     C           0.318
 ALA       O     O     O          -0.422
```

Monomer library (_chem_comp_bond)

```
loop_
_chem_comp_bond.comp_id
_chem_comp_bond.atom_id_1
_chem_comp_bond.atom_id_2
_chem_comp_bond.type
_chem_comp_bond.value_dist
_chem_comp_bond.value_dist_esd
 ALA       N    H       single      0.860    0.020
 ALA       N    CA      single      1.458    0.019
 ALA       CA   HA      single      0.980    0.020
 ALA       CA   CB      single      1.521    0.033
 ALA       CB   HB1     single      0.960    0.020
 ALA       CB   HB2     single      0.960    0.020
 ALA       CB   HB3     single      0.960    0.020
 ALA       CA   C       single      1.525    0.021
 ALA       C    O       double      1.231    0.020
```

Monomer library (_chem_comp_chir)

```
loop_
_chem_comp_chir.comp_id
_chem_comp_chir.id
_chem_comp_chir.atom_id_centre
_chem_comp_chir.atom_id_1
_chem_comp_chir.atom_id_2
_chem_comp_chir.atom_id_3
_chem_comp_chir.volume_sign
ALA  chir_01  CA  N  CB  C    negativ
```

positiv, negativ, both, anomer

Current status of the dictionary

Currently, there are about

- 9000 monomers with a complete description
- 100 modifications
- 200 links

Cis-peptides, S-S bridges, sugar-, DNA-, RNA-links are automatically recognized.

What happens when you run *REFMAC*5?

You have only monomers for which there is a complete description

the program carries on and takes everything from the dictionary

You have a monomer for which there is no description (or only a minimal description)

Minimal description or no description

In the case you have monomer(s) in your coordinate file for which there is no description (or minimal description) ***REFMAC5*** **generates for you a complete library description (monomer.cif) and then it stops so you can check the result.**

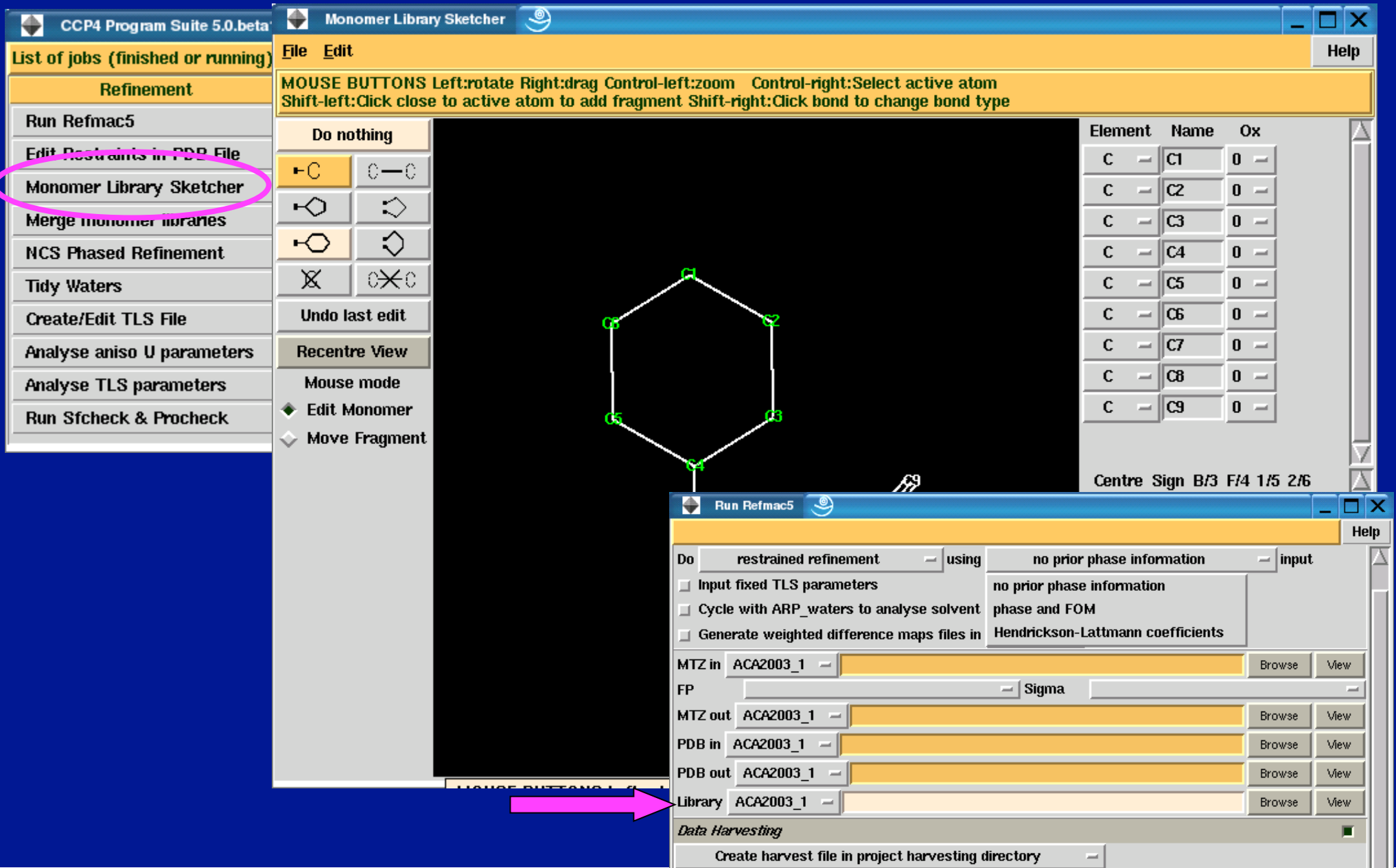
If you are satisfied you can use monomer.cif for refinement. The description generated in this way is good only if your coordinates are good (CSD, EBI, any program that can do energy minimization).

A more general approach for description generation requires the use of the graphical program *SKETCHER* from CCP4*i*. *SKETCHER* is a graphical interface to *LIBCHECK*.

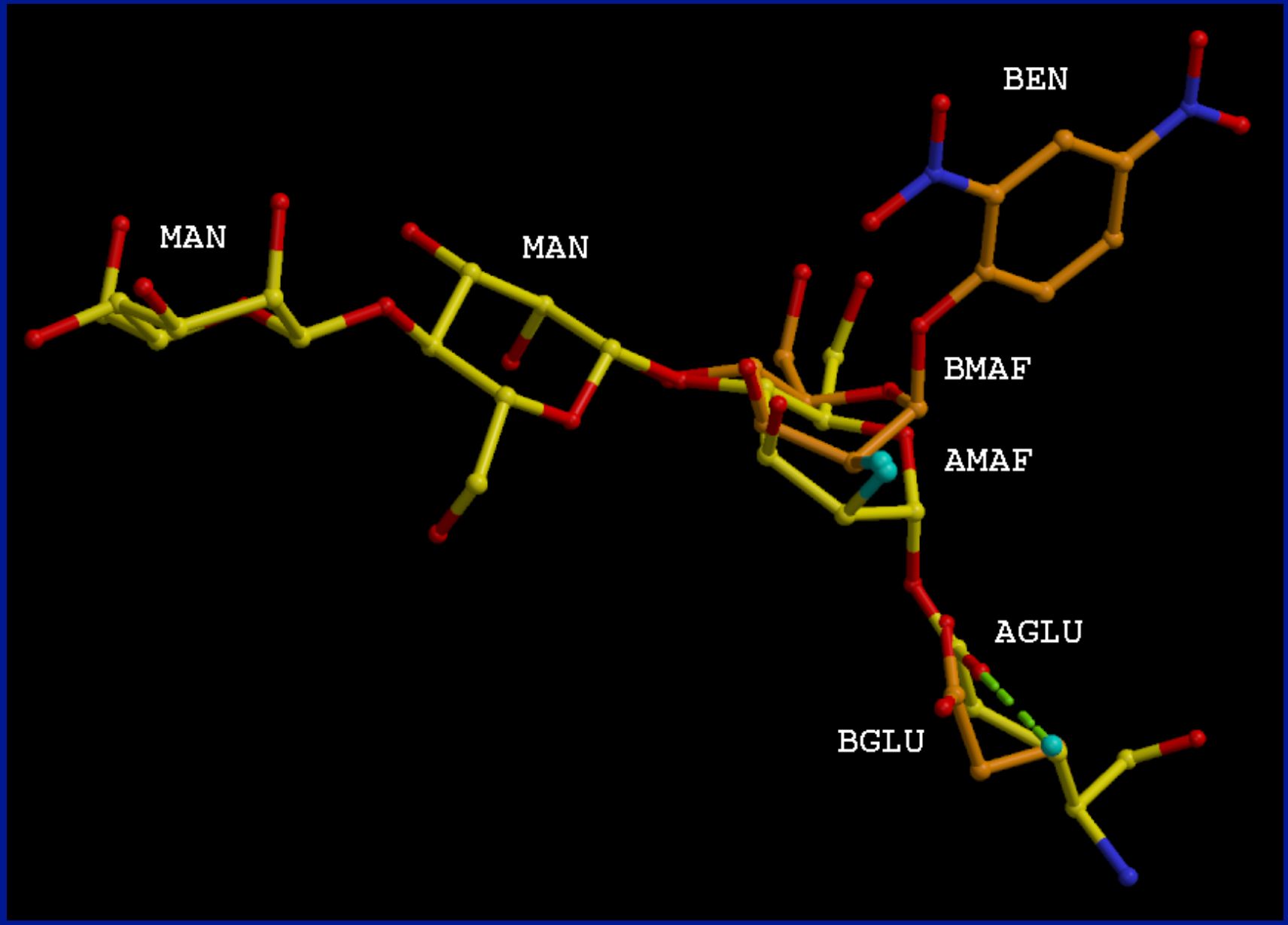
Alternatively, you can use the *PRODRG2* server
http://davapc1.bioch.dundee.ac.uk/programs/prodrg/prodrg.html

Even better use the GRADE server (Global Phasing)
http://grade.globalphasing.org/cgi-bin/grade/server.cgi

SKETCHER


CCP4 Program Suite 5.0.beta
List of jobs (finished or running)
Refinement
Run Refmac5
Monomer Library Sketcher
NCS Phased Refinement
Tidy Waters
Create/Edit TLS File
Analyse aniso U parameters
Analyse TLS parameters
Run Sfcheck & Procheck
Monomer Library Sketcher
File Edit
Help
MOUSE BUTTONS Left:rotate Right:drag Control-left:zoom Control-right:Select active atom
Shift-left:Click close to active atom to add fragment Shift-right:Click bond to change bond type
Do nothing
Undo last edit
Recentre View
Mouse mode
Edit Monomer
Move Fragment
Element Name Ox
C C1 0
C C2 0
C C3 0
C C4 0
C C5 0
C C6 0
C C7 0
C C8 0
C C9 0
Centre Sign B/3 F/4 1/5 2/6
Run Refmac5
Help
Do restrained refinement using no prior phase information input
Input fixed TLS parameters
Cycle with ARP_waters to analyse solvent
Generate weighted difference maps files in
no prior phase information
phase and FOM
Hendrickson-Lattmann coefficients
MTZ in ACA2003_1 Browse View
FP Sigma
MTZ out ACA2003_1 Browse View
PDB in ACA2003_1 Browse View
PDB out ACA2003_1 Browse View
Library ACA2003_1 Browse View
Data Harvesting
Create harvest file in project harvesting directory

REFMAC5 can handle complex chemistry



Links and Modifications in practice

At the top of the PDB file:

```
0         1         2         3         4         5         6         7
12345678901234567890123456789012345678901234567890123456789012345678901234567890
LINK         C6   BBEN B   1                 O1   BMAF S   2                   BEN-MAF
LINK         OE2  GLU A  67         1.895    ZN     ZN R   5                   GLU-ZN
LINK              GLY H 127                       GLY H 133                   gap
LINK              MAF S   2                       MAN S   3                   BETA1-4
SSBOND   1 CYS A  298     CYS A  298                          4555
MODRES     MAN S     3   MAN-b-D                                               RENAME
```

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

JLigand: a graphical tool for the *CCP4* template-restraint library

Andrey A. Lebedev,[a]* Paul Young,[b] Michail N. Isupov,[c] Olga V. Moroz,[d] Alexey A. Vagin[d] and Garib N. Murshudov[e]

[a]CCP4, STFC Rutherford Appleton Laboratory, Harwell Oxford, Didcot OX11 0QX, England, [b]York Digital Library, University of York, Heslington, York YO10 5DD, England, [c]Henry Wellcome Building for Biocatalysis, Biosciences, College of Life and Environmental Sciences, University of Exeter, Stocker Road, Exeter EX4 4QD, England, [d]Structural Biology Laboratory, University of York, Heslington, York YO10 5DD, England, and [e]Structural Studies Division, MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, England

Correspondence e-mail: andrey.lebedev@stfc.ac.uk

Biological macromolecules are polymers and therefore the restraints for macromolecular refinement can be subdivided into two sets: restraints that are applied to atoms that all belong to the same monomer and restraints that are associated with the covalent bonds between monomers. The *CCP4* template-restraint library contains three types of data entries defining template restraints: descriptions of monomers and their modifications, both used for intramonomer restraints, and descriptions of links for intermonomer restraints. The library provides generic descriptions of modifications and links for protein, DNA and RNA chains, and for some post-translational modifications including glycosylation. Structure-specific template restraints can be defined in a user's additional restraint library. Here, *JLigand*, a new *CCP4* graphical interface to *LibCheck* and *REFMAC* that has been developed to manage the user's library and generate new monomer entries is described, as well as new entries for links and associated modifications.

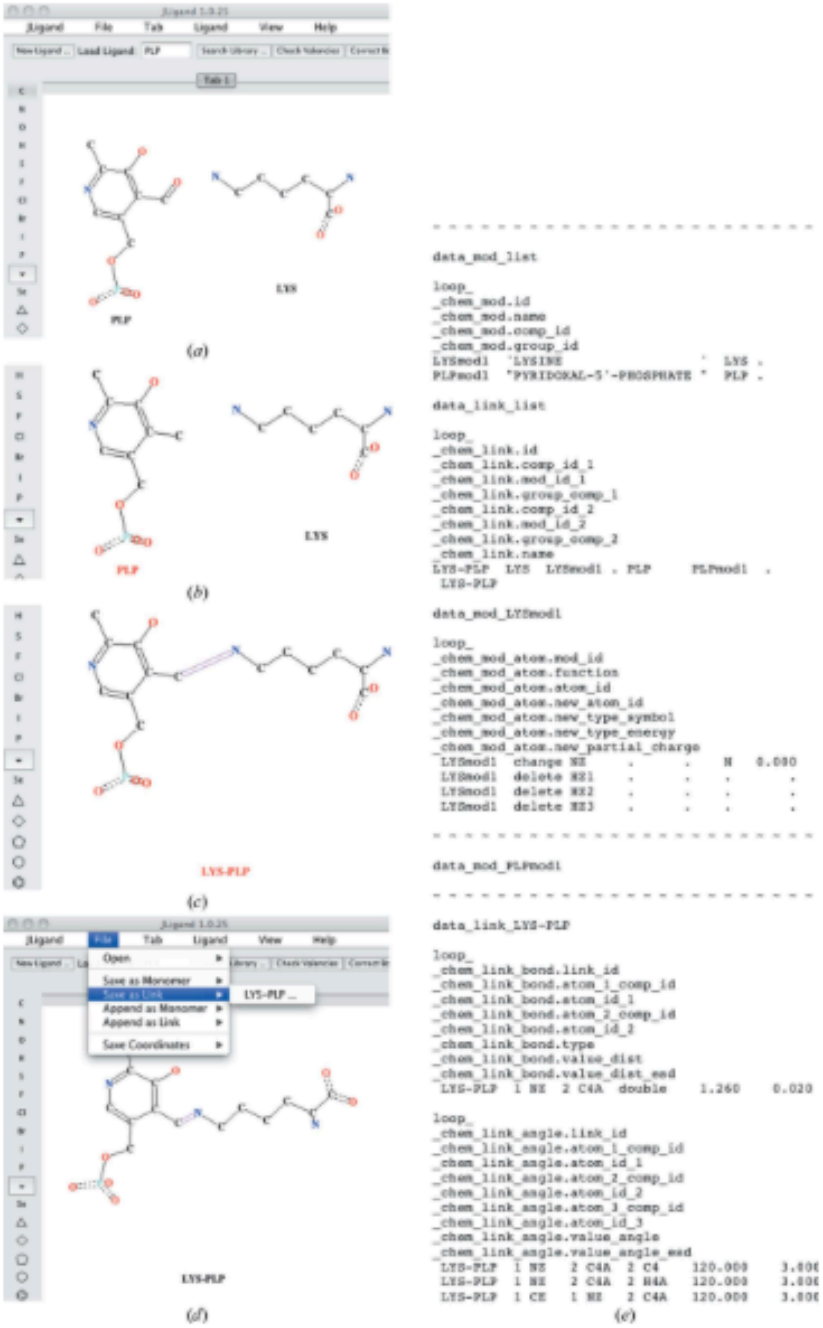


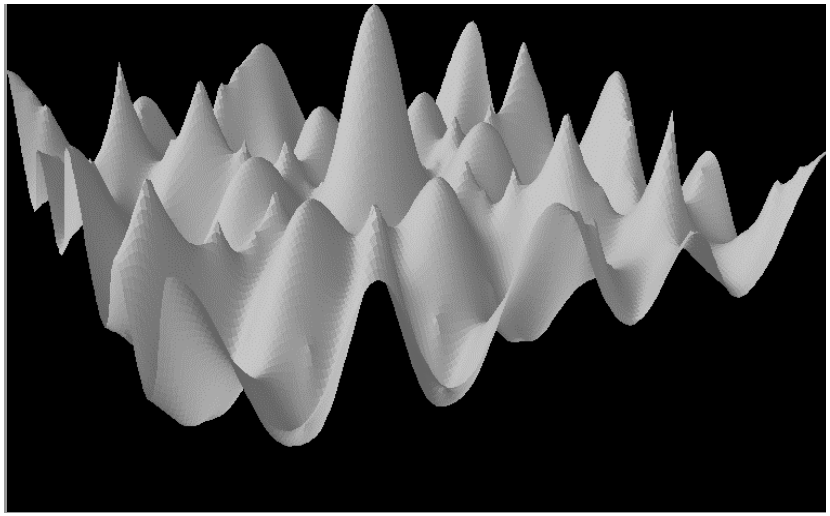
Figure 4
Defining restraints for the covalent linkage Lys–PLP using *JLigand*. Snapshots show the state of the *JLigand* interface after the following user actions: (*a*) loading monomers LYS and PLP from the restraint library, (*b*) removing O4A from PLP, (*c*) connecting C4A (PLP) with NZ (LYS) and defining the bond order and (*d*) regularization. The Save As Link menu has to be used to generate (*e*) an additional library file containing the following five data blocks: list of modifications, list of links, modifications LYSmod1 and PLPmod1 to be applied to LYS and PLP, respectively, and link LYS–PLP. The file header, tables of restraints for LYSmod1 and all the tables for PLPmod1 are omitted, as indicated by lines filled by tildes. H atoms are dealt with automatically. It is also possible to visualize them and handle them explicitly.

Key aspects of refinement

- Objective function
- Method of optimization
- Model parametrization
- Prior knowledge

Refinement convergence

- Landscape of a refinement function is very complex



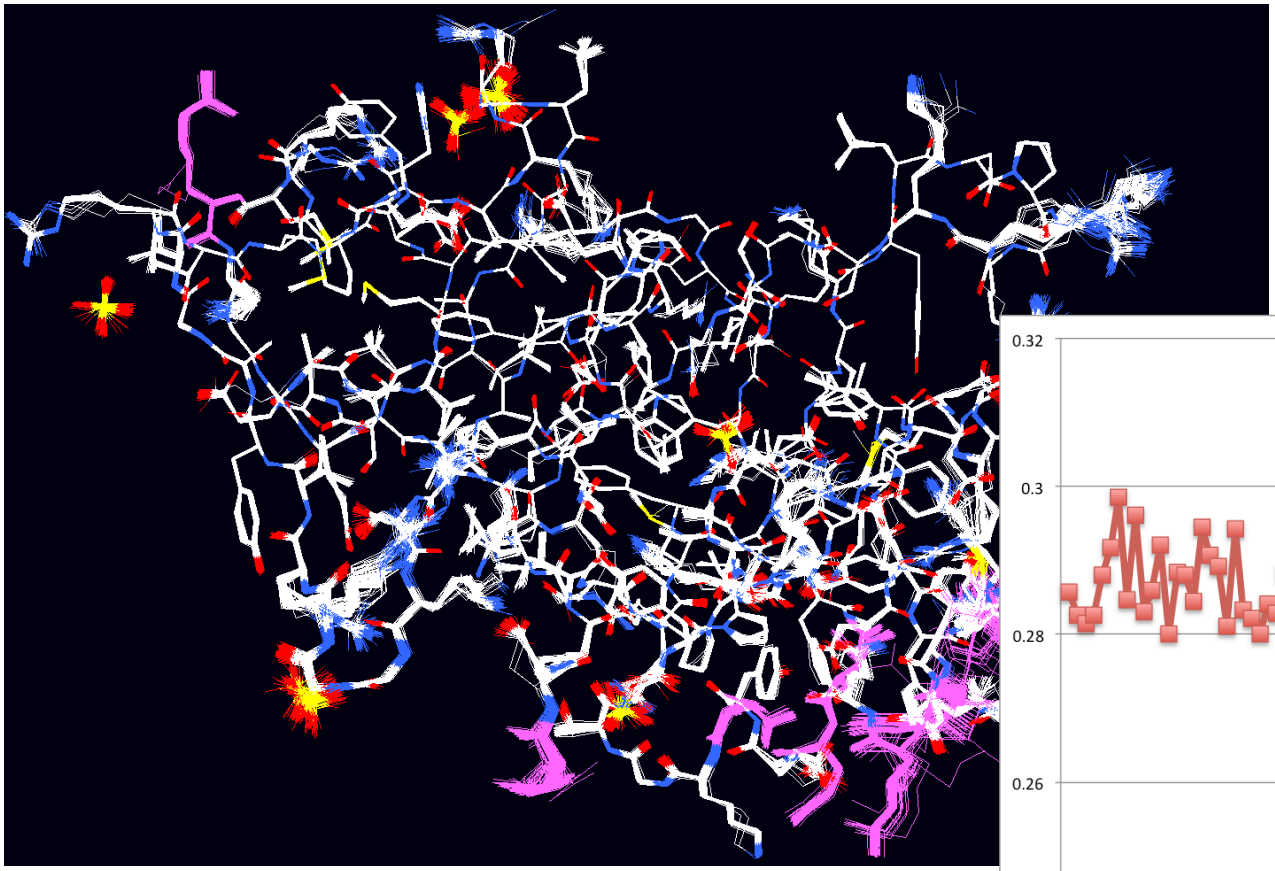
Picture stolen from Dale Tronrud

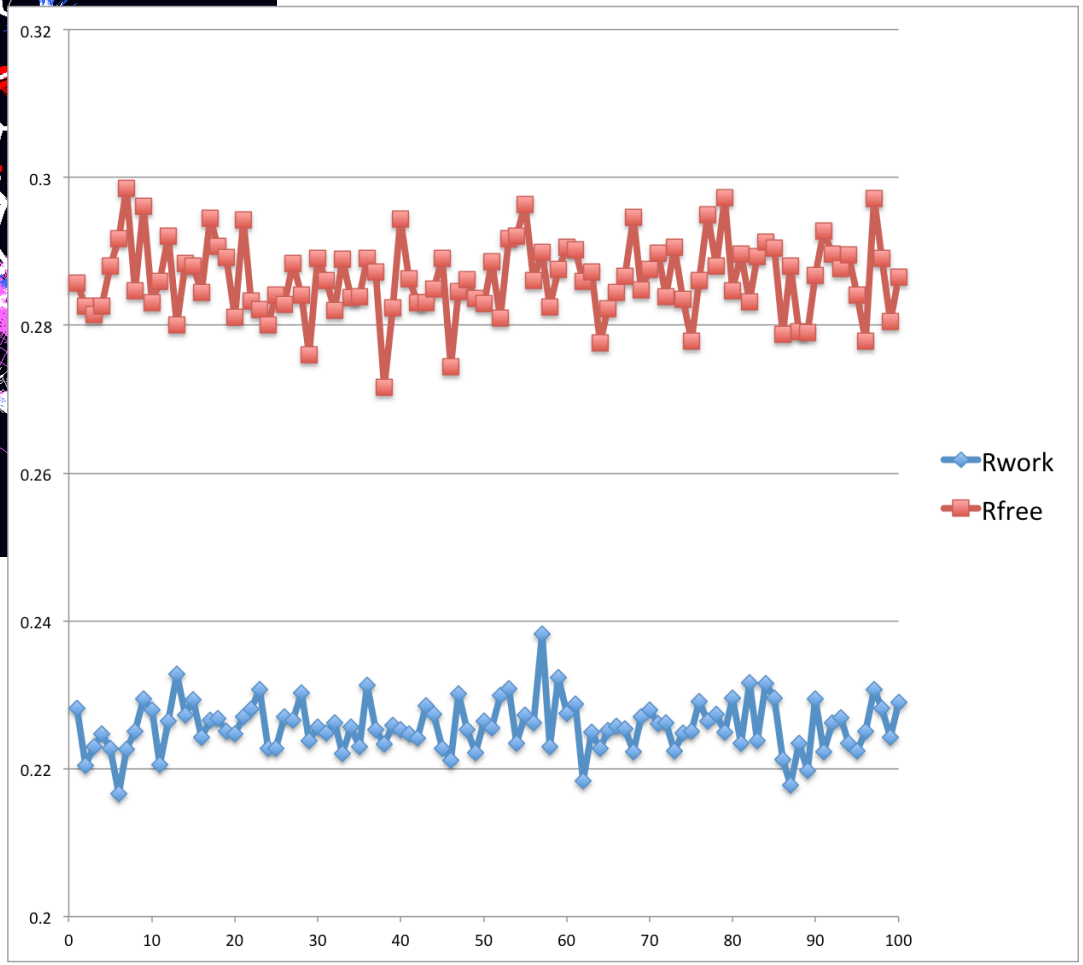
- Refinement programs have very small convergence radii compared to the size of the function profile
  - Depending where you start, the refinement engine will bring the structure to one of the closest local minimum

- What does it mean in practice ? Let's do the following experiment: run 100 identical Simulate Annealing refinement jobs, each staring with different random seed…

Refinement convergence

- As result we get an ensemble of slightly different structures having small deviations in atomic positions, B-factors, etc… R-factors deviate too.





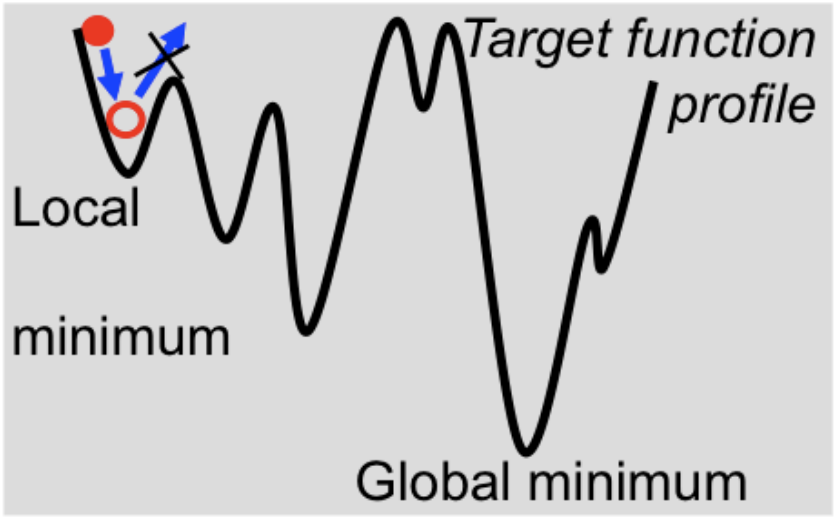
Refinement convergence

- Interpretation of the ensemble:
  - The variation of the structures in the ensemble reflects:
    - Refinement artifacts (limited convergence radius and speed)
    - Some structural variations
  - Spread between the refined structures is the function of resolution (lower the resolution – higher the spread), and the differences between initial structures
  - Obtaining such ensemble is very useful in order to asses the degree of uncertainty the comes from refinement alone

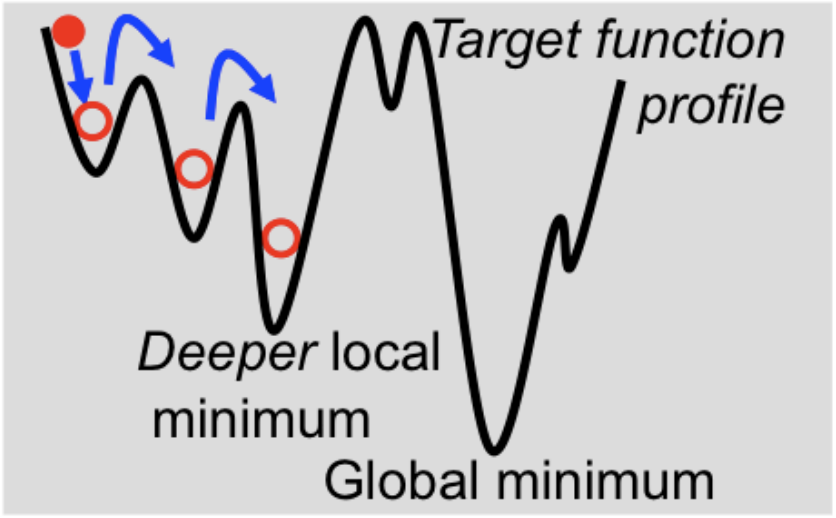
It is not uncommon to find that structures characterized by small differences in *R* statistics have essentially the same information content. Biology is more robust than R factors.

Refinement target optimization methods

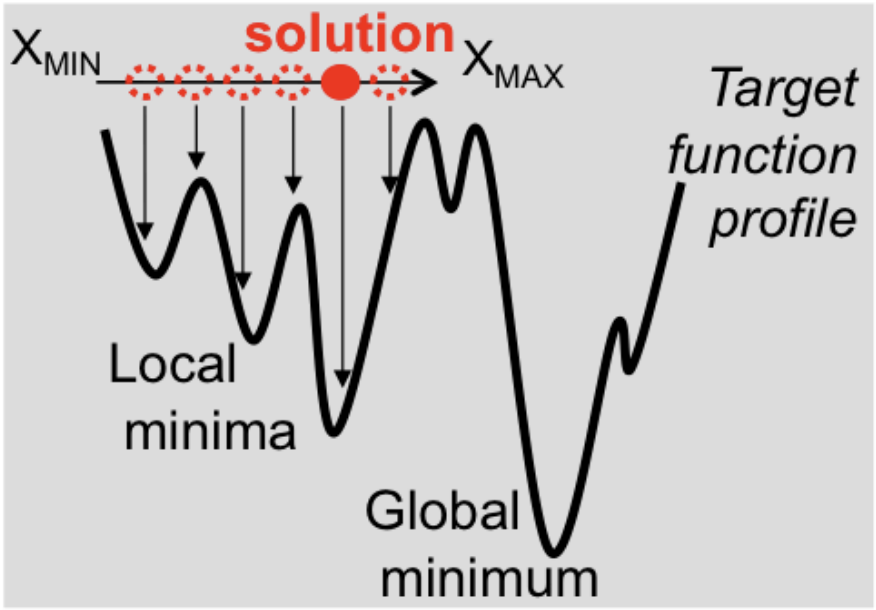
- **Gradient-driven minimization**



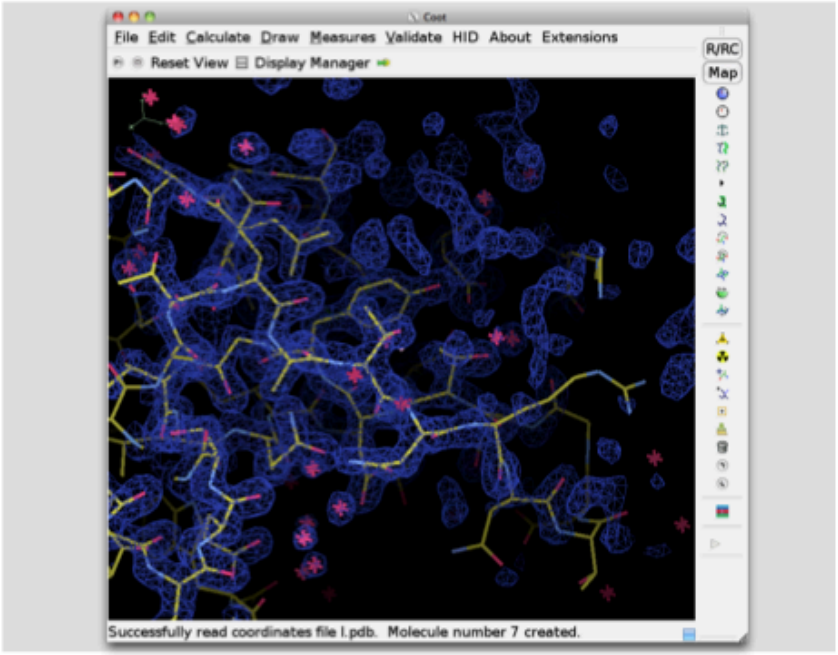
- **Simulated annealing (SA)**



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



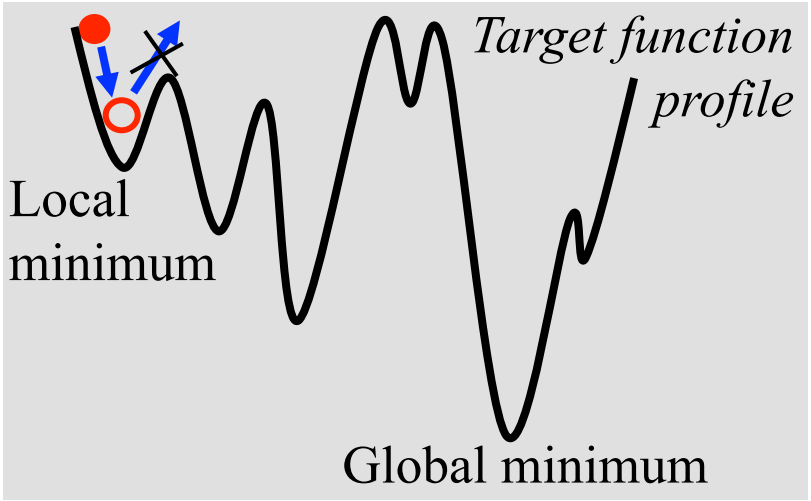
- **Hands & eyes** (Via Coot)



Refinement target optimization methods

- **Gradient-driven minimization**

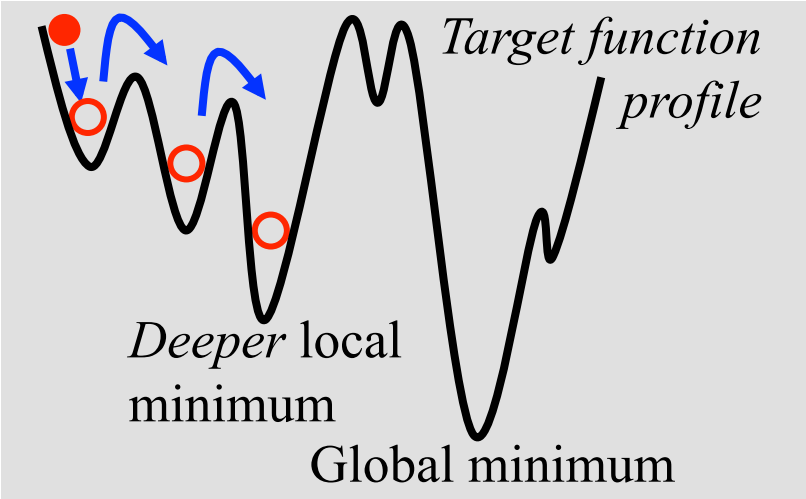
- Follows the local gradient.
- The target function depends on many parameters – many local minima.



Refinement target optimization methods

- **Simulated annealing (SA)**

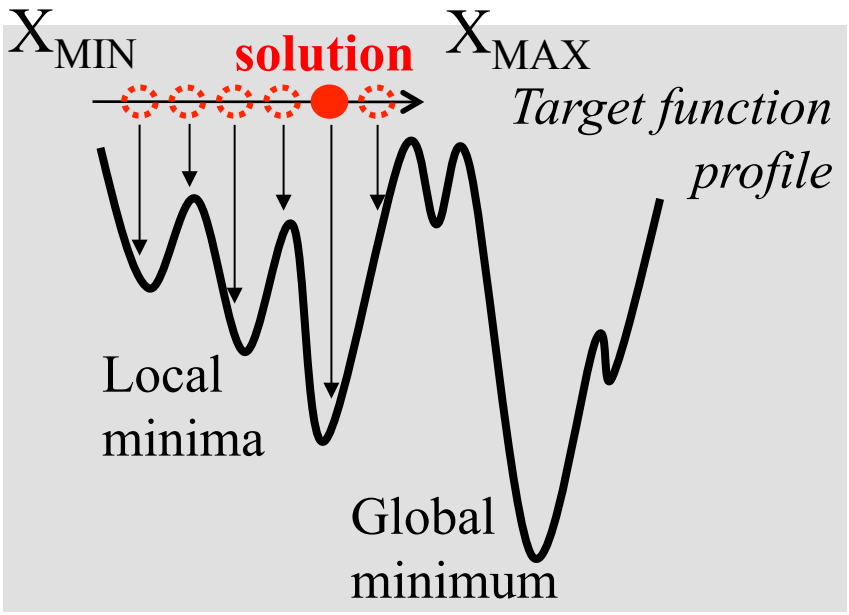
- SA is an optimization method which is good at escaping local minima.
- Annealing is a physical process where a solid is heated until all particles are in a liquid phase, followed by cooling which allows the particles to move to the lowest energy state.
- Simulated annealing is the simulation of the annealing process.
  - Increased probability of finding a better solution because motion against the gradient is allowed.
  - Probability of uphill motion is determined by the temperature.



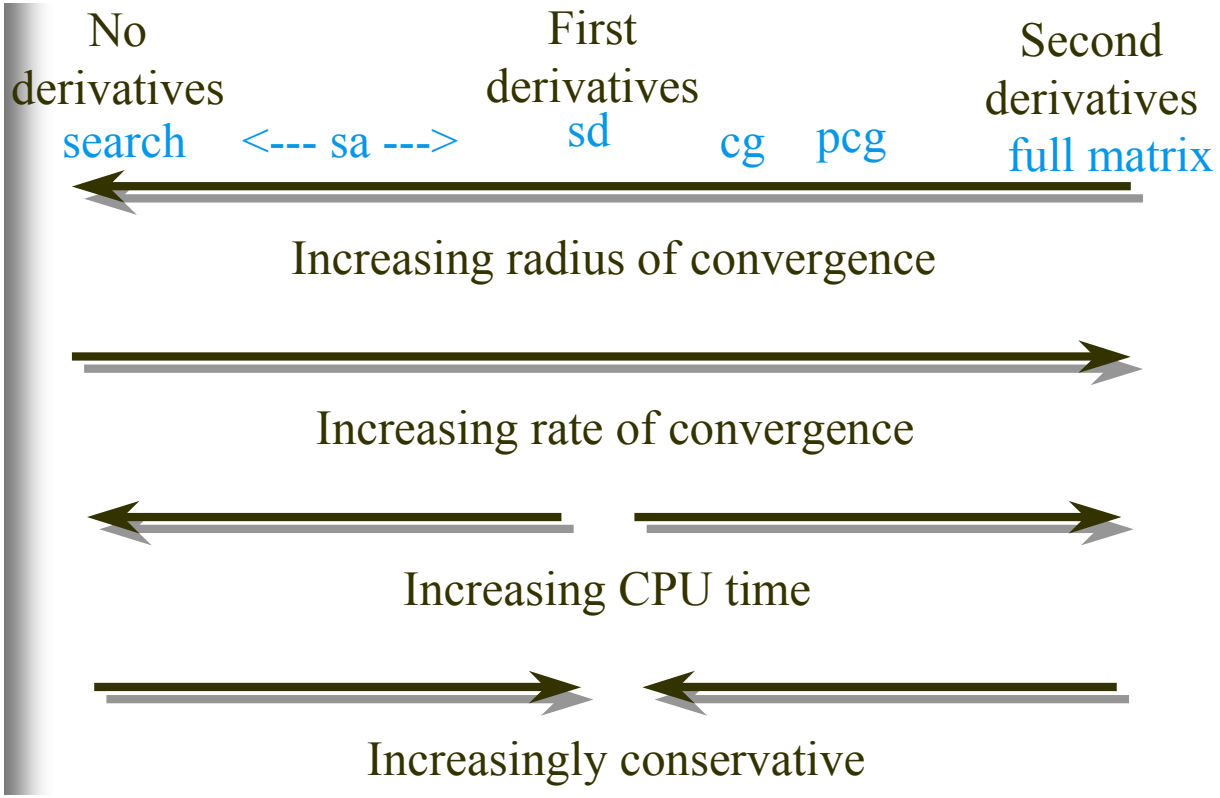
Refinement target optimization methods

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

Robust but may be time inefficient for many parameter systems, and not as accurate as gradient-driven. Good for small number of parameters (1-3 or so), and impractical for larger number of parameters.



Summary on optimization tools



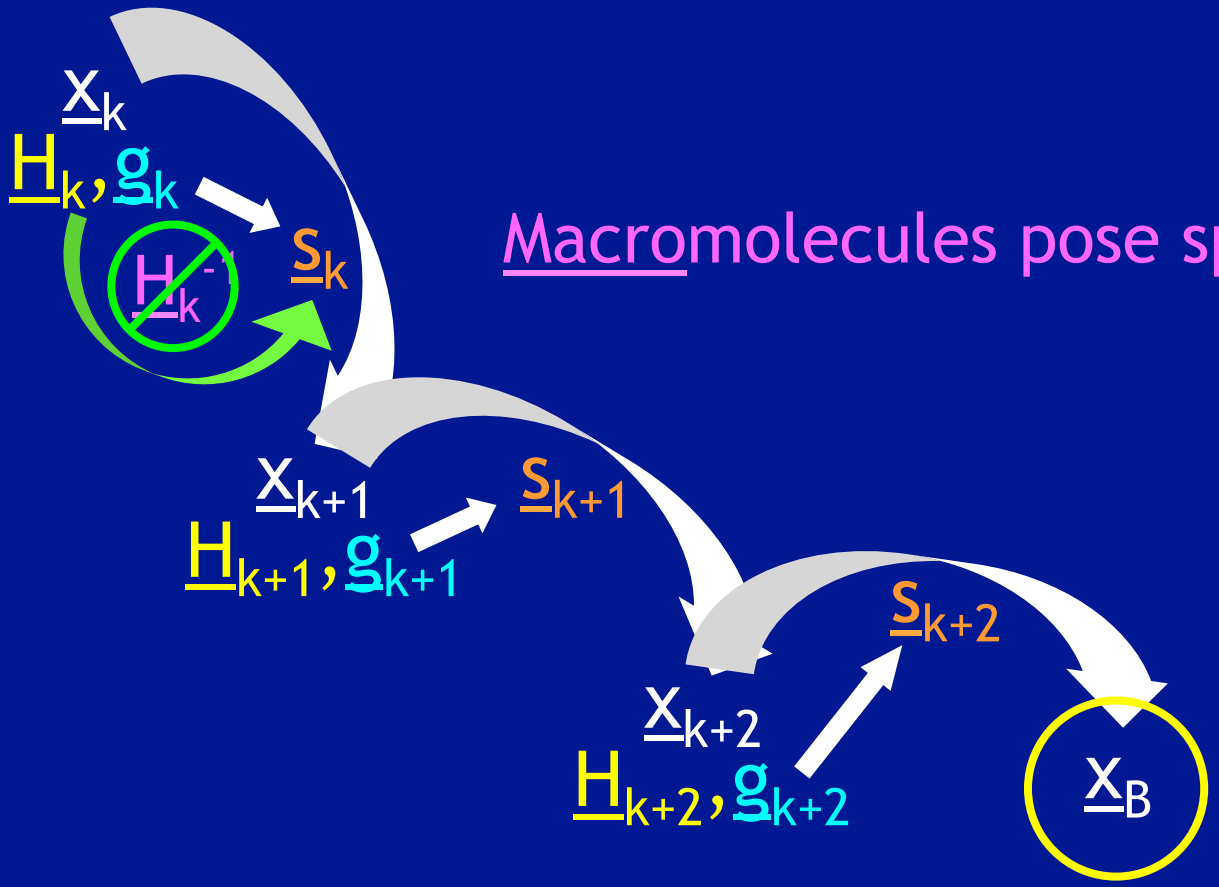
Picture stolen from Dale Tronrud

Newton's method

Taylor expansion of the objective function $f(\underline{x})$ around a working $\underline{x}_k$

$$\underline{H}\underline{s} = -\underline{g}$$

$\underline{H}$	Second-derivative matrix of $f(\underline{x})$
$\underline{s}$	Shift vector to be applied to $\underline{x}_k$
$\underline{g}$	Gradient of $f(\underline{x})$



Macromolecules pose special problems

Macromolecules

The calculation and storage of $\underline{H}$ ($\underline{H}^{-1}$) is very expensive

$$\begin{pmatrix} \frac{\partial^2 f}{\partial p_1 \partial p_1} & \cdots & \frac{\partial^2 f}{\partial p_1 \partial p_{10000}} \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & \vdots \\ \frac{\partial^2 f}{\partial p_{10000} \partial p_1} & \cdots & \frac{\partial^2 f}{\partial p_{10000} \partial p_{10000}} \end{pmatrix}$$

$\underline{H}$ in isotropic refinement has $4N \times 4N$ elements
2500 atoms → 100 000 000 elements

$\underline{H}$ in anisotropic refinement has $9N \times 9N$ elements
2500 atoms → 506 250 000 elements

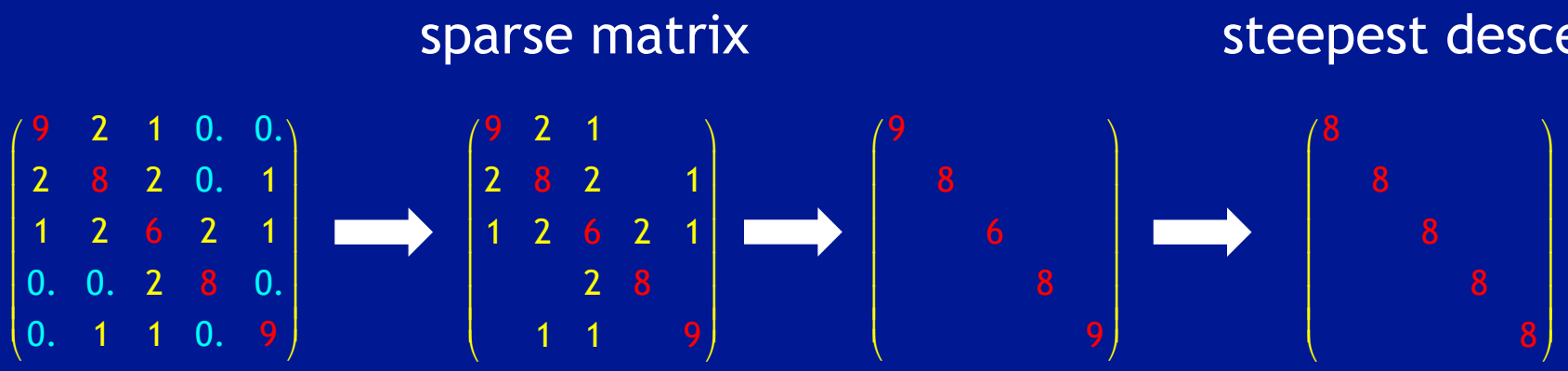
Direct calculation $\quad$ time $\propto N_{el} \times N_{refl}$

FFT methods $\quad$ time $\propto c_1 N_{el} + c_2 N_{refl} \log N_{refl}$

[Agarwal, 1978]
[Murshudov et al., 1997]
[Tronrud, 1999]
[Urzhumtsev & Lunin, 2001]

Approximations

The magnitude of matrix elements decreases with the lenghtening of the interatomic distance



*REFMAC*5 uses the scoring method of minimisation

The objective function $f(\underline{x})$ is the likelihood L

$$\underline{H}\underline{s}=-\underline{g} \longrightarrow \underline{I}\underline{s}=-\underline{g}$$

$$\underline{g}=\frac{\partial L}{\partial \underline{p}}$$ Score vector

$$\underline{H}=\frac{\partial^2 L}{\partial \underline{p}\partial \underline{p}^T}$$ Observed information matrix

$$\underline{I}=\left\langle \frac{\partial^2 L}{\partial \underline{p}\partial \underline{p}^T} \right\rangle$$ Fisher's information matrix

$$\underline{I}=\left\langle \underline{g}(\underline{p})\underline{g}(\underline{p})^T \right\rangle$$

$$\langle \xi \rangle = \int \ldots \int \xi e^{-L} \mathrm{d}o_1 \ldots \mathrm{d}o_n$$

Positive semidefinite

REFMAC5 uses the scoring method of minimisation

research papers

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Fisher's information in maximum-likelihood macromolecular crystallographic refinement

Roberto A. Steiner, Andrey A. Lebedev and Garib N. Murshudov*

Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5YW, England

Correspondence e-mail: garib@ysbl.york.ac.uk

Received 13 June 2003
Accepted 21 August 2003

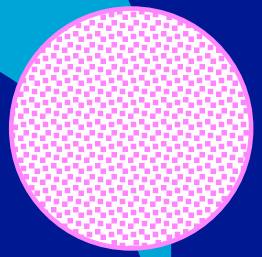
Fisher's information is a statistical quantity related to maximum-likelihood theory. It is a matrix defined as the expected value of the squared gradient of minus the log-likelihood function. This matrix is positive semidefinite for any parameter value. Fisher's information is used in the quasi-Newton scoring method of minimization to calculate the shift vectors of model parameters. If the matrix is non-singular, the scoring-minimization step is always downhill. In this article, it is shown how the scoring method can be applied to macromolecular crystallographic refinement. It is also shown how the computational costs involved in calculation of the Fisher's matrix can be efficiently reduced. Speed is achieved by assuming a continuous distribution of reciprocal-lattice points. Matrix elements calculated with this method agree very well with those calculated analytically. The scoring algorithm has been implemented in the program *REFMAC5* of the *CCP4* suite. The Fisher's matrix is used in its sparse approximation. Tests indicate that the algorithm performs satisfactorily.

Properties of the scoring method

- As Fisher's information is positive semidefinite for any parameter value the shift s is always downhill (if the matrix is non-singular)
- The scoring method is linearly convergent at a rate which depends on the relative difference between the observed and expected information [Smyth, 1996]
- In short runs the scoring method often converges faster than Newton's method especially if the number of observations is big [Kale, 1961]
- Fisher's information is easier to calculate than the Hessian

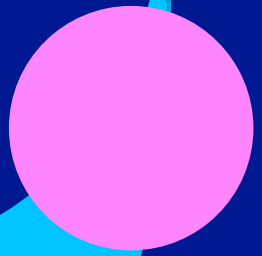
Integral approximation of *I*

$$I_{p_i(n)p_j(m)} \cong K_{p_ip_j} q_n q_m \sum_{\underline{h}} w_s H_{p_ip_j} f_n^0 f_m^0 t_n t_m \mathrm{trig}_{p_ip_j}\left(2\pi \underline{h}\underline{D}_{nm}\right)$$



Discrete reciprocal space

I depends on atom types, ADPs, interatomic distance

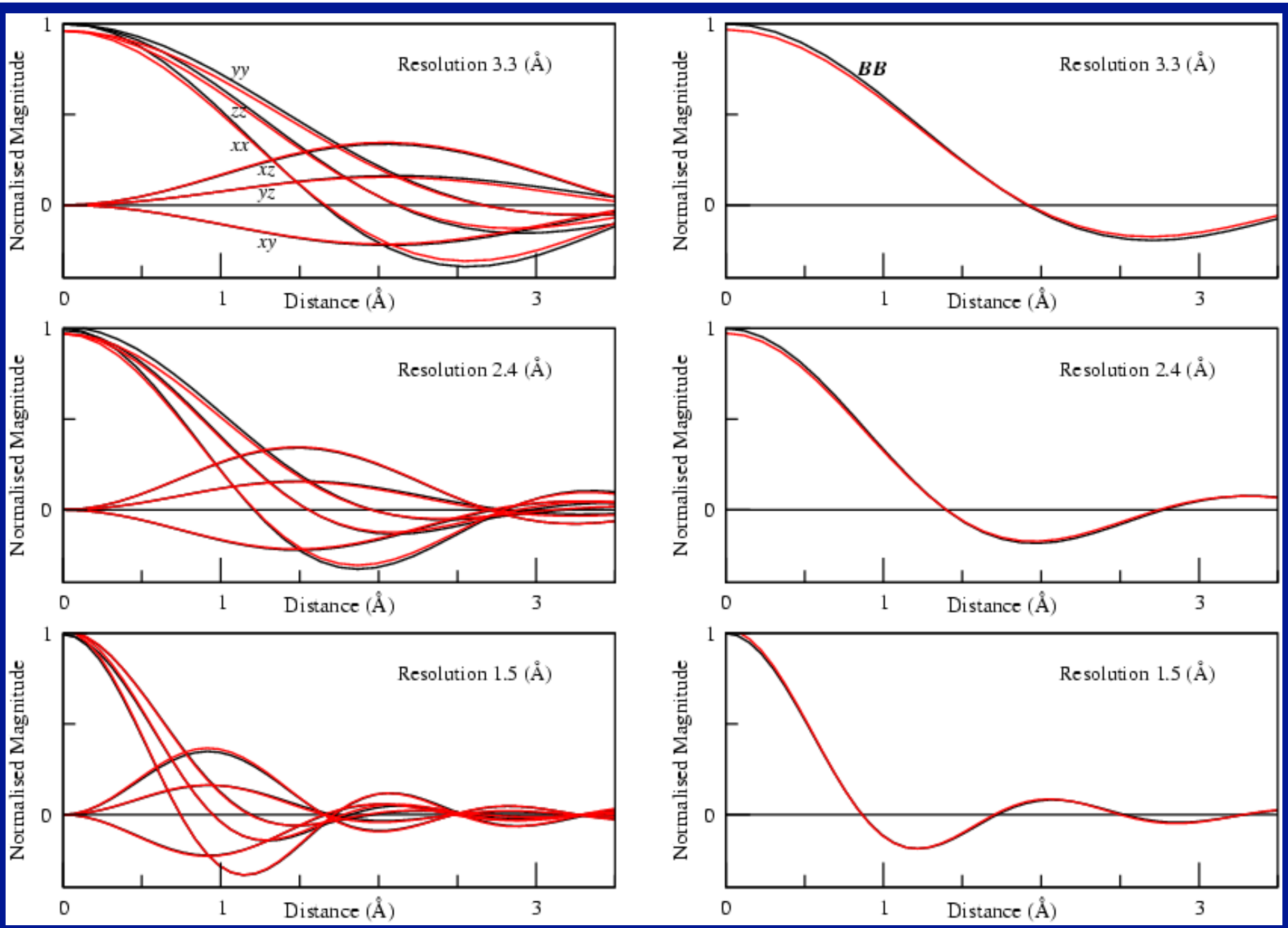


Continuous reciprocal space

$$I_{p_i(n),p_j(m)} \cong K_{p_ip_j} q_n q_m \int_{res.\ sphere} w_s H_{p_ip_j} f_n^0 f_m^0 t_n t_m \mathrm{trig}_{p_ip_j}\left(2\pi \underline{h}\underline{D}_{nm}\right)$$

[Agarwal, 1978]
[Dodson, 1981]
[Templeton, 1999]

Analytical / versus integral /



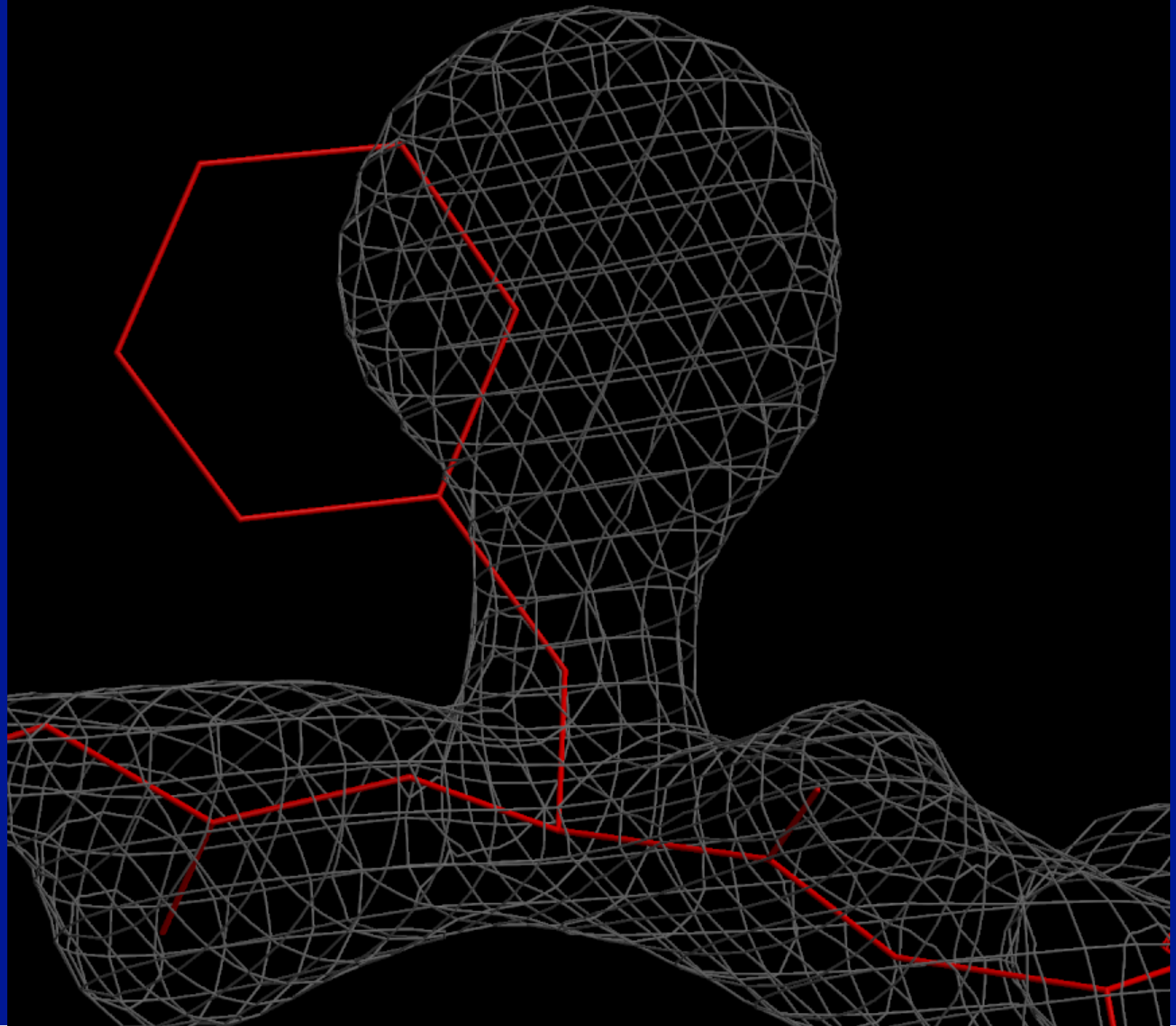
Fast evaluation of I

Two-step procedure

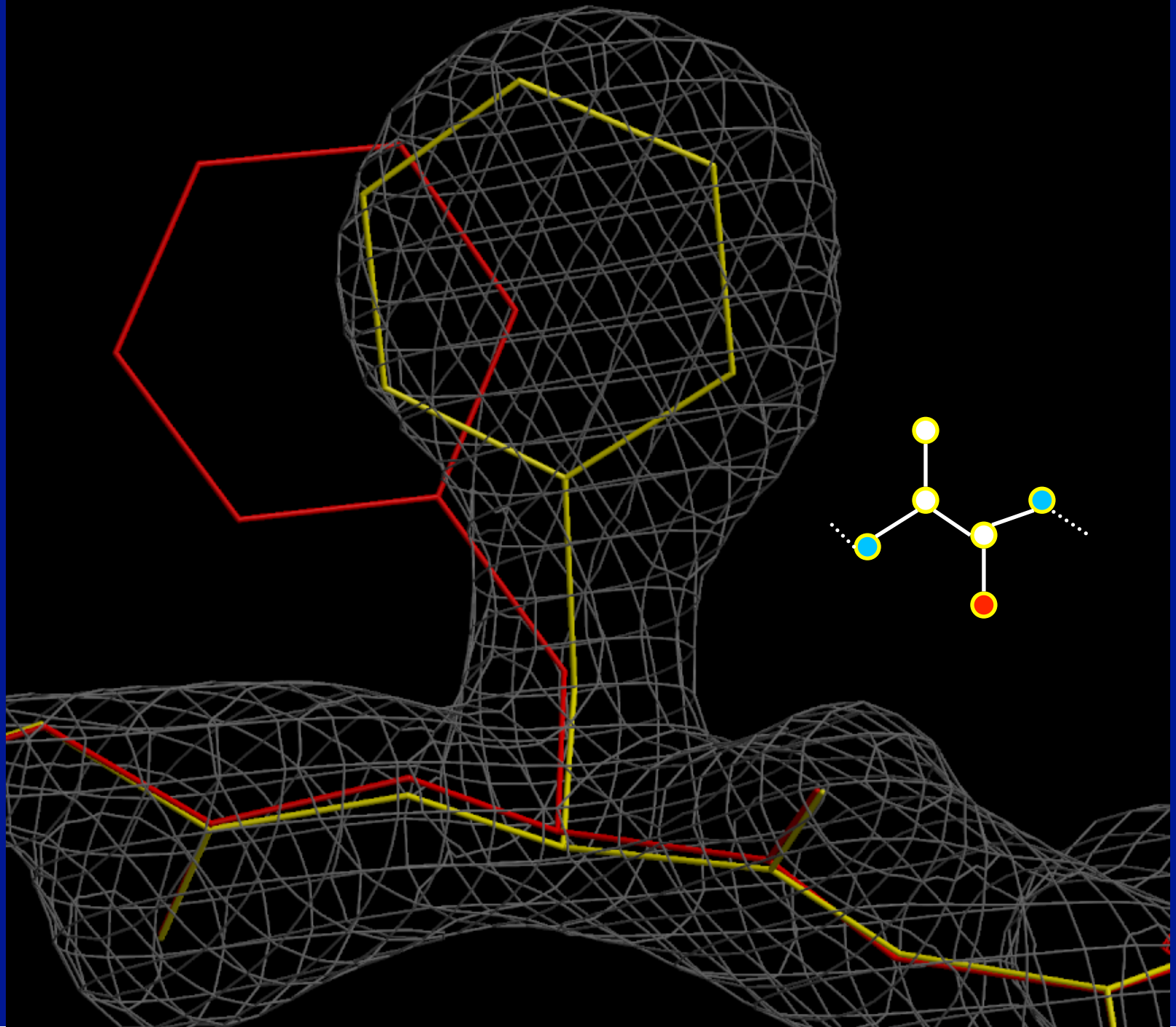
1. Tabulation step – a limited set of integrals are tabulated for different elements as a function of $\underline{D}_{nm}$ and B in a convenient coordinate system

2. Rotation step – the matrix element in the crystal system is calculated from the tabulated values using a rotation matrix

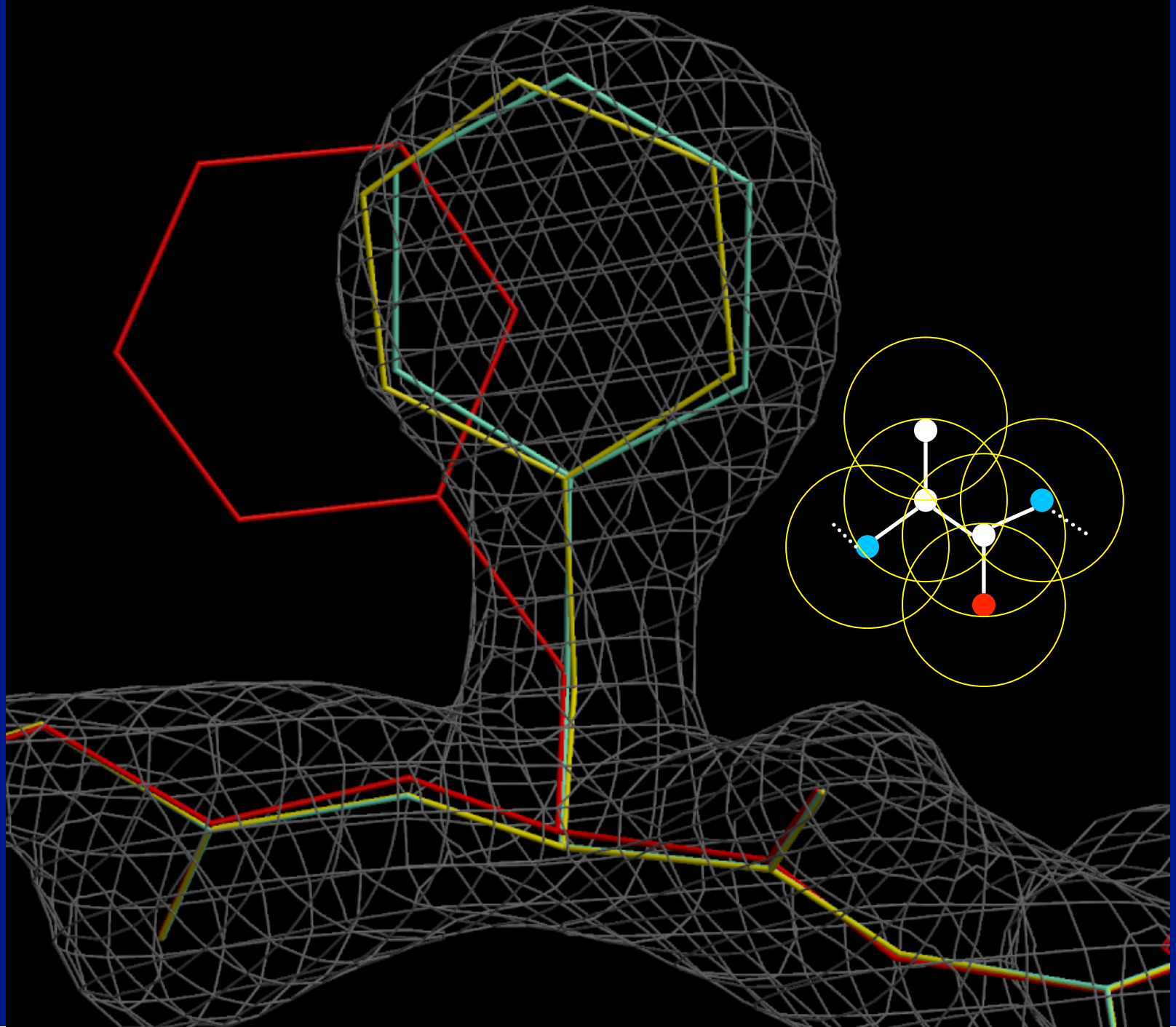
Use of different off-diagonal D_{max} cut-off



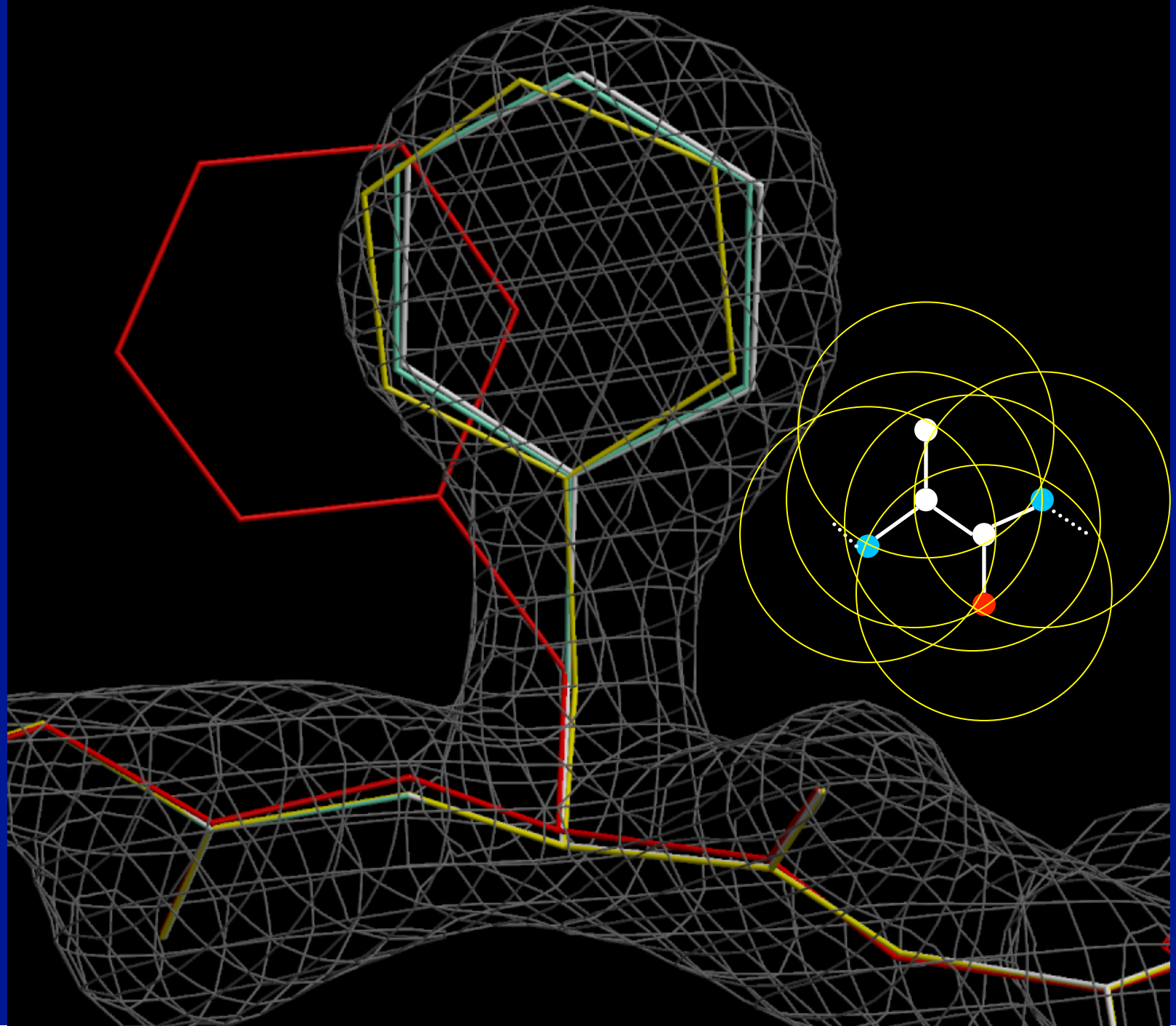
Use of different off-diagonal D_{max} cut-off



Use of different off-diagonal D_{max} cut-off



Use of different off-diagonal D_{max} cut-off



Key aspects of refinement

- Objective function
- Method of optimization
- Model parametrization
- Prior knowledge

Refinement summary

- **Model parameterization:**
  - quality of experimental data (resolution, completeness, …)
  - quality of current model (initial with large errors, almost final, …)
  - data-to-parameters ratio (restraints)
  - individual *vs* grouped parameters
  - knowledge based restraints/constraints (NCS, reference higher resolution model, etc…)

- **Refinement target:**
  - ML target is the option of choice for macromolecules
  - Real-space *vs* reciprocal space
  - Use experimental phase information if available

- **Optimization method:**
  - Choice depends on the size of the task, refinable parameters, desired convergence radius

Typical refinement steps

- **Input data and model processing:**
  - Read in and process PDB file
  - Read in and process library files (for non-standard molecules, ligands)
  - Read in and process reflection data file
  - Check correctness of input parameters
  - Create objects that will be reused in refinement later on (geometry restraints,…)
- **Main refinement loop (macro-cycle; repeated several times):**
  - Bulk solvent correction, anisotropic scaling, twinning parameters estimation
  - Update ordered solvent (water) (add or remove)
  - Target weights calculation
  - Refinement of coordinates (rigid body, individual) (minimization or Simulated Annealing)
  - ADP refinement (TLS, group, individual isotropic or anisotropic)
  - Occupancy refinement (individual, group, constrained)
- **Output results:**
  - PDB file with refined model
  - Various maps (2mFo-DFc, mFo-DFc) in various formats (CNS, MTZ)
  - Complete statistics
  - Structure factors

Refinement - summary

- **Refinement is:**
  - Process of changing model parameters to optimize a target function
  - Various strategies are used (restraints, different model parameterizations) to compensate for imperfect experimental data

- **Refinement is NOT :**
  - Getting a ‘low enough’ R-value (to satisfy supervisors or referees)
  - Getting ‘low enough’ B-values (to satisfy supervisors or referees)
  - Completing the sequence in the absence of density